FALA 2010

"VI Jornadas en Tecnología del Habla" & "II Iberian SLTech"

Speech and Language Technologies for Iberian Languages

PROCEEDINGS

10-12 November 2010

Centro Social Caixanova, Vigo, Spain

UniversidadeVigo

Multimedia Technology Group (GTM)

Spanish Thematic Network on Speech Technology (RTTH)

ISCA Special Interest Group on Iberian Languages (SIG-IL)

PROCEEDINGS

FALA 2010

"VI Jornadas en Tecnología del Habla"

and

II Iberian SLTech Workshop

November 10-12, 2010
Centro Social Caixanova
Vigo, Spain

Editors: *Carmen García Mateo*
Francisco Campillo Díaz
Francisco Méndez Pazó

Universidade de Vigo

fala2010.uvigo.es

ISBN: 978-84-8158-510-0

Abstracts and Proceedings Production by Diego Estévez González
diego.estevez.gonzalez@gmail.com

Welcome Message

On behalf of the entire FALA2010 Organizing Committee, the Spanish Thematic Network on Speech Technology (RTTH) and the ISCA-Special Interest Group on Iberian Languages, it is our pleasure to welcome you to the FALA2010 Conference in Vigo, Spain, held from November 10 to 12, 2010. Our Conference is the joint event of "VI Jornadas en Tecnologías del Habla" and "II Iberian Workshop on Speech and Language Technologies for Iberian Languages".

The Spanish Thematic Network on Speech Technology ("Red Temática en Tecnología del Habla, (RTTH)) has been promoting every other year the "Jornadas en Tecnologías del Habla" since 2000. Previous workshops were held in Sevilla (2000), Granada (2002), Valencia (2004), Zaragoza (2006) and Bilbao (2008). On the other hand, the ISCA Special Interest Group on Iberian Languages (SIG-IL) decided to organize the "I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages". The I Iberian SLTech was held in Porto Salvo, nearby Lisbon, on September 2009. Both events pursue the aims of being a meeting point to present and discuss the results of the research on speech and language technologies on Iberian languages. They also aim at promoting industry/university collaboration.

This year these two separate events joint efforts in order to take advantage of synergies and aiming at gathering most of members of the scientific community working on speech and language technologies for the Iberian Languages. Therefore, all the community working on processing of the Iberian Languages has been invited to participate in this three day event planned to promote interaction and discussion. There will be a wide variety of activities: technical papers presentations, keynote lectures, presentation of project reports and laboratories activities, demos, and recent PhD thesis presentations.

The 48 accepted papers were organized into 4 oral sessions and 3 poster sessions. The review task was complex and challenging and we appreciate Program Chairs and Area Coordinators' efficient and timely handling of this process as well as all the reviewers who helped in the process.

Special thanks are due to three keynote speakers. On 10 November, Dr. Heiga Zen from Toshiba Research Europe Ltd. (UK) will give a talk entitled "Fundamentals and recent advances in HMM-based speech synthesis". On 11 November, Dr. Alex Acero from Microsoft Research (USA) will describe some new approaches to speech recognition in his keynote entitled "New Machine Learning approaches to Speech Recognition". On November 12, Dr. Bill Byrne from the University of Cambridge will give the talk entitled "Hierarchical phrase-based statistical machine translation with weighted finite state transducers".

In order to promote the development of language and speech technologies and to encourage research labs to assess the performance of their research systems, the Spanish Thematic Network on Speech Technology has been performing in conjunction with the "IV y V Jornadas en Tecnologías del Habla" the Albayzín 2006 and Albayzín 2008 evaluation campaigns. Following the success of the above Albayzín evaluation campaigns, this year the Albayzín 2010 evaluation campaign has been performed in the topics: Speech Synthesis, Audio Segmentation, Speaker Diarization and Language Recognition. The results of evaluations will be presented during at a special session on November 10.

We would like to acknowledge to all those who made this FALA2010 possible and hopefully successful. We are very grateful to the University of Vigo, Caixanova, ISCA, the Spanish Ministry of Science and Innovation and the city Council of Vigo that have supported FALA2010 in various ways, for their important cooperation.

We thank the Spanish Thematic Network on Speech Technology and the ISCA Special Interest Group on Iberian Languages for their sponsorship of 7 awards: 4 awards for the 4 best student papers presented at FALA 2010, 1 award for the best demo/project presented at the demo/project session and 2 awards for the best PhD thesis presented at the PhD thesis session.

Last but not least, we thank you for all the contributions to the FALA 2010 conference: for attending sessions, presenting papers, being session chairs, being committees' members and performing all those other necessary functions. We once again welcome you to the exciting city of Vigo and wish you a fruitful Conference.

Enjoy FALA2010 in Vigo!

Carmen Garcia Mateo

Ruben San Segundo Hernández

Antonio Teixeira

Chairs of FALA2010 Conference

Organizing Committee

General Chairs:

Carmen García Mateo, University of Vigo - Spain
Rubén San Segundo Hernández, UPM- Spain
Antonio Teixeira, University of Aveiro - Portugal

Program Chairs:

José A. Rodríguez Fonollosa, UPC - Spain
Antonio Cardenal López, University of Vigo - Spain
Francisco Campillo Díaz, University of Vigo - Spain

Demos Chairs:

Daniela Braga, Microsoft (MLDC)- Portugal
Rubén San Segundo Hernández, UPM- Spain
Eduardo Rodríguez Banga, University of Vigo - Spain

Awards Chairs:

Miguel Sales Dias, Microsoft (MLDC)- Portugal
Climent Nadeu, UPC- Spain

Evaluation Chairs:

Laura Docio-Fernandez, University of Vigo - Spain
Francisco J. Méndez Pazó, University of Vigo - Spain

Technical Program Committee

Area Coordinators:

Speech Technology and Applications
Antonio Cardenal López, University of Vigo
Francisco Campillo Díaz, University of Vigo

Speech production
Adrià de Gispert, University of Cambridge

Natural Language Processing
Adrián Rodríguez Fonollosa, UPC

Resources, Standarization and Evaluation
Lluís Padró Cicera, UPC

Reviewers

Name	Forename	Affiliation
Aldezabal Roteta	Izaskun	Ixa Research Group - EHU / UPV
Arriola Egurrola	José María	EHU / UPV
Barbosa	Plínio	Universidade Estadual de Campinas (UNICAMP)
Benedí Ruiz	José Miguel	Universidad Politécnica de Valencia
Benitez Ortúzar	Mª Carmen	Teoría de la Señal, Telemática y Comunicaciones- UGR
Bonafonte Cávez	Antonio	Teoría de la señal y comunicaciones- UPC
Braga	Daniela	MLDC/Microsoft
Campillo	Francisco	Universidad de Vigo
Cardenal	Antonio	Universidad de Vigo
Cardeñoso Payo	Valentín	Universidad de Valladolid
Carvalho	Paula	Faculdade de Ciências, Universidade de Lisboa
Casacuberta Nolla	Francisco	Dpto. Sistemas Informáticos y Computación –UPV
Castro Bleda	María José	Universidad Politécnica de Valencia
Cordoba Herralde	Ricardo de	Speech Technology Group. UPM
de Gispert	Adrià	University of Cambridge
de la Mota Gorriz	Carmen	Universitat Autònoma de Barcelona
de la Torre Vega	Ángel	Universidad de Granada
Docío Fernández	Laura	Teoría de la Señal y Comunicaciones- UVIGO
Escudero Mancebo	David	University of Valladolid
Faúndez Zanuy	Marcos	Telecomunicaciones y arquitectura de computadores
Ferreiros López	Javier	Grupo de Tecnología del Habla – UPM
Gallardo Antolín	Ascensión	Universidad Carlos III de Madrid
García Granada	Fernando	Sistemas Informáticos y Computación –UPV
García Mateo	Carmen	University of Vigo
Gatius Vila	Marta	Lenguajes y sistemas – UPC
Godino Llorente	Juan Ignacio	Universidad Politécnica de Madrid
Gojenola Galletebeitia	Koldo	EHU / UPV
Gómez Vilda	Pedro	Universidad Politécnica de Madrid

Hernández Gómez	Luis A.	SSR- Universidad Politécnica de Madrid
Hernando Pericas	Francisco Javier	Teoría de la señal y comunicaciones – UPC
Hurtado Oliver	Lluís Felip	Universitat Politécnica de Valencia
Iglesias	Gonzalo	University of Cambridge
Klautau	Aldebaro	Universidade Federal do Pará
Lleida Solano	Eduardo	Universidad de Zaragoza
Llisterri Boix	Joaquím	Universitat Autònoma de Barcelona
López Soto	María Teresa	Universidad de Sevilla
López-Cózar Delgado	Ramón	Dpto. Lenguajes y Sistemas Informáticos – UGR
Machuca Ayuso	María Jesús	Universitat Autònoma of Barcelona
Mamede	Nuno	IST / INESC-ID Lisboa
Mariño Acebal	José B.	Universitat Politècnica de Catalunya
Martínez Barco	Patricio Manuel	Universidad de Alicante
Martínez Hinarejos	Carlos David	Universidad Politécnica de Valencia
Montero Martínez	Juan Manuel	Universidad Politécnica de Madrid
Montoyo Guijarro	Andrés	Lenguajes y Sistemas Informáticos
Moreno Bilbao	Asunción	Universitat Politecnica de Catalunya
Muñoz Guillena	Rafael	Lenguajes y Sistemas Informáticos
Nadeu Camprubi	Climent	Teoría de la señal y comunicaciones – UPC
Navarro Mesa	Juan Luis	Universidad de Las Palmas de Gran Canaria
Neto	João	INESC-ID/IST
Nolazco Flores	Juan	Tecnológico de Monterrey, Mexico
Ortega Giménez	Alfonso	Ingeniería Electrónica y Comunicaciones – U. Zaragoza
Palazuelos Cagigas	Sira Elena	Electrónica
Peinado Herreros	Antonio Miguel	Universidad de Granada
Peláez Moreno	Carmen	Universidad Carlos III de Madrid
Pérez Córdoba	José Luis	Universidad de Granada
Pla Santamaría	Ferrán	Sistemas Informáticos y Computación – UPV
Quaresma	Paulo	Universidade de Évora
Rauber	Andreia	Universidade Católica de Pelotas
Rigau Claramount	German	University of the Basque Country
Rodríguez Fonollosa	José Adrián	Universitat Politècnica de Catalunya
Rodríquez Banga	Eduardo	Teoría de la Señal y Comunicaciones - UVIGO
Rodríquez Fuentes	Luis Javier	Universidad del País Vasco
Rubio Ayuso	Antonio José	Teoría de la Señal, Telemática y Comunicaciones - UGR
Saiz Noeda	Maximiliano	University of Alicante
Sánchez Calle	Victoria Eugenia	Teoría de la Señal, Telemática y Comunicaciones - UGR
Sánchez Peiró	Joan Andreu	Dpto. Sistemas Informáticos y Computación - UPV
Sanchís Arnal	Emilio	Sistemas Informáticos y Computación - UPV
San-Segundo Hernández	Rubén	Grupo de Tecnología del Habla UPM
Sarasola Gabiola	Kepa	Euskal Herriko Unibertsitatea
Segarra Soriano	Encarnación	Sistemas Informáticos y Computación - UPV
Silva	Mário	Universidade de Lisboa
Simões	Alberto	ESEIG/IPP
Teixeira	Carlos	Faculdade de Ciências, Universidade de Lisboa
Teixeira	António	IEETA/Universidade de Aveiro
Torre Toledano	Doroteo	ATVS, Escuela Politécnica Superior UAM
Varona Fernández	María Amparo	Electricidad y Electrónica - EHU / UPV
Vicedo González	José Luis	Lenguajes y Sistemas Informáticos
Vivaracho Pascual	Carlos Enrique	Informática- U. Valladolid

Conference Program at a Glance

Wednesday November 10	Thursday November 11	Friday November 12
08:00 On-Site Registration		
08:50 *Opening Ceremony*	08:45 On-Site Registration	08:45 On-Site Registration
09:20 O1- Oral Session Speech Production and Synthesis	09:00 O2- Oral Session Speech Recognition	09:00 O4 - Oral Session Machine Translation and Technology Development
10:40 Coffee Break	10:40 Coffee Break	10:40 Coffee Break
11:00 *Keynote Talk* *Heiga Zen*	11:00 *Keynote Talk* *Alex Acero*	11:00 *Keynote Talk* *Bill Byrne*
12:00 P1 - Poster Session	12:00 P2 - Poster Session	12:00 P3 - Poster Session
13:30 Lunch	13:30 Lunch	**13:30 *Closing Ceremony***
15:30 Albayzin'10 Evaluation Session (Oral)	15:30 O3 - Oral Session Speaker Characterization	
17:10 Coffee Break	17:10 Coffee Break	16:00 Optional Visit Trip to Santiago de Compostela and guided visit to the old town and cathedral.
17:30 Albayzin'10 Evaluation Session (Posters)	17:30 Thesis/Project/DemoSession	
18:30 RTTH Assembly		
20:00 - Welcome Reception Pazo Museo "Quiñones de León"	20:30 Gala Dinner Hotel Bahía	

Table of Contents

Oral Session 1: Speech Production and Synthesis

Defining analogy for non-native inclusions in Spanish TTS

Tatyana Polyákova, Antonio Bonafonte

Universitat Politècnica de Catalunya, Barcelona, Spain
tatyana.polyakova@upc.edu, antonio.bonafonte@upc.edu

Abstract

Mass media globalization introduces the challenge of multilingualism into most popular speech applications such as text-to-speech synthesis and automatic speech recognition. In Spain as well as in the other countries, the usage of English words is rapidly growing, however due to the linguistic diversity of the languages spoken across the country, Spanish is not less influenced by inclusions from the four official languages. This work is focused on the pronunciation of Catalan inclusions in Spanish utterances. Our goal was to approach the nativization phenomenon by data-driven methods, making it easily transferable to other languages without loss in performance. For this particular task, training and test nativization corpora were manually crafted and the task itself was approached using pronunciation by analogy. The results were encouraging and showed that even small corpus of 1000 words allows to capture the analogy in the nativization process. The resulting pronunciations allowed significant improvements in the intelligibility of Catalan inclusions in Spanish utterances.

Index Terms: nativization of Catalan words, grapheme-to-phoneme conversion, phoneme-to-phoneme conversion, Spanish TTS, pronunciation by analogy

1. Introduction

Speech technologies in the framework of their rapidly expanding usage must be adapted to the multilingual scope allowing a higher level of flexibility and answering the modern users' needs. The text-to-speech synthesis finds many important applications on the emerging market of speech technologies. Voices capable of embracing more than one language are highly demanding in the era of mass media globalization. The TTS systems are used in telephone companies, smart phones, car navigation systems and recently in speech-to-speech translation, a technology that is highly demanded due to the globalization of the world industry and mass media.

Every language receives a constant incoming flow of new words. In addition to the natural process of appearance of neologisms, by morphological or semantic word and word meanings creations, a lot of new words come to the current language from other languages. The are several ways that the words of foreign origin are incorporated into a receptor language.

Very few databases containing non-native pronunciation are available, while the nativization corpora is simply inexisting. This need for training data lead us to a creation a minimalistic nativization corpus described in Section 2.

In order to have a synthesizer always up-to-date we need an ultimate automatic method for the derivation of the nativized pronunciation. The problem of foreign words, more particularly, of proper names of foreign origin was studied in [1]. The goal in [1] was to transcribe proper names of different origins correctly from the point of view of English phonetics. The nativization problem and different influencing factors were also described in [2] and [3]. Summarizing all possible influence factors and the difficulties encountered for the correct nativization of non-native words we are betting on an approach that can combine the knowledge of the orthographic and phonetic forms in the language of origin with pronunciation adaptation rules to the target language. In [4] it was demonstrated that the analogy between the nativized pronunciation and the original one can be inferred in a reliable and simple way since the nativization of English words in a Spanish text given the English pronunciation is an easier task than the native pronunciation of unknown English words and yet all human attempts to nativization are highly dependable on the analogy between known and unknown words. In [4] the final goal of the nativization was to be able to correctly pronounce English phrases in Spanish utterances as well as those out-of-dictionary proper names, commercial trademarks, etc., such as *Bruce* or *PlayStation* that were incorporated into Spanish in an already nativized form. This work is extended to Catalan inclusions in Spanish. The resulting nativized pronunciations should be well accepted by general Spanish audiences.

This paper is organized as follows, Section 2 explains the creation of the training and test datasets, Section 3 explains the algorithm used for nativization of Catalan words. Section 4 followed up by conclusions summarizes the experimental results obtained.

2. Nativization database

The main idea of this work was to train a nativization model to convert Catalan pronunciations to acceptable Spanish ones, adapting in a suitable way the pronunciation of the phonemes that do not exist in Spanish depending on such factors as frequency of usage of the word, and Spanish phonemization rules. Two nativization scopes were exploited: 1)training of a nativization model using the information about the orthographic form and the nativized phonetic transcription and 2) usage of the original Catalan and nativized to Spanish pronunciations for training. In order to apply data-driven techniques to nativization a need for training and test data raises. For usual grapheme-to-phoneme conversion tasks large pronunciation corpora of 100 thousands words and their corresponding pronunciations are available. Since we did not find any existing nativization database we chose to manually create a minimalistic corpus that would not require expert linguistic knowledge. For our task the training corpus was orthographically balancedin order to have all possible letter bi-grams in the corpus, we selected s total of 1000 words. The original phonetic transcriptions of these words were manually nativized according to the criteria described by Llorente in the book of styles for one of the Spanish TV chan-

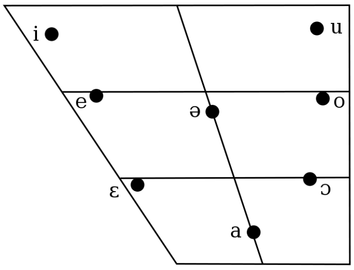


Figure 1: Vowels of standard Eastern Catalan

nels [5]. It is necessary to emphasize that the phoneme inventory used for nativization was limited to the Spanish phoneset. The test data was manually collected from the available on-line sources. Since a thousand words was selected for training, it was found appropriate that the test data comprised 10% of the training corpus. None of the test words were present in the training dictionary. It was intended that the test words were frequently used and with simple meaning in order for the results to be unbiased by other factors. Here are some examples of train *agredolça, boirumós, migjorn* and test words *enllaç, desig, forjar*.

2.1. Phonetic differences

The sounds of a language are defined by a phoneme inventory or phoneset. A phenomenon called extension of the phoneset often occurs in bilingual communities and speakers; however, it is impossible to study foreign word pronunciation on the level of the individual. In bilingual societies, it is much easier to observe general tendencies. In the particular case of Catalan, both of nativization and phoneset extension phenomena occur. It is curious to note that Spanish words in Catalan are pronounced using the regular Spanish phoneset, due to the fact that the majority of Catalan speakers are perfectly fluent in Spanish. For example, the Spanish name *Jorge* in Catalan is pronounced `/x 'o r x e/` and not `/dZ 'o r dZ @/` as Catalan phonetics would stipulate even though the phoneme `/x/` is absent from Catalan it is used for Spanish names. On the contrary, the pronunciation of Catalan words in Spanish is adapted according to Spanish pronunciation rules and the phoneset extension phenomenon is rare. Spanish and Catalan have several major phonetic differences which depend on the dialect of the latter. Most varieties of Catalan possess seven stressed vowels that are: `/a/`, `/e/`, `/o/`, `/u/`, `/i/`, `/E/`, `/O/`,`/@/`. The open vowels `/a/`, `/E/` and `/O/` as well as the unstressed `/@/` do not occur in Spanish. In Spanish medium vowels can be realized as open only in particular contexts while in the rest of the cases all vowels are articulated as closed. In Catalan, however there is an important phonological difference between open and closed vowels which can not be attributed to the context and therefore is not predictable. For example, homographs *seu (yours) vs. seu (headquarters)* `/s 'e u/` vs. `/s 'E u/` have different meanings depending on the vowel articulation point. A diagram of Catalan vowels can be found in Figure 1 As well as in Spanish, in Catalan there are six plosives `/b/`, `/d/`, `/g/`, `/p/`, `/t/`, `/k/` (3 voiced and 3 unvoiced) at three different articulation points. Catalan does not have any dental, uvular or velar fricative consonants sounds, but has two alveolo-palatals `/Z/` (voiced) e.g. *vigent* `/b i Z 'e n/` and `/S/` (unvoiced) e.g. *caixa* `/k 'a S a/`. The labiodental `/v/` exists in Catalan as a result of sonorization of any `/f/` before a voiced consonant or a vowel at the beginning of the word. Besides, in Catalan, all unvoiced fricatives are sonorized if followed by a voiced consonant. The fricative `/z/` which is very frequent in Catalan, exists in Spanish as an allophone but not as a phoneme. Catalan has four affricates which is 3 more than Spanish, the voiced affricates `/dz/`, `/dZ/` and the unvoiced `/ts/`, `/tS/`. The phonemes `/ts/` and `/dz/` arise mainly from compounding such as in *potser* /p u ts 'e/, but may as well occur at any other position as in *dotze* `/d 'o dz @/`. Similarly to Spanish the nasals are adapted to the articulation point of the following consonant, however in Catalan both `/m/` and `/J/` can occur at the end of the word e.g. *any* `/'a J/`. There are two laterals in Catalan as well as in Spanish, the alveolar `/l/` and the alveolopalatal `/L/`. Additionally Catalan has the double *l.l* that is pronounced as a prolonged alveolar articulation `/l:/`. Both languages possess 2 trills, the simple `/r/` and the multiple `/rr/`. In contrast with Spanish the `/r/` in Catalan can only appear at the intervocalic position or after a plosive or fricative that forms part of the same syllable e.g. $frau$ `/ f r 'a uw/` or $cara$ `/k 'a r a/` [6]. A complete set of Catalan consonants can be found in Figure 2. Taking into account most of the above

	Bilabial	Labiodental	Interdental	Dent.-alv.	Alveolar	Alveol.-palatal	Palatal	Velar
Ocl.	p b			t d				k̟k k̠ g
Fric.		f v		s̺ z̺	s z	ʃ ʒ		
Afric.					t͡s d͡z	t͡ʃ d͡ʒ		
Aprox.	β		ð̞				j	ɣ
Nasal	m	ɱ		n̪	n	nʲ ɲ		ŋ
Later.				l̪	l	lʲ ʎ		ɫ
Vibr. simp.					ɾ ɹ			
Vibr. múlt.					r			

Figure 2: Catalan consonants [6].

mentioned phonetic differences, we developed nativization criteria in order to find the best pronunciation for Catalan words in Spanish utterances.

2.2. Criteria

The challenge of this task consisted of developing solid criteria for nativization, taking into account local specifications of certain words, pronunciation and word popularity factor, among others. Some of the criteria could not be easily formulated that is why using a training corpus clearly has an advantage over the rule-based approach. Several examples of the criteria used are described below. All open vowels were mapped to the closed ones, while the unstressed `/@/` was mapped to `/a/` in most of the cases except for those words that are were similar to Spanish where it was transcribed as `/e/` e.g. *adrearà* from `/a D r a s a r 'a /` to `/a D r e s a r 'a/` in the nativized form. For consonants, some difficulties were found when transcribing `/Z/ /S/`. Their nativization depended both on letter and phoneme context. The voiced fricative `/Z/` at the beggining of the word was nativized to `/jj/` e.g $jutge$, to `/tS/` before a nasal e.g. $taronja$, and to `/j/` in other cases as in $vorejar$. The unvoiced fricative `/S/` was transcribed or to `/j s/` when it corresponded to the digraph ix e.g. $coix$; to `/tS/` when it corresponded to the same phoneme in a similar Spanish word e.g $anxoves(cat.)$ vs. $anchoas(sp.)$; or to `/s/` in the rest of the cases. The affricate `/dZ/` was nativized to `/tS/` as in the word $migdia$. Affricates `/ts/` and `/dz/` were mapped to the corresponding double phonemes `/t s/`

and /d z/. The multiple trill /rr/ was conserved only in the cases when it corresponded to the Spanish phonetic rules, in all other cases it was changed to the simple trill /r/. The nasal /N/ and the voiced /z/ were conserved as they were present in our voice database. Silent r at the end of the verb in a compound verb-pronoun construction such as $afegir - n'hi$, was restored in the nativized form for the sake of comprehension. The database nativization task was carried out by the authors using both the source language orthographic form and pronunciation. In the following section we describe the functionality of a multilingual grapheme-to-phoneme system used in this work.

3. Pronunciation by analogy

Data-driven approaches were proven to be more efficient than the ones based on the explicit linguistic modeling and they undoubtedly gain in adaptability [7]. For g2p conversion the best results were obtained using data-driven corpus-based methods. Pronunciation by analogy method previously used in [8, 9] was found to be the most efficient for grapheme-to-phoneme task. In this section we review the pronunciation by analogy algorithm. Our implementation is based on [8] with the new strategies introduced in [9]

3.1. Algorithm description

After the training dictionary has been aligned, the matcher starts to search for common substrings between the input word and the rest of the dictionary entries. Every input word is then compared to all the words in the lexicon in order to find common "arcs". Let us call the substrings in the grapheme context "letter arcs" and the corresponding substrings in the phoneme context "phoneme arcs". All the possible letter arcs with the minimum length of 2 letters and the maximum length equal to the input word length are generated and then searched in the dictionary. For every letter arc from the input word, matching with the same letter arc in a dictionary word, the corresponding pronunciation or the phoneme arc is extracted. The frequency of appearance of each phoneme arc corresponding to the same letter arc is stored along with the starting position and length for each arc.

Each time that for the same letter arc we find the same phoneme arc; the frequency of the phoneme arc is incremented. The matching phoneme arcs are introduced into the pronunciation lattice that can be represented by nodes and connecting arcs. If an arc starts at a position i and ends at a position j, and if there is yet no arc starting or ending at position j, the nodes L_i and L_j are added to the graph. An arc is drawn between them. All the nodes are labeled with the corresponding "juncture" phoneme and its position in the word. The arcs are labeled with the remaining phonemes and their frequency of appearance.

Each complete path through the lattice is called "pronunciation candidate". We considered only the shortest paths through the lattice [8]. If there is a unique shortest path, it is chosen as the best pronunciation and the algorithm stops. Usually there are several shortest paths through the lattice, and a decision function is necessary to choose the best pronunciation candidate among them.

Each candidate can be represented as C_j={F_j, D_j, P_j}, where $F_j = \{f_1,\ldots,f_n\}$ are the phoneme arc frequencies along the jth path, $D_j = \{d_1,\ldots,d_n\}$ are the arc lengths and $P_j = \{p_1,\ldots,p_k\}$ are the phonemes comprising the pronunciation candidate, being k the pronunciation length. Marchand and Damper in 2000 [8] proposed to use 5 scoring strategies in order to choose the best pronunciation.

In our previous work [9] we proposed 6 additional strategies for choosing the best candidate which in combination with the others outperformed the original ones. The scoring strategies are based on the following parameters, frequency of appearance of a given phoneme arc in the dictionary, its length and the actual phonemes which constitute the candidate. Different strategies work with different aspects of analogy. High arc frequency is considered to be a major advantage over the low arc frequency. The frequency of suffixes and prefixes are prioritized by different strategies. The more common phonemes the candidate shares with the others the higher will be its final score. If a candidate has exactly the same pronunciation as the other one both of them are prioritized. These measures are used separately or combined across the strategies. All the strategies previously used in grapheme-to-phoneme conversion are described below [8, 9]. The pronunciation by analogy algorithm was previously applied to grapheme-to-phoneme conversion [8, 9]. In this work it was extended to the nativization task.

4. Experimental results

The experimental results are given below for each method.

4.1. Previous results: nativization tables

In our previous work [10] we developed a nativization system based on nativization tables (Ntab). Pronunciations were derived according to the scheme shown in Figure **??**. The nativization was carried out in a phoneme-to-phoneme manner, using nativization tables for source→target phoneme transformations. The source language was Catalan and the target language was Spanish. Therefore all Catalan phonemes were mapped to the closest Spanish ones. The nativization tables were able to convert 79.74% phonemes and 21.78% words correct. These results are given for the same 100 word test corpus described in 2. However, these results are much better than those obtained without using nativization, applying the Spanish g2p to derive the pronunciation of Catalan words, Spanish g2p scored only 33.97% correct in phoneme and 3.96% on word nativization on a 100 word test corpus. The only words that this kind of system can "nativize" correctly are those that are pronounced very closely to Spanish orthography, for example $aqu\acute{}\,to$ /a k 'i/ or $sac\,to$ /s 'a k/.

4.2. Grapheme-to-phoneme nativization (g2p_nat)

The first hypothesis to be tested was prediction of nativized pronunciation by analogy in the orthographic context. Out of eleven strategies available in the PbA for choosing the best pronunciation candidate it was necessary to determine the best strategy combination for our data. All possible strategy combinations were considered and compared. For grapheme-to-phoneme nativization (g2p_nat) the resulting best strategy combination was the following: 0000101101. For each of the eleven strategies described in [9] 1 means that the strategy corresponding to that position was included and 0 means it was left out. The best results obtained on training data equaled to 86.51% in phoneme and 38.61% in word accuracy. When we considered each strategy individually the best results were obtained for the eleventh strategy that combines the frequency product with the frequency of the same pronunciation. The lowest scoring strategy is seventh strategy that prioritizes the candidates with very frequent first arc. The results for each single strategy and the best strategy combination can be found in Table 1

Table 1: Single strategy results for g2p_nat and best strategy combination

strategy mask	ph. acc.	word. acc.
10000000000	85.80	35.64
01000000000	83.96	31.68
00100000000	83.82	31.68
00010000000	83.79	31.68
00001000000	84.62	32.67
00000100000	85.65	33.66
00000010000	83.38	30.69
00000001000	85.53	33.66
00000000100	83.96	35.64
00000000010	83.69	30.69
00000000001	85.86	36.63
00001011011	86.51	38.61

4.3. Phoneme-to-phoneme nativization (ph2ph_nat)

For Catalan, it makes a lot of sense to perform grapheme-to-phoneme nativization, in fact, non-Catalan speakers apply Spanish grapheme-to-phoneme rules when reading Catalan; however, in order to find the best pronunciation for Catalan phonemes absent from Spanish the phonetic transcription available in the source language may be quite helpful. Finding automatic correspondences between source and target (nativized) phonemes is a more consistent task than in the case of letters, being g2p conversion already a difficult task for Catalan especially for such a reduced training corpus. For phoneme-to-phoneme nativization experiments the PbA was modified in order to receive phoneme input. The best strategy combination (11010101010) as in the g2p_nat case was determined performing n-fold evaluation of all possible strategy combinations. The results obtained on 100 word test set of common names are 92.09% phonemes and 56.44% words correct. These results show that p2p_nat nativization outperforms g2p_nat nativization by 22% in word accuracy terms. Performing single strategy experiments for phoneme-to-phoneme nativization we can also observe that the best scoring strategies are the sixth and the eight one, while the worst places belongs to the ninth. For more results see Table 2.

Table 2: Single strategy results for p2p_nat and best strategy combination

strategy mask	ph. acc.	word. acc.
11000000000	91.51	53.47
10000000000	91.51	52.48
01000000000	90.09	48.51
00100000000	90.51	48.51
00010000000	90.95	50.50
00001000000	90.79	49.50
00000100000	91.65	54.46
00000010000	89.94	46.53
00000001000	91.51	54.46
00000000100	89.18	42.57
00000000010	90.38	48.51
00000000001	90.92	50.50
11010101010	92.09	56.44

5. Conclusions

In this paper we proposed to use pronunciation by analogy for the nativization of Catalan words in Spanish utterances in the framework of a multilingual TTS system. The best results were achieved using phoneme-to-phoneme nativization based on the analogy in the phoneme context. The nativization results obtained using analogy only in the letter context were rather poor, due to the reduced corpus size. It is worth mentioning that even in the case of grapheme-to-phoneme nativization the results show very significant improvements in comparison to those obtained by direct phoneme-to-phoneme table-based mapping. Nativized pronunciations are more tolerant to the vowel and consonant substitution and the persisting errors are minor and are not crucial for intelligibility. Even though the test corpus that consisted of 100 hundred frequently used common Catalan names can be considered somewhat tiny, for both g2p_nat and p2p_nat methods n-fold evaluation was performed on the training corpus of 1000 rather infrequent common names (selected by the greedy corpus balancing tool) and the results obtained were quite similar to those obtained on the test data. Simple mapping rules were proven to be insufficient for the task because some of the criteria could not be easily formulated.

6. Acknowledgements

This work has been partially funded by the Spanish Government under grant TEC2009-14094-C04-01 (BUCEADOR project)

7. References

[1] A. Font Llitjos and A. Black, "Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names," in *Proc. the of European Conference on Speech Communication and Technology*, Genova, Italy, Sep. 2001.

[2] I. Trancoso, "Issues in the pronunciation of proper names: the experience of the Onomastica project," in *In Proceedings of Workshop on Integration of Language and Speech*, Moscow, Russia, 1995.

[3] I. Trancoso, C. Viana, I. Mascarenhas, and C. Teixeira, "On deriving rules for nativised pronunciation in navigation queries," in *Proc. the of European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999.

[4] T. Polyákova and A. Bonafonte, "Nativization of english words in spanish using analogy," in *The Seventh ISCA Tutorial and Research Workshop on Speech Synthesis, Kyoto, Japan, September 22-24*, 2010.

[5] J. Llorente and L. Díaz Salgado, *Libro de estilo de Canal Sur TV y Canal 2 Andalucía*. Radiotelevisión de Andalucía, 2004.

[6] A. Planas, *Así se habla: nociones fundamentales de fonética general y española: apuntes de catalán, gallego y euskara*. Horsori Editorial, Sl, 2005.

[7] A. van den Bosch and W. Daelemans, "Data-oriented methods for grapheme-to-phoneme conversion," in *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 1993, pp. 45–53.

[8] Y. Marchand and R. Damper, "A multistrategy approach to improving pronunciation by analogy," *Computational Linguistics*, vol. 26, no. 2, pp. 195–219, 2000.

[9] T. Polyakova and A. Bonafonte, "New strategies for pronunciation by analogy," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[10] T. Polyákova and A. Bonafonte, "Further improvements to pronunciation by analogy," in *Actas de las V Jornadas en Tecnologas del Habla*, Bilbao, Spain, Nov. 2008, pp. 149–152.

Analysis of Cat-ToBI indices intertranscriber inconsistencies: implications for automatic labelling

David Escudero-Mancebo, Lourdes Aguilar

Department of Computer Science, Universidad de Valladolid, Spain
Department of Hispanic Studies, Universidad Autonoma de Barcelona, Spain
descuder@infor.uva.es, Lourdes.Aguilar@uab.cat

Abstract

In this paper we present an experience to measure the inter-transcribers consisteny where a number of observer have been required to identify ToBI events in the same set of sentences. We computed the pairwise transcribers agreement with its corresponding confusion matrix and the kappa coefficients. The goal was to identify the main sources of confusion resulting: (1) bad trained observers (2) problematic symbols. The identification of those problematic symbols supports the practical decision to merge them into an alternative class when automatic approaches to ToBI labelling are focused; in this case for ToBI break indices.

Index Terms: prosody, ToBI, inter-transcriber consistency, automatic recognition

1. Introduction

ToBI is a standard for representing and labelling prosodic events including tones (accent tones and boundary tones) and breaks [1]. The tones level is used to mark the occurrence of phonological tones at appropriate points in the F0 contour. The break level is used to mark break indices, which are numbers representing the strength of the boundary between two orthographic words. The number 0 represents no boundary, 4 represents a full intonation phrase boundary and the rest of indices are breaks with intermediate strength. In this paper we focus on breaks as the genesis of this work was the potential interest of breaks for the representation of the utterance rhythmic structure with applications in text-to-speech systems.

ToBI has been implemented for several languages including English, German and Japanish. Concerning to Iberian languages, it exists active groups responsible for the Cat-ToBI and Sp-ToBI for Catalan and Spanish respectively. The need of a reference corpus similar as the ones existing for other languages (e.g. the Boston Radio Corpus for English [2]) is still a need both for Catalan and Spanish. The activity presented in this paper is included in the Glissando project [1], that has the aim to record and label with ToBI marks a bilingual Spanish and Catalan corpus containing Radio news recordings and spontaneous dialogs.

Labelling a corpus with ToBI tags is an expensive procedure. In [3] it is estimated that the ToBI labelling commonly takes from 100-200 times real time. To speed up the process, automatic or semiautomatic methods seem to be a productive resource. [4] or [5] are good examples of the state of art on automatic labelling of ToBI events. For Catalan we presented a work for labelling break indices [6]. In that work we reduced the set of break indices merging together some of them with the aim to increase the identification results. This merging strategy is common in other studies such the ones already mentioned of [4] or [5] that combine the different type of accent tones transforming the labelling problem into a binary one to decide weather an accent is present or not.

In this work we show that grouping different labels is a coherent procedure according to the diversity of judges observed in an inter-transcriber experiment. We present an experiment where different labellers are required to assign different ToBI tags in the same reduced set of sentences. Results seems to indicate that some of the ToBI tags are easier to confuse for the labellers. The more the confusion between a pair of classes the more the evidence that this pair of classes is a good candidate to be merged. We present a tool to compute and visualize the inter-transcriber inconsistency and we discuss about the inter ToBI labels confusion values.

First we present the experimental procedure with the corpus used, next the experimental procedure indicating which metrics have been applied and the procedure to visualize information. Finally we conclude with discussion and future work.

2. Experimental Procedure

A test of labeling consistency was conducted to measure inter-transcriber consistency in the Cat-ToBI prosodic transcription system in order to asess the system and to detect if there are labels frequently confused. Twenty utterances were excerpted from four different speech styles produced by twelve different speakers and transcribed by ten labelers differing in their levels of experience with Cat-ToBI.

2.1. Speech database

To assess the labeling conventions of Cat-ToBI and to demonstrate that these conventions are applicable to various types of speech, we selected twenty utterances representing four different discourse types: spontaneous speech excerpted from the database of the Atles interactiu de l'entonacio catala[2], in particular, from the intonation survey and the Map Task dialogue corpus; radio news and text reading (from the Festcat database[7]). Twelve speakers (5 male and 7 female) produced the sentences. These sentences contained a total of 264 words, and lasted a total of 89.8 seconds. Nine of the sentences are interrogative questions, four are emphatic declarative, and the rest, neutral declarative.

[1]Partially founded by the Ministerio de Ciencia e Innovacion, Spanish Government Glissando project FFI2008-04982-C003-02

[2]http://prosodia.upf.edu/atlesentonacio/metodologia/index-english.html

2.2. Subjects

The subjects span a variety of levels of experience with prosody and experience with Cat-ToBI ranging from absolute beginners to contributors to its development. The labelers were divided into three groups: Group 1 (Experts), Group 2 (Familiar with prosodic annotation systems), and Group 3 (Beginners, completely new to any model of intonation or prosodic transcription). Each group included four labelers, except for Group 2 which had two labelers. All subjects are native speakers of Catalan, with two dialects represented (Central Catalan, Balearic Catalan).

Each transcriber was provided with a document describing the Cat-ToBI system[8] and with the Cat-ToBI training materials[3]. The training materials contain a tutorial explaining each of the labels in Cat-ToBI, along with recorded examples of transcribed utterances. There are also exercises to practice the labels described in the text. The training materials were designed to be self-explicative. Moreover, absolute beginners attended a course (three sessions of three hours each) on the basics of AM model and the ToBI labelling systems.

All the labellers were given a document with basic instructions and a package with the sound files and the textgrids, with the Praat tool[4]. The selected speech has not been previously labeled by any of the transcribers; each transcriber worked alone on the samples and they were not allowed to discuss utterances in the experimental data-set. After they completed the transcription, their textgrid files were collected and statistics for labeler agreement were applied to the data.

2.3. Transcription procedure

The manual annotation was performed using the Praat tool. The transcribers were looking at a computer screen with a display of the signal (F0 curve and waveform) and they rely on auditive and visual information to take their prosodic decisions. The key elements to be labeled are prominence, prosodic boundary strength and pitch accent and boundary tone types.

In the ToBI framework, the transcribers have to perform the following tasks:

1. Mark syllables which are carrying a clear prominence, that is, decide if there is a pitch accent
2. If there is a pitch accent, decide the pitch accent type
3. Mark important between-word interruptions of the normal speech stream as either weak (signalling intermediate phrases) or strong breaks (that is, intonational phrases)
4. Decide the boundary tone type

2.4. Reliabilty measurements

2.4.1. Pairwise transcriber agreement

Agreement was measured by counting the number of labeling agreement for all pairs of transcribers. That is, 4 transcribers (T1, T2, T3, T4) would produce 6 possible transcriber pairs (T1T2, T1T3, T1T4, T2T3, T2T4, T3T4), and the criterion is conservative: if 3 of 4 transcribers agree, only 3 of 6 pairs will match, making the agreement rate 50% (agreement = agree / (disagree + agree)).

For example, if a particular word boundary was labeled by the first transcriber as 2, by the second transcriber as 3, and by transcribers 3 and 4, as 2, the number of transcriber pairs who agree with each other is three (T1T3, T1T4, T3T4) and the number of transcriber pairs who disagree with each other is also three (T1T2, T2T3, T2T4).

We are representing this information as a confusion matrix where rows and columns index the ToBI symbol. The main diagonal is indicates the coincidences and the rest of the elements are the discrepances.

2.4.2. Kappa coefficient

Cohen's kappa[9], which works for two raters, and Fleiss' kappa[10], an adaptation that works for any fixed number of raters, improve upon the pairwise transcriber agreement in that they take into account the amount of agreement that could be expected to occur through chance.

Agreement can be thought of as follows, if a fixed number of people assign numerical ratings to a number of items then the kappa will give a measure for how consistent the ratings are. The kappa, κ , can be defined as,

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{1}$$

Where $\bar{P}$ is an array measuring the agreement for the different symbols and $\bar{P}_e$ is the hypothetical probability of chance agreement. The factor $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance, and, $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

The inter-transcriber consistency for prominence, break strength and pitch accent and boundary tone inventory was quantified by means of kappa coefficient. According to [11], a kappa between 0.61 and 0.80 is considered to point at a substantial consistency. [12] considered a good level of agreement when the value obtained from the kappa statistic is greater than 0,7.

2.5. Visualizing the inconsistency

We use the Kappa Fleiss coefficient to obtain a symmetric matrix of distances between the ToBI events. This matrix is indexed in term of the type of break and represents the inconsistency between every pair of breaks. The more the distance the easier it was to distinguish this pair of breaks by the labellers. Given a pair of break indices, the whole set of decisions is binarized setting to `un-available data` the decisions that do not concern the selected pair and the Kappa Fleiss coefficient is computed. The more coincidences between the labellers referring to the given pair of breaks, the higher the corresponding κ value in the distance matrix.

Multidimensional scaling (MDS) is a set of related statistical techniques often used in information visualization for exploring similarities or dissimilarities in data. An MDS algorithm starts with a matrix of itemitem similarities, then assigns a location to each item in N-dimensional space, where N is specified a priori. For sufficiently small N, the resulting locations may be displayed in a graph or 3D visualisation. Multidimensional scaling will be used to display our distance matrix of break indices in a 2D plot. The closer the breaks, the more the confusion.

A similar procedure will be applied to obtain a distance matrix between the labellers. The more the agreement between a

[3]http://prosodia.uab.cat/cat_tobi/en/index.php
[4]http://www.praat.org

	B0	B1	B2	B3	B4
B0	414	164	6	23	1
B1		332	29	61	1
B2			8	22	4
B3				105	39
B4					163

Figure 1: Confusion matrices where cells compute the number of ocurrencies of the pair indexed by row and column

pair of labellers the closer will be displayed in a 2D plot. The distance between every pair of labellers will be computed as $1-\kappa$, been κ the inter-transcriber agreement of the pair of labellers.

We use classical multidimensional scaling [13], in particular its implementation in R^5 `cmdscale` procedure. We used the Interrater Reliability and Agreement. (irr) R package to compute the kappa coeficients.

3. Results

The global intertranscriber rate of agreement is 74.49 % which is a moderate result when compared with the test performed with consolidate ToBI systems in the rates trainning process: Previous works on intertranscriber reliability of ToBI-framework systems have certified between 81% and 92% of agreement in determining pitch accents for English [12], overall mean scores of 88.9% of agreement for German [14], and agreement percentages of between 59% and 91% (depending on accent categories) for Korean ([15]).

When the intertranscriber rate of agreement is split in the corresponding confusion matrix (table 1) we see clearly that there are important differences among the indexes. Thus, breaks 0 and 4 are identified easily, meanwhile the break 2 is identified with about 10% agreement. With respect to the breaks 1 and 3, the confusion with the other indexes is also high.

To display these discrepancies, we use the kappa coefficient. Table 2(top) shows a table of the kappa Fleiss coefficient for every pair of break indices according to the procedure explained in section2.4.2. Figure 2(down) interprets the kappa Fleiss coefficients as distances to apply multidimensional scaling. We can observe that it seems to be three groups of breaks: break 0, break 4 and a third group formed by the breaks 1, 2 and 3.

We obtain a kappa coefficient of 0.666 that corresponds to a substantial agreement in the commonly used kappa scale. As it was explained in section 2.2, there are three groups of raters: Experts, Beginners and Intermediates. If we separate the rates assigned by these two groups we obtain a kappa coefficient of 0.75 in the expert group. To display these discrepancies among the taggers we refer again to section 2.4.2 to build the table and figure in 3. We remark here the important differences among the different labellers. This type of figures could potentially be used to check the labeller reliability, under the supposition, that the closer the labeller is to an expert, the more accurate his or her rates are. Thus, the labeller **i1** behaves as a goat (in biometric terminology) meanwhile others behave as sheeps, always close to an expert.

[5] The R Project for Statistical Computing `http://www.r-project.org/`

	B0	B1	B2	B3	B4
B0		0.715	0.667	0.639	0.746
B1			0.586	0.602	0.659
B2				0.524	0.727
B3					0.746
B4					

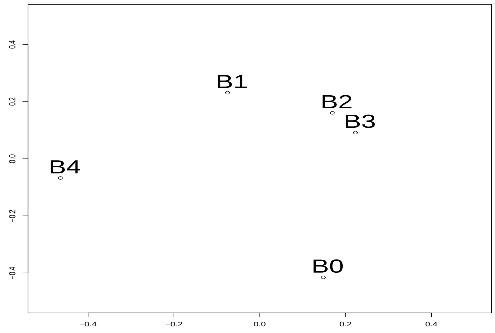


Figure 2: Multidimensional 2D plot of the the distances between de different Breaks (Kappa coefficients in the table above). B0..B4 are the different break indexes.

Figure 4 illustrates the inter transcribers confusion when the different breaks are isolated. Once again we see that the breaks 0 and 4 are the less problematic. Break 2 is very confusing. Table 1 shows that the break 2 appears very few and when it does, it is marked by 1 or 2 labellers most of the times. This observation makes it congruent the merging of break 2 with other classes in practical situations.

With respect to the break 3 and break 1 distinction, figure 4 seems to indicate that there are two groups of taggers (as marked on the figure). This cluster of labellers could be indicating the use of different criteria and it is something to analyze in future works. In the time this divergent criteria is solved, it seems congruent to merge the labels into only one reducing the number of break tags from five to only three as it was done in our previous work [6] on automatic identification of break indices.

	1	2	3	4	5	6	7	8
B0	25	16	3	5	4	5	9	48
B1	11	11	9	7	5	27	13	20
B2	24	3	1	2	1			
B3	10	6	6	11	5	4	5	4
B4	2		6	3	1			22

Table 1: Frecuency of different ToBI labels: the cell quantity is the number of times that the break index of the row was labelled by the number of labellers indicated by the column.

4. Conclusions

In this paper we have run a test of inter-transcribers consistency consisting on the ToBI labelling of a set of Catalan sentences by a number of observers.

Results show that one of the main sources of confusion has its origin in bad trained labellers which observations separates clearly from the experts labellers. Another source of noise is

	b1	b2	b3	E1	E2	i1	E3	E4
b1		0.665	0.645	0.653	0.79	0.597	0.668	0.665
b2			0.699	0.832	0.69	0.537	0.631	1
b3				0.658	0.705	0.504	0.57	0.699
E1					0.728	0.516	0.619	0.832
E2						0.547	0.727	0.69
i1							0.556	0.537
E3								0.631
E4								

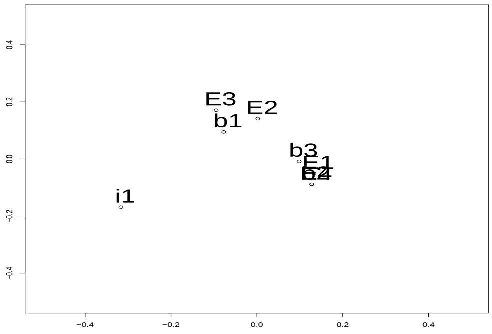


Figure 3: Intertranscriber discrepancy for Breaks among the different labellers. The Kappa indices of the table are projected in a 2D multidimensial scaling plot. E is expert, b is beginner and i is intermediate.

the use of symbols (like break index 2) with an apparently fuzzy definition leading to a scarce use in only rare situations.

The use of the visualization tools has shown to be useful to identify potential diverse tagging criteria. The 2D plots have shown to be useful to detect clusters of labellers as an evidence of possible different labelling criteria.

All these observations lead us to conclude that the merging of labels is a congruent procedure with practical advantages supported by the observation run on perceptual tests. The visualization tools permits to identify the closest symbols to be merged.

5. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “ToBI: A standard for labelling English prosody,” in *Proceedings of ICSLP-1992*, 1992, pp. 867–870.

[2] M. Ostendorf, P. Price, and S. Shattuck, “The boston university radio news corpus,” Boston University, Tech. Rep., 1995.

[3] A. K. Syrdal, J. Hirshberg, J. McGory, and M. Beckman, “Automatic ToBI prediction and alignment to speech manual labeling of prosody,” *Speech Communication*, no. 33, pp. 135–151, 2001.

[4] S. Ananthakrishnan and S. Narayanan, “Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 216–228, January 2008.

[5] V. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, “Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 797–811, May 2008.

[6] L. Aguilar, A. Bonafonte, F. Campillo, and D. Escudero, “Determining Intonational Boundaries from the Acoustic Signal,” in *Proceedings of Interspeech 2009*, 2009, pp. 2447–2450.

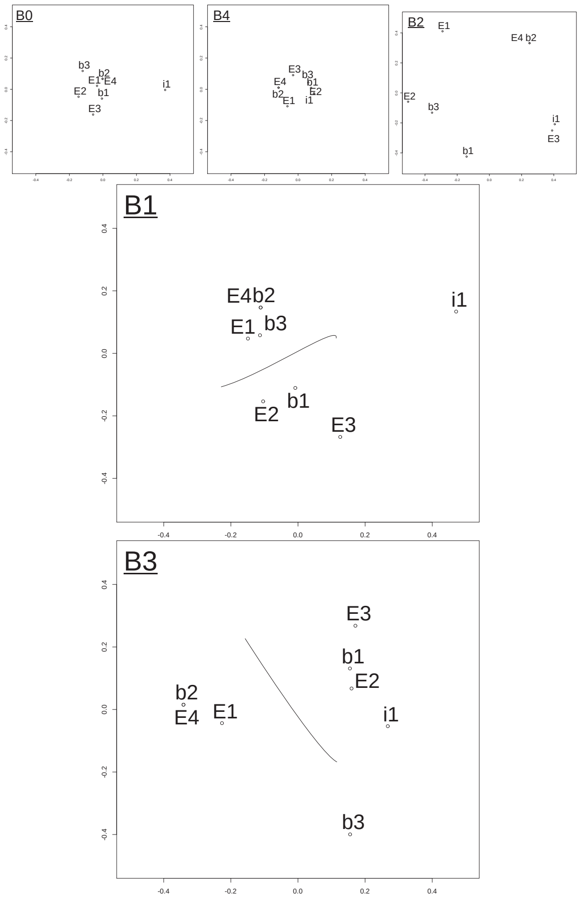


Figure 4: Intertranscriber discrepancy among the different labellers isolating the decisions for every type of Break

[7] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Perez, “Corpus and Voices for Catalan Speech Synthesis,” in *Proceedings of LREC 2008*, 2008.

[8] P. Prieto, L. Aguilar, I. Mascar, F. Torres, and M. Vanrell, “L’etiquetatge prosdic Cat-ToBI,” *Estudios de Fonetica Experimental*, vol. XVIII, pp. 287–309, 2009.

[9] J. Cohen, “A coefficient for agreement for nominal scales,” *Education and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[10] J. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[11] J. Landis and G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, pp. 159–174, 1977.

[12] T. Yoon, S. Chavarria, J. Cole, and M. Hasegawa, “Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI,” in *Proceedings of Interspeech 2004*, 2004.

[13] I. Borg and P. Groenen, *Modern Multimensional Scaling: theory and applications*. Springer-Verlag New York, 2005.

[14] M. Grice, R. Reyelt, R. Benzsuller, and A. Batliner, “Consistency in transcription and labelling of German intonation with GToBI,” in *Proceedings of ICSLP 1996*, 1996, pp. 1716–1719.

[15] S. Jun, S. Lee, K. Kim, and Y. Lee, “Labeler agreement in transcribing Korean intonation with K-ToBI,” in *Proceedings of ICSLP 2000*, 2000.

MFCC+F0 Extraction and Waveform Reconstruction using HNM: Preliminary Results in an HMM-based Synthesizer

D. Erro, I. Sainz, I. Saratxaga, E. Navas, I. Hernáez

AHOLAB Signal Processing Laboratory, University of the Basque Country, Bilbao (Spain)
derro@aholab.ehu.es

Abstract

The most widespread techniques for speech synthesis and voice conversion are currently based on probabilistic frameworks. Particularly, Hidden Markov Models (HMMs) play a relevant role in speech synthesis, whereas Gaussian Mixture Models (GMMs) are almost standard in voice conversion. Consequently, in both cases the performance of the systems is limited by three main factors: 1) the suitability of the statistical models; 2) the over-smoothing phenomenon; 3) the accuracy of the underlying speech parameterization and reconstruction method. This paper focuses on the third issue, still open at present: translating speech frames into parameter vectors with good properties for the mentioned statistical frameworks, and reconstructing waveforms properly. The proposal presented in this paper uses the Harmonics plus Noise Model (HNM) to extract MFCC+f_0 and reconstruct speech frames from them. The results of a perceptual evaluation show that the tool is valid for state-of-the-art HMM-based speech synthesis systems.

Index Terms: speech parameterization, statistical parametric speech synthesis, voice conversion, harmonics plus noise model

1. Introduction

Speech parameterization and reconstruction is a hot topic at present, mainly because of the great development of speech synthesis systems based on HMMs [1][2] and voice conversion systems based on GMMs [3][4][5][6]. These statistical frameworks require the input signals to be translated into tractable sets of vectors with good properties. Thus, Mel-frequency Cepstral Coefficients (MFCCs), which are known to work well in many areas of speech technologies, are also widely used for modeling spectra in synthesis and conversion systems [1][5]. Apart from their spectral modeling capability, one of their main advantages is that they allow using diagonal covariance matrices, since the individual components in each vector are highly uncorrelated. Other types of parameters such as Line Spectral Frequencies (LSFs) are often used in voice conversion [4][6]. Nevertheless, there is not a unique way of extracting parameter vectors from speech frames, and even less a unique reconstruction procedure. Vocoding is still an open topic for research, as both, parameter extraction from speech signals and speech reconstruction from parameters, have an immediate impact on the overall performance of the systems. This problem can be considered to be more important in speech synthesis than in voice conversion, where an original utterance of a source speaker is available (apart from the statistical models) and provides some information that can be used as a starting point. Therefore, this paper and the research work behind it have been focused especially on the former.

In the particular case of HMM-based speech synthesizers, many ways of parameterizing speech signals have been put into practice during the last fifteen years. In the basic implementation of HTS (the publicly available HMM-based Speech Synthesis System [7] based on HTK [8] and originally conceived at Nitech), the spectrum was modeled through Mel-frequency Cepstral Coefficients (MFCCs) obtained via Mel-generalized cepstral analysis [9], whereas a very simple pulse/noise excitation based on f_0 was used [10]. Subsequent improvements on that primary model consisted in using a more sophisticated mixed excitation [11][12]. Maia et al. [13] used an even more sophisticated trainable mixed excitation based on state-dependent filters for pulses and noise. In a recent work, Drugman et al. [14] used a two-band mixed excitation in which the upper band contained noise and the lower band was modeled through deterministic waveforms chosen via principal component analysis. In [15] and [16], a harmonics + noise decomposition of the signal itself (instead of the excitation) was used as a support for parameter extraction and waveform reconstruction. In both of them, the parameters used for training were based on linear prediction. Some other works focused on glottal source and vocal tract instead of spectrum and excitation [17][18][19]. Some attempts were also made to integrate the parameter extraction step into the statistical modeling step [20]. Probably, the most popular solution is the one based on Straight, a high-quality vocoder that decomposes signals into a spectral envelope (free of interferences from f_0) and an excitation given by f_0 and a so-called aperiodic envelope [21]. Straight's outputs are usually converted into adequate parameters such as MFCCs and band-aperiodicities [22]. However, it is worth mentioning that Straight is a proprietary software.

This paper presents a tool that extracts MFCC+f_0 from speech frames, and vice versa, assuming a Harmonics plus Noise Model for speech waveforms [23]. The tool has been specifically designed to be integrated into HTS. The implemented method has the following interesting properties:

- It allows extracting high-order MFCCs.
- It does not require excitation parameters other than f_0.
- It achieves considerably high perceptual quality in resynthesis.
- It allows several speech manipulations and modifications.
- The waveform reconstruction procedures can be implemented to be very efficient, which is helpful at synthesis time.

The perceptual tests performed to evaluate the tool in a speech synthesis application show that its performance is comparable to that of Straight, and thus can be used in state-of-the-art synthesizers. Moreover, we plan to make the tool freely available during the following months. The mentioned method is described in detail in Section 2, and the results of its preliminary evaluation are presented in section 3. Finally, Section 4 shows the conclusions of this paper.

2. Description of the method

The method is based on the decomposition of speech frames into a harmonic part and a stochastic part, which was proposed by Laroche et al. [24]. The harmonic component captures the locally periodic part of the signal that results from the vibration of the vocal folds. It is modeled through a set of harmonically related sinusoids. The stochastic component contains all the signal events that cannot be captured by the harmonic one, such as aspiration noise, bursts, etc. It is usually modeled as white Gaussian noise passing through a shaping filter.

$$s(t) = \sum_i A_i(t) \cdot \cos(2\pi i f_0(t) t + \varphi_i(t)) + e(t) \tag{1}$$

This mature speech model and its associated algorithms and methods [23] (a different implementation for operating under a constant frame rate, which is more appropriate for this task, can be found in [25]) provide a valid high-quality parameterization for speech analysis, modification and reconstruction. However, such a parameterization is hardly usable in a statistical framework for several reasons, being the most important ones the following [16]:

- The number of harmonics inside the analysis band is variable and depends on f_0.
- The resulting number of parameters is high (f_0 =100Hz means 50 harmonics between 0 and 5 kHz, each one given by its own amplitude and phase).
- The variability of the amplitudes and phases with respect to f_0 is extremely high.

Therefore, the model is not suitable for direct speech parameterization in the mentioned statistical frameworks, although it can be used as a support for extracting other types of parameters, as done in previous works [15][16]. The next subsections describe the proposed analysis and reconstruction procedures.

2.1. Parameter extraction

During the analysis step, given an input signal, the analysis frame rate, and the order of the parameterization, the system calculates one f0 value and one MFCC vector for each frame.

The first step of the analysis procedure is pitch detection. In this case, a modified version of the autocorrelation-based algorithm presented in [26] is used for extracting the local f_0 and determining whether the current frame is voiced or unvoiced. The modifications introduced into the original algorithm aim at increasing the estimation accuracy through a-posteriori local refinements using shorter analysis windows and considering the slopes of the complex amplitudes of the harmonics at low frequencies, as proposed in [24].

Voiced and unvoiced frames are treated in a different way to extract their MFCC representation. If the input frame has been classified as voiced by the pitch detector, a typical harmonic analysis (based on least squares optimization [23]) is performed on the full analysis band to get the log-amplitudes of the harmonics at multiple frequencies of f_0. Note that the amplitudes can be interpreted as discrete samples of the actual spectral envelope. Even at high frequencies (close to the Nyquist frequency), which carry noise-like signal components according to conventional HNM, the harmonic analysis is assumed to provide valid samples of the spectral envelope. Unvoiced frames are analyzed through a simple fast Fourier transform (FFT). Optionally, the resulting spectrum can be smoothed within certain bands. In order to homogenize both types of output, the envelope given by the harmonic amplitudes obtained for voiced frames is resampled at the FFT resolution via interpolation. Although past research shows that linear interpolation between log-amplitudes is accurate enough for some applications such as pitch modification [27][25], sinc-based interpolation is used here to increase the consistence of the analysis (see Figure 1 for details). As there is no reliable spectral information at frequencies below f_0, an extra artificial harmonic with the same amplitude as the fundamental one is added at 0 Hz before interpolating. A similar strategy was followed in [27] and gave good perceptual results. The resulting spectral envelopes should be very similar to those calculated by Straight [21] (see Figure 2), and therefore have the same potential advantages, mainly the fact that they allow estimating high-order MFCCs.

Next, the amplitude spectra are amplitude-normalized according to a multiplicative factor $f_0^{-½}$ (in unvoiced frames, f_0 is given the value f_s/L, where f_s is the sampling frequency and L is the analysis window length). This normalization is necessary to eliminate the dependency of the amplitude from f_0, which allows resynthesizing the signal at f_0 values other than the measured one. Note that two signals having the same energy and spectral envelope show harmonic amplitudes proportional to their pitch. The explanation is simple: for a given bandwidth, at higher f_0 the energy of the signal has to be supplied by fewer harmonics, so their amplitude has to be also higher.

During the last step of the analysis, cepstral coefficients are extracted from each amplitude spectrum as follows. First, the traditional cepstrum is obtained as the inverse Fourier transform of the log-amplitude spectrum, and then its dimension is reduced and the warping factor of the cepstral parameterization is transformed to match the Mel scale using the recursion described in [9]. Although other ways of calculating MFCCs from discrete points of the spectrum were also explored [28], informal tests consisting of visualizing the ripple of the MFCC curves at low frequencies led to the choice of the mentioned solution.

2.2. Speech waveform reconstruction

The first step consists of generating the noise part of the signal, which is present in both, voiced and unvoiced frames. The noise is obtained through inverse FFT after rebuilding the FFT spectrum from the MFCCs. The FFT module is obtained by sampling the MFCC envelope at a reasonable resolution (100 Hz), interpolating linearly to increase the resolution up to the one desired for the FFT, and de-normalizing by factor $(f_s/L)^{½}$, where L is now the FFT size. The phase is randomly generated following a uniform distribution in the range $[-\pi, \pi)$.

If the current frame is unvoiced, the synthetic frame is equal to the generated noise. Otherwise, the noise is high-pass filtered in the frequency domain (before the inverse FFT) according to a constant maximum voiced frequency (5 kHz is an adequate value, as reported in [25]). The fact that a constant-shape filter is used in voiced segments instead of an explicit modeling of the noise part is motivated by the good performance of such an HNM implementation in many applications [29]. Next, the harmonic component is generated as follows. The amplitudes of the harmonics are calculated by sampling the MFCC envelope and de-normalizing by factor $f_0^{½}$. Their phases are obtained through a minimum-phase approach [30]. Moreover, a linear-in-frequency phase term calculated from f_0 (for more details, see [16], for instance) is added at each frame in order to keep the phase relation between adjacent frames coherent. Apart from that, some artificial phase dispersion is included in the harmonics above 3.5 kHz in order to reduce the buzziness that may appear on the synthetic speech. It is worth mentioning that other types of phase manipulations based on all-pass filters were tried in order to increase the naturalness of the synthetic signals [31][32], but none of them produced better results than the described method according to listening tests.

The synthetic signal is reconstructed by overlap-add (OLA) using triangular windows. Thus, it can be expressed as:

$$\begin{aligned} s(kT+t) &= \tfrac{T-t}{T}\cdot s^{(k)}(t) + \tfrac{t}{T}\cdot s^{(k+1)}(t-T)\ ,\quad 0 \le t < T \\ s^{(k)}(t) &= \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(2\pi i f_0^{(k)} t + \varphi_i^{(k)}\right) + e^{(k)}(t) \end{aligned} \quad (2)$$

where $\{A_i^{(k)}\}$, $\{\varphi_i^{(k)}\}$, and $e^{(k)}(t)$ are the amplitudes, phases and noise at frame k, respectively, and T is the distance between frames.

3. Preliminary Evaluation

An open-source software toolkit named HMM-based speech synthesis system, HTS, has been publicly released since 2002 by the so called HTS working group, led by Nitech, to provide a research and development platform for the speech synthesis community [33]. During training, given a parametric representation of a number of speech signals and sets of labels describing their phonetic and prosodic context, HTS models the acoustic features of the different phonemes together with their duration using context-dependent HMMs (CD-HMMs). During synthesis, given the context labels of the signal to be generated, HTS creates a sentence-HMM by concatenating the corresponding CD-HMMs, and then generates the output waveform by inverse parameterization of the vector sequence whose likelihood with respect to the sentence-HMM is maximal.

The current HTS distribution includes demo scripts for training speaker-dependent and speaker-adaptive systems. The parameterization and reconstruction functions provided in the HTS demo are the traditional one, which uses MFCCs and a simple pulse/noise excitation, and the Straight-based one, currently used in state-of-the-art systems. In order to evaluate the method proposed in section 2, we built a synthesizer based on HTS and measured the naturalness of the synthetic utterances by means of a mean opinion score (MOS) test. Seven listeners were asked to listen to five different synthetic sentences for each of the three methods to be compared (namely, “Traditional”, “Straight” and “Proposed”) and rate them in a 1-to-5 MOS scale. The database used for this evaluation consisted of 2K short sentences (around 2 hours of speech) spoken by a Basque female speaker in neutral style. The features used for training were the following: f0 + 25 MFCCs for the traditional method, f0 + 40 MFCCs + 5 band-aperiodicities for Straight, and f0 + 40 MFCCs for the proposed method.

The MOS results shown in Figure 3 (at 95% confidence intervals) reveal that the performance of the parameterization method presented in this paper is significantly better than that of the traditional one. Straight still yields the best results, though the differences are much smaller in this case. We believe that one reason for this small gap is related to the explicit modeling of the aperiodic component. Informal experiments consisting of manipulating the Straight band-aperiodicities to match the shape of the HNM high-pass filter for noise (note that aperiodicity can be identified with the noise part of HNM) led to the conclusion that some important unvoiced information is lost under the current HNM implementation. These small differences were not perceived in resynthesized natural speech, probably because the inter-frame variability (not present in synthetic speech generated from statistical models) seems to compensate for the lack of a more sophisticated noise model. Future works will aim at studying other variants of HNM that assume a full-band noise component (such as [25]).

4. Conclusions

This paper has presented a method for extracting MFCCs and f_0 from speech and reconstructing the waveform from this parametric representation. The proposed method, which is based on the HNM, yields highly satisfactory results when compared to state-of-the-art techniques in a HMM-based speech synthesis application. Particularly, the preliminary results reported in this paper are not far from those of Straight-based parameterization, even without an explicit modeling of the aperiodic component. It is expected that further improvements on that part of the system will lead to even more promising results. A more formal evaluation with a higher number of listeners and synthetic voices will be carried out in future works.

5. Acknowledgements

This work has been partially supported by UPV/EHU (Ayuda de Especialización de Doctores), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and the Basque Government (Berbatek, IE09-262).

6. References

[1] H. Zen, K. Tokuda, A.W. Black, “Statistical parametric speech synthesis”, Speech Communication, vol.51, no.11, pp.1039-1064, 2009.

[2] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, S. Renals, “A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis”, IEEE Trans. Audio, Speech, & Language Processing, vol.17, no.6, pp.1208-1230, 2009.

[3] Y. Stylianou, O. Cappé, E. Moulines, “Continuous Probabilistic Transform for Voice Conversion”, IEEE Trans. Speech & Audio Processing, vol.6, no.2, pp.131-142, 1998.

[4] A. Kain, “High Resolution Voice Transformation”, Ph.D. thesis, OGI School of Science & Engineering at Oregon Health & Science University, 2001.

[5] T. Toda, A.W. Black, K. Tokuda, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory”, IEEE Trans. Audio, Speech & Language Processing, vol.15 , no.8, pp.2222-2235, 2007.

[6] D. Erro, A. Moreno, A. Bonafonte, “Voice Conversion Based on Weighted Frequency Warping”, IEEE Trans. Audio, Speech, & Language Processing, vol.18 , no.5, pp.922-931, 2010.

[7] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, “The HMM-based speech synthesis system version 2.0”, Proc. 6th ISCA Speech Synthesis Workshop, 2007.

[8] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, “The HTK Book Version 3.4”, Cambridge University Press, 2006.

[9] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, “Mel-generalized cepstral analysis - a unified approach to speech spectral estimation”, Proc. Int. Conf. Spoken Language Processing, vol.3, pp.1043-1046, 1994.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, Proc. Eurospeech, pp.2347–2350, 1999.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Mixed excitation for HMM-based speech synthesis”, Proc. Eurospeech, pp.2263–2266, 2001.

[12] X. Gonzalvo, J.C. Socoro, I. Iriondo, C. Monzo, E. Martinez, “Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish”, Proc. 6th ISCA Speech Synthesis Workshop, pp. 362–367, 2007.

[13] R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, “An excitation model for HMM-based speech synthesis based on

residual modeling", Proc. 6th ISCA Speech Synthesis Workshop, pp.131-136, 2007.

[14] T. Drugman, G. Wilfart, T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis", Proc. Interspeech, pp.1779-1782, 2009.

[15] C. Hemptinne, "Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system", Master thesis, IDIAP Research Institute, 2006.

[16] E. Banos, D. Erro, A. Bonafonte, A. Moreno, "Flexible harmonic/stochastic modeling for HMM-based speech synthesis", Proc. V Jornadas en Tecnologías del Habla, pp.145-148, 2008.

[17] J.P. Cabral, S. Renals, K. Richmond, J. Yamagishi, "Glottal Spectral Separation for Parametric Speech Synthesis", Proc. Interspeech, pp.1829-1832, 2008.

[18] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku, "HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering", IEEE Trans. Audio, Speech, & Language Processing, 2010 (in press).

[19] P. Lanchantin, G. Degottex, X.Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method", Proc. ICASSP, pp.4630-4633, 2010.

[20] T. Toda, K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM", Proc. ICASSP, pp.3925–3928, 2008.

[21] Hideki Kawahara, "Straight, exploration of the other aspect of Vocoder: perceptually isomorphic decomposition of speech sounds", Acoustic Science and Technology, vol.27, no.6, pp.349-353, 2006.

[22] H. Zen, T. Toda, M. Nakamura, K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005", IEICE Trans. Inf. Syst. E90-D (1), pp.325–333, 2007.

[23] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification", PhD thesis, École Nationale Supèrieure des Télécommunications, Paris, 1996.

[24] J. Laroche, Y. Stylianou, E. Moulines, "HNM: a simple, efficient harmonic+noise model for speech", IEEE Workshop on Apps. Signal Proc. to Audio & Acoustics, pp.169-172, 1993.

[25] D. Erro, A. Moreno, A. Bonafonte, "Flexible harmonic/ stochastic speech synthesis", Proc. 6th ISCA Speech Synthesis Workshop, pp.194-199, 2007.

[26] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", Proc. of the Institute of Phonetic Sciences, University of Amsterdam, vol.17, pp.97–110, 1993.

[27] E.R. Banga, C. García-Mateo, X. Fernández-Salgado, "Concatenative Text-to-Speech Synthesis based on Sinusoidal Modeling", chapter in "Improvements in Speech Synthesis", John Wiley and Sons, pp.52-63, 2001.

[28] O. Cappé, J. Laroche, E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points", IEEE Workshop on Apps. Signal Proc. to Audio & Acoustics, pp.213-216, 1995.

[29] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE Trans. Audio, Speech, & Language Processing, vol.9, no.1, pp.21-29, 2001.

[30] R. McAulay and T. Quatieri, "Sinusoidal Coding", chapter in "Speech Coding and Synthesis", Elsevier, pp.121-173, 1995.

[31] S. Ahmadi, A.S. Spanias, "Low bit-rate speech coding based on an improved sinusoidal model", Speech Communication, vol.34, pp.369-390, 2001.

[32] X. Sun, F. Plante, B.M.G. Cheetham, K.W.T. Wong, "Phase modelling of speech excitation for low bit-rate sinusoidal transform coding", Proc. ICASSP, vol.3, pp.1691-1694, 1997.

[33] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

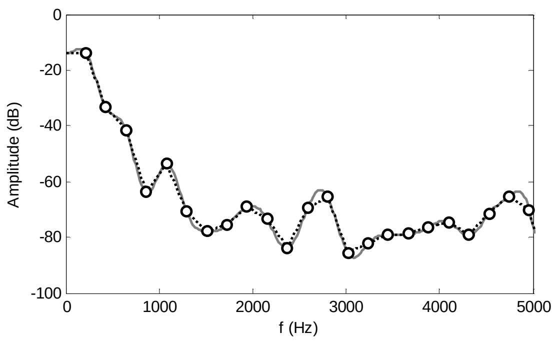


Figure 1: *Sinc-based interpolation (solid line) vs. linear interpolation (dotted line) between amplitudes.*

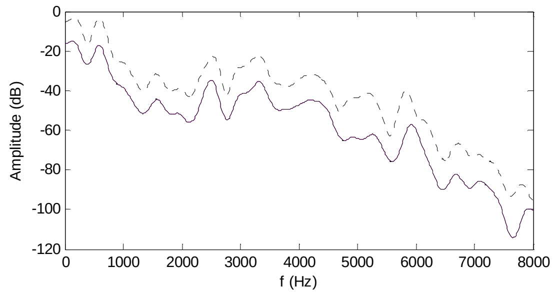


Figure 2: *Spectrum given by the proposed method (solid line) and Straight spectrum (dotted line) at a voiced frame.*

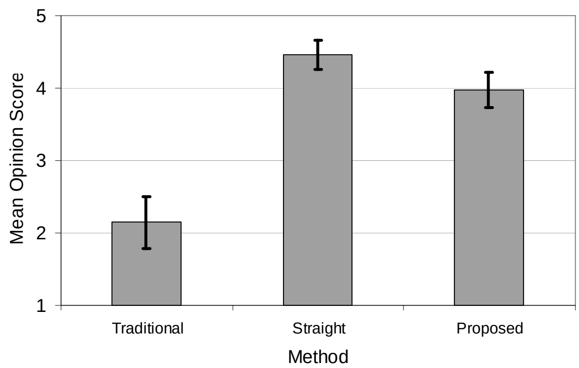


Figure 3: *Results of the MOS test at 95% confidence intervals.*

Articulatory Characteristics of European Portuguese Laterals: a 2D & 3D MRI Study

Paula Martins[1], Catarina Oliveira[1], Augusto Silva[2], António Teixeira[2]

[1]Escola Superior de Saúde/IEETA, University of Aveiro, Aveiro, Portugal
[2]Dep. Electrónica Telec. Informática/IEETA, University of Aveiro, Aveiro,Portugal
pmartins@ua.pt, coliveira@ua.pt, augusto.silva@ua.pt, ajst@ua.pt

Abstract

Magnetic Resonance Imaging (2D and 3D) has been acquired in seven EP native speakers during the production of EP lateral sounds /l, L/, with the purpose of obtaining articulatory data and acquiring further knowledge regarding the production of the lateral consonants. To attain these goals the acquired MRI corpus includes the /l/ sound in different syllabic positions (onset, coda and intervocalic) in the context of the three EP cardinal vowels, and the palatal lateral /L/ in intervocalic context. The results indicate the existence of a high degree of inter-subject variability (more evident for /l/). /l/ velarization can take place in all syllabic positions, including in onset, which is probably a speaker-dependent characteristic. A detailed articulatory description of the palatal lateral is also provided.

Index Terms: Speech Production, European Portuguese, Laterals, MRI, Image processing

1. Introduction

1.1. European Portuguese Laterals

Phonologically European Portuguese (EP) has two lateral consonants: /l/ and /L/. The production of /l/ involves a linguo-alveolar contact and one or two lateral channels along the lateral sides of the tongue [1, 2]; the /L/ is produced by one movement of the front of the tongue against the alveolo-palatal zone.

For the last two decades, it has been argued that Portuguese /l/ is categorically associated with a non-velarized ('light') allophone, which typically occurs syllable initially, and a velarized ('dark') one in coda position. However, there are conflicting positions in the literature regarding this light-dark dichotomy.

Unfortunately, the majority of the descriptions was based on gross impressionistic observations, and the only extensive empirical descriptions of Portuguese /l/ come from the acoustic data of Andrade [3]. Her results show that /l/ velarization does take place in onset position, for speakers of the Lisbon variety of EP, although the degree in which it is manifested varies across individual subjects. In agreement with this acoustic data, Recasens and Espinosa [4] stated that EP, together with Russian and Leeds British English, belong to a group of sound systems where /l/ presents essentially the same dark realization word-initially and word-finally.

Previous MRI studies of EP [5] seem to confirm the theory of the existence of a single dark-l for EP. However, the corpus was not exclusively designed for the study of laterals. More recently, in a study based in EMA [6], some syllable position effects were found: coordination patterns for syllable-initial /l/ are distinct from those observed for syllable-final /l/; the tongue dorsum is more retracted in the syllable-final /l/; /l/s in coda revealed, for one of the speakers, a reduction in magnitude of the tongue tip gesture.

There are no extensive and up-to-date data on the articulatory characteristics of /L/. It is traditionally recognized as a palatal lateral, but according to MRI data [5], the articulation appears to happen a little further forward.

1.2. MRI studies with laterals

MRI articulatory data for American English confirmed the existence of some differences between dark and light /l/. Although, both allophones present a similar overall configuration of the tongue body, with alveolar contact, lateral compression and a convex tongue body, there are differences between the two allophones: greater linguo-alveolar contact in light /l/ and a more retracted tongue body in the production of dark /l/ with repercussions in the area functions. The dark /l/ (when compared with the light /l/) presents large areas in the palatal region (behind the constriction), less lateral contacts and small areas in the pharyngeal and velar regions [1].

More recently, [2] also based on MRI data, reported a relatively shorter linguo-alveolar contact for dark /l/(0.8 cm) than for light /l/ (1.7 cm) and shorter lateral channels for the dark /l/. The two lateral channels formed for the light /l/ are asymmetrical (4.9 cm longer on the right and 2.1 cm on the left). A separate supra-lingual space was also found.

Section 1 introduces the problem and reviews related work on laterals. In section 2, the corpus, speakers, and MRI acquisition setup are described. Section 3 describes image segmentation and image processing techniques used to extract information from MRI images. The results, including preliminary 2D and 3D data, are given in section 4. This section is followed by the discussion and the conclusions that can be withdrawn from the present study.

2. Method

2.1. Corpus

MRI corpus included the lateral-alveolar /l/ in different positions: word-initial position (e.g laca "hairspray" [lak6]), intervocalic position (e.g sala "room" [sal6]) and coda position (e.g sal "salt"[sal]) in the context of the three EP cardinal vowels [i,a,u] (in stressed position). For the lateral /L/, due to phonotactic constraints in Portuguese, only intervocalic context was considered (e.g palha "straw" [paL6]).

2.2. Speakers

MRI data was acquired with seven (3 female, 4 male) speakers: six native mono-lingual EP speakers and a bilingual male speaker (EP and Spanish), ages ranging from 21 to 39, with no history of hearing or speech disorders. The speakers were all volunteers, from the midland of the country. Only two of the speakers (CO and JH) had Phonetic or Linguistic knowledge, the other speakers were instructed by the research team and had the opportunity of having a training session before the MRI acquisition session. An MRI screening form and informed consent were obtained before their participation in the study.

2.3. MRI Acquisition

The MRI experiment was carried out in a Magnetic Resonance Imaging Unit at Coimbra (IBILI).The images were acquired using a 3.0 T MR scanner (Magneton Tim Trio, Siemens, Erlanger, Germany) equipped with high performance gradients (Gmax = 45mT/m, rise time= 0.2s, slew Rate= 200 T/m/s; and FOV =50 cm). A standard 12-channel Head and Neck phased-array coils and parallel imaging (GRAPPA) were used in all data acquisition sessions. The Imaging protocol used in the present study was based in a previous MRI study conducted by our research team [5]. The subjects were positioned comfortably in a supine position using headphones. After acquiring reference images, a T1 W 5 mm thickness midsagittal MRI slice of the vocal tract was obtained using a TSE sequence (TR/TE/FA=400 ms/7.8 ms/120), FOV=240x240 mm; matrix (256x256) resulting in a pixel size of (0.938, 0.938). The acquisition time was 6 seconds. After that, a volume covering the entire vocal tract was obtained in the sagittal plane with a T1W 3D Spoiled GE sequence (VIBE), resulting in an acquisition time of 19 seconds; matrix (224x256); voxel size (1.055, 1.055, 2). The speakers sustained the sound during the period of acquisition; the sequence was launched when the /l/ was produced (e.g. sallllllll). Finally, a 3D high resolution sequence (VIBE) in the axial plane was obtained for each of the speakers, without phonation, to allow the extraction and co-registration of the mandible and dental casts.

3. Image Processing

3.1. Image processing techniques

Image segmentation was performed using two open source image processing tools: MevisLab [7] and ITK-Snap toolkit [8]. Some routines and data analysis were performed in Matlab. The codes used were implemented by one of the authors of the paper.

2D contours were extracted from midsagittal MR images using a semi-automatic technique (Live wire routine implemented in Mevislab). The tongue images were segmented using the same technique (e.g. fig. 4) and also using another semi-automatic tool based on a level sets framework (ITK-Snap, e.g.fig. 1).

Segmentation of the vocal tract images was more demanding and time consuming than the tongue. The overall process involves the following steps: 1) Mandible/maxilla segmentation and extraction from an MRI volume acquired for each speaker 2) Mandible and maxilla masks were resampled and co-registered with vocal tract data sets 3) Curved Multiplanar Reconstruction (MPR) of the vocal tract was performed to obtain areas perpendicular to the midline of the tract 4) Segmentation of the re-sliced volume was performed (every 5 mm from the glottis to the velum and every 3 mm throughout the oral cavity). This process was established using different pipelines implemented in MevisLab. Live wire technique was the method used in segmentation. After that, contour lists (CSO) were ex-

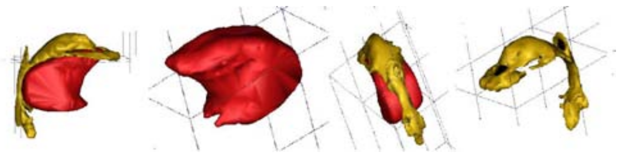


Figure 1: From left to right: ITK-Snap visualizations of the vocal tract and tongue (lateral view), tongue (oblique view), vocal tract and tongue (posterior-oblique view) and vocal tract (anterior-oblique view), during the production of /l/.

ported to Matlab allowing the extraction of area functions and tract visualizations. The inclusion of the teeth is a fundamental step when the goal is the modelation of the small lateral channels established in the production of the lateral sounds.

4. Results

In this section we present 2D MRI data for seven speakers and preliminary 3D information for 2 speakers (CO, female and JPM, male). Place of articulation, tongue configuration (2D and 3D), contextual variability, 3D visualization of the tract and area functions are exploited.

4.1. The /l/

4.1.1. 2D data

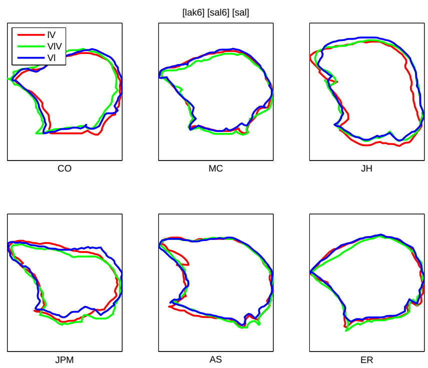


Figure 2: Superimposition of 2D contour for /l/ (6 speakers) in the three syllabic positions (onset, intervocalic and coda) with [a] in stressed position.

Analysis of the midsagittal images with contour superimposition revealed that the [l] was produced with contact of the tongue blade (i.e. laminal articulation) or tongue tip (i.e. apical articulation) against the dento-alveolar region. Laminal articulation was more frequent than apical articulation. Some speakers produced a laminal articulation in onset and a more apical articulation in coda. Some inter-speaker variability was observed regarding tongue body configuration, behind the alveolar contact. The tongue body shape in the three syllabic positions (onset, intervocalic and coda) was quite similar for most of the speakers. Only JPM and JH exhibited more variability (see fig. 2).

When considering the effect of vowel context in tongue configuration for [l] (fig. 3), it was observed that, for some speakers, vowel context has relatively little effect on tongue configuration. The variability appeared to be greater in speakers

JPM, AS and CO.

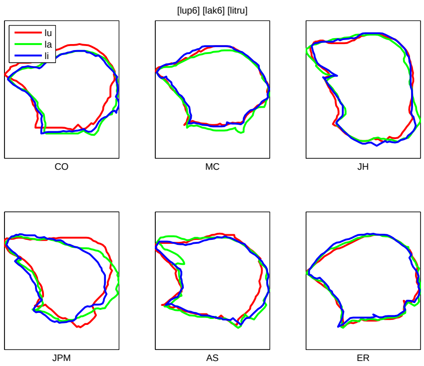


Figure 3: 2D contour superimposition for onset /l/ in vowel context [i, a, u].

4.1.2. 3D data

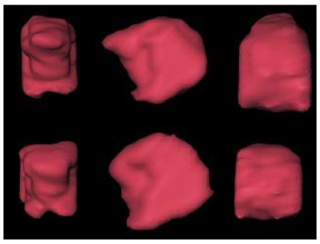


Figure 4: 3D rendering of the tongue (speaker CO) for onset /l/ in the word [lak6] (top) and vowel [a](bottom). From left to right: anterior, lateral and posterior views. Live wire segmentation implemented in MevisLab.

Analysis of 3D information for two of the speakers (JPM and CO) allowed the following considerations: Although the midsagittal contours of the /l/ are, in general, very similar to those of the vowels, 3D data revealed some differences (fig. 4) in tongue body configuration between the lateral and the vowels. The /l/ presents inward-lateral compression and a convex shape of the posterior tongue body; the tongue tip is raised towards the alveolar ridge. On the contrary, the tongue body for the vowel ([a]) has a slightly concave and spread configuration.

Lateral channels (fig. 5) formed around the tongue sides are longer for onset [l] than for coda /l/. For the speaker CO, lateral passages are short and start slightly posterior and through the sides of the alveolar contact, both in onset and coda position. For JPM, lateral channels are longer in onset than in coda: starting back at the velo-palatal area, extending through the oral cavity and continuing along the sides of the alveolar contact. The calculated area for the lateral passages is always less than $0.5cm^2$ for both speakers. The length of the alveolar contact is higher in onset than in coda.

Analysis of JPMs area functions (fig. 6) shows that the main differences between onset and coda /l/ occur at velar and upper pharyngeal regions. At this level the areas for coda /l/ are always lower than those obtained in onset. This information reflects a retracted tongue and tongue back elevation towards the velum, in coda position. For the speaker (CO), the only difference between the two syllabic positions was a slightly lowering of the areas at the pharyngeal region, in coda.

4.2. The /L/

4.2.1. 2D data

In /L/, the contact is made between the tongue blade and predorsum against the alveolo-palatal region for most of the speakers. Only one speaker (ER) articulated the /L/ exclusively at the palatal region. Figure 7 shows a superimposition of the contours for the intervocalic /L/ in the context of vowels /a, i, u/. Mostly, the consonant revealed to be only slightly influenced by the vowel: when observed (speakers JPM, AS, LCR), the variability occurred always behind the contact, being more evident at the tongue root level, an area not involved in the articulation of the lateral.

4.2.2. 3D data

For 3D data, the most relevant findings are related with a quite extensive contact area. Speaker CO presents a higher contact length (2.4 cm) than speaker JPM (2.1cm), which became more evident due to vocal tract length differences (about 3.5 cm shorter for CO than JPM).

Lateral channels are long and can be observed for both speakers. They are relatively symmetric, starting behind the back molars until and throughout the alveolo-palatal contact. Lateral channel areas are higher than those obtained for the lateral-alveolar /l/.

Speaker CO presents large areas at the pharyngeal region (due to a more anterior tongue position), which start to decrease from the velum towards the constriction. The other speaker shows relatively small areas at the low pharyngeal level, reaching the maximum area at the velum, and then following a pattern similar to that observed for CO.

As showed in fig. 8, tongue shape is characterized by an inward compression towards the midline, a convex shape and a lowered tongue tip position, for both speakers.

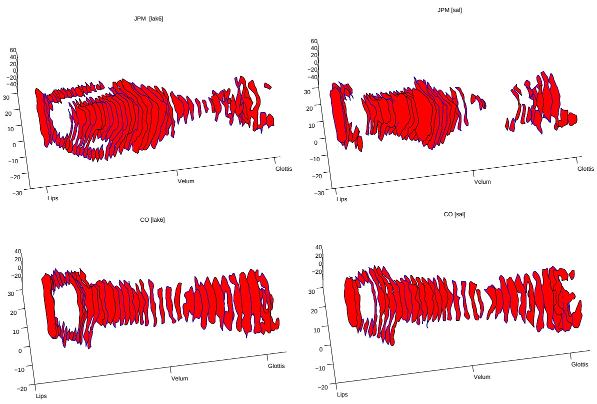


Figure 5: Vocal tract areas and lateral passages (3D data) for speakers JPM (top) and CO (bottom). At left, onset/l/ ([lak6]), and at right coda /l/ ([sal]).

5. Conclusions

This paper investigated articulatory characteristics of EP laterals /l/ an /L/, from a large dataset of MRI images. New 3D data allowed an articulatory description, representing a fundamental step towards the modelation of the lateral sounds.

MRI images revealed the existence of some inter-speaker variability in the production of these consonants, namely in the

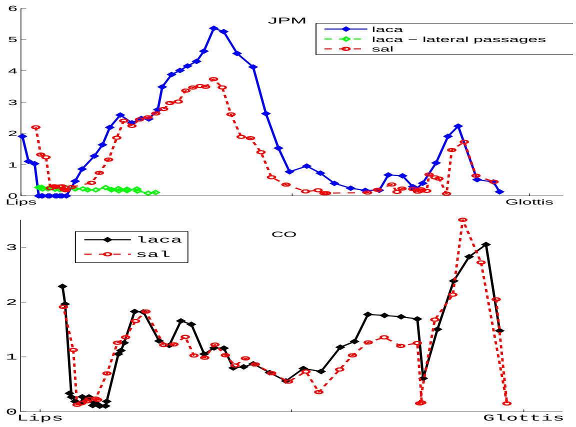


Figure 6: Vocal tract area functions of /l/ in onset (line) and coda (dashed) for JPM (top) and CO (bottom).

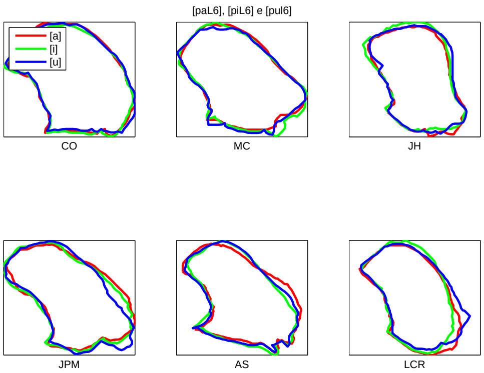


Figure 7: 2D contour superimposition for /L/ in vowel context [i, a, u].

alveolar lateral. Common characteristics for /l/ are: linguo-alveolar contact, either laminal or apical; inward-lateral compression and convex shape of the posterior region of the tongue [1] (though more evident in JPM than in CO), and lateral channels alongside the tongue.

Lateral channels are shorter in coda than in onset, as well as the length of the linguo-alveolar contact.These results are in line with previous studies for American English [1, 2].

Speaker CO presented short lateral channels and small alveolar contact in both syllabic positions. On the other hand, JPM showed extensive lateral channels in onset and short ones in coda. These findings pointed out to a dark realization for CO [3], in both syllabic positions, and two different allophones for JPM.

The palatal lateral is articulated at the alveolo-palatal region and not exclusively at the palatal area. This is in agreement with what has been reported for other Romance languages, as emphasized by [9]. This consonant is also characterized by an extensive contact area, lowered tongue tip, inward compression towards the midline and posterior convex shape resulting in long lateral channels and large pharyngeal and/or velar areas (depending on the speaker).

One of the limitations of this study has to do with difficulties in the co-registration and segmentation of the dental casts. Due to this fact we may assume that the areas of the lateral

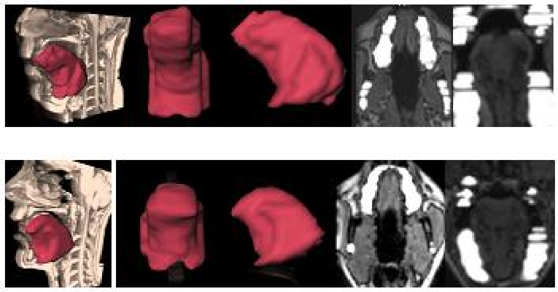


Figure 8: 3D data from JPM (top) and CO (bottom) for /L/ ([paL6]). From left to right: 3D rendering of the tongue (left oblique view) inside the volume, frontal view, lateral view, axial slice (after curved MPR) shows the pattern of the lingual contact and coronal reformatation.

channels were not precisely estimated.

Next steps for complementing this work will include: the segmentation and analysis of 3D data for all the speakers in all contexts, the improvement of segmentation tools (to improve accuracy and efficiency, particularly for vocal tract analysis) and the validation of the methods used to obtain information (e.g mandible and maxilla co-registration on the volumes).

6. Acknowledgements

This research was supported by FCT (Portuguese Research Agency) PhD grant of the first author (SFRH/BD/65183/2009). The authors would like to thank Professor Miguel Castelo-Branco (IBILI Director) and the technical staff of the MRI unit. We are also very grateful to our speakers.

7. References

[1] S. Narayanan, A. Alwan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. part I. The laterals," *Journal of the Acoustical Society of America*, vol. 101, pp. 1064–2007, 1997.

[2] X. Zhou, "An mri-based articulatory and acoustic study of american english liquid souns /r/ and /l/." Ph.D. dissertation, Faculty of the Graduate School, University of Maryland, Maryland, EUA, 2009.

[3] A. Andrade, "On /l/ velarization in European Portuguese," in *International Congress of Phonetic Sciences (ICPhS)*, San Francisco, August 1999, pp. 543–546.

[4] D. Recasens and A. Espinosa, "Articulatory, positional and coarticulatory characteristics for clear /l/ and dark /l/: evidence from two Catalan dialects," *Journal of the International Phonetic Association*, vol. 35, no. 1, pp. 1–25, 2005.

[5] P. Martins, I. Carbone, A. Pinto, A. Silva, and A. Teixeira, "European Portuguese MRI based speech production studies," *Speech Communication*, vol. 50, no. 11-12, pp. 925–952, 2008.

[6] C. Oliveira, A. Teixeira, and P. Martins, "Towards an articulatory characterization of european portuguese /l/," in *ISCA Workshop on Experimental Linguistics*, Athens, Greece, 2010.

[7] M. S. MeVis, "Mevislab (medical image processing and visualization) version 2.1," 2003-2010.

[8] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3d active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[9] D. Recasens and A. Espinosa, "Articulatory, positional and contextual characteristics of palatal consonants: Evidence from majorcan Catalan," *Journal of Phonetics*, vol. 34, pp. 295–318, 2006.

Poster Session 1

Information Extraction from Portuguese Hospital Discharge Letters

Liliana Ferreira[1], *António Teixeira*[1], *João Paulo da Silva Cunha* [1]

[1]Institute of Electronics and Telematics Engineering of Aveiro
Department of Electronics, Telecommunications and Informatics
University of Aveiro, Portugal
{lsferreira, ajst, jcunha}@ua.pt

Abstract

In this paper we describe MedAlert, a system which automatically extracts information from free-text discharge summaries written in Portuguese. We introduce a corpus of 915 hypertension related discharge letters and the method used to create the discharge letters representation model. MedAlert is based on an open-source framework and its components use natural language processing principles to discover elements of the knowledge model. We evaluate MedAlert precision using a set of 10 discharge letters from the MedAlert corpus from which 339 named entities were recognized. MedAlert achieves an entity recognition precision of 1 for entities such as anatomical sites, evolutions and dates and 0.93-0.99 for conditions, findings and therapeutics. A precision value of 0.69 is reported for examination entities due to to the recognition of active substances, such as insulin, as laboratory examinations.
Index Terms: information extraction, medical language processing, medical knowledge representation

1. Introduction

In order to perform research and to improve standards of health care it is required the access to a variety of data sources. The knowledge contained in unstructured textual documents like clinical notes and discharge summaries is critical to achieve these goals.

To bridge the gap between free-text and structured information, an automatic and highly accurate mapping of free-text reports onto a structured representation is required. Natural Language Processing (NLP) systems can retrieve named entities such as diseases and anatomical sites, and may be able to provide links between them in case of a relationship.

In this paper we describe the method used to create a structured representation of the discharge summaries and we describe our system, MedAlert, which automatically recognizes the entities mentioned in the reports. We also report on the performance of MedAlert, measured on a set of discharge summaries of patients admitted with hypertension related disorders.

This system has great clinical value: it can increase the classification of patients for practice management (*how many patients with cerebral hemorrhage were received*), for research (*how many patients used a specific of drug with success*), quality control (*how many patients with cerebral hemorrhage were received and which were their outcome*). In addiction, it would allow physicians to continue to practice using their current descriptive language in free-text reports without a requirement to enter structured data in a complex, time consuming computer-based system.

This paper is organized as follows: Section 1.1 presents some related work in the area of medical language processing systems and its resources. Section 2 discusses the terminologies used in the system. In particular, Section 2.1 provides details about the corpus and the manual annotation process while Section 2.2 presents the knowledge sources used. The architecture of MedALERT, the Medical Language Processing System, is the focus of Section 3. Section 4 presents the evaluation results of entity recognition from the set of free-text reports. We conclude in Section 5 with some future work directions.

1.1. Related work

The goal of information extraction (IE) is to extract structured and semantically well defined concepts from unstructured data sources to facilitate access and retrieval of information [1]. In the clinical domain, information extraction has the potential to help clinicians rapidly answer questions such as *How many patients were diagnosed in 2007 with cerebrovascular diseases?*, *What percentage of these patients had also hypertension?* There are multiple approaches to building IE systems. In general, such systems have NLP components such as tokenizers, part-of-speech taggers and parsers. Two separate frameworks for building information extraction systems were developed and made available as open-source components. One is the Generalized Architecture for Text Engineering (GATE) [3] and the other is the Unstructured Information Management Architecture (UIMA) [2]. Some of the state-of-the-art IE systems in the biomedical and clinical domain are presented in [4] and [5]. The caTIES system [6] extracts several types of named entities (NE) such as histology, anatomical site, size and grade and is based on the GATE framework. No results have yet been published. MedLEE, another clinical natural language processing IE system, extracts domain knowledge from a variety of unstructured reports, such as discharge summaries, radiology reports and pathology reports [7]. MedLEE focus on extracting named entities but there seems to be no results published for extracting information from pathology reports using MedLEE. MedTAS/P, Medical Text Analysis System/Pathology, is a system for the automatic conversion of unstructured pathology reports into a structured and codified knowledge source according to a Cancer Disease Knowledge Representation Model. It is part of a clinical NLP-based system as described in [1]. MedTAS/P achieves F1-scores of 0.97-1.0 for instantiating classes in the knowledge representation model such as histologies or anatomical sites. To the best of our knowledge there as been no effort to develop a system for the extraction of information from clinical reports written in Portuguese and in Portuguese hospitals.

2. Resources

2.1. MedAlert corpus

The biomedical community hasn't yet, to our knowledge, developed a gold-standard training and test corpus of annotated clinical reports written in English which can be used as a shared standard for evaluating automatic knowledge extraction system. The same is valid when referring to Portuguese reports. Therefore, the development of detailed manually annotated corpus for training, validating and evaluating the system is of extreme importance.

For the training, validation and evaluation of the MedAlert system we created a corpus of 915 free text discharge letters written in Portuguese from the Infante D. Pedro Hospital in Aveiro, Portugal. These discharge letters refer to patients admitted with hypertension related problems.

These discharge summaries contain the information added by the clinician during the period the patient was admitted in the hospital. They are divided into 6 sections, each containing information about the patient admission motive, clinical history, physical examination, evolution, therapeutics applied during admission and destiny recommended to the patient after discharge, respectively. The documents do not have any personal data and therefore it is not possible to identify the patient through the analysis of the discharge letters. Table 1 shows the distribution of information in the corpus, in particular the amount of texts, sentences and tokens for each structure.

Table 1: Corpus MedAlert

	Tokens	Sentences	Texts
Admission motive	1 989	225	162
Clinical history	22 386	1 066	185
Physical Examination	7 865	711	134
Evolution	8 998	506	154
Therapeutic	6 299	219	159
Destiny	4 158	262	120
Total	51 695	2 989	914

Given the expense of human annotation, the gold standard set has to be a relatively small subset of the whole corpus of 915 documents. Therefore, to create the gold standard corpus we used a random sample of approximately 10% of the corpus, making a total of 90 documents.

In order to ensure consistency, critical to the quality of the gold standard, it is important that all documents are annotated to the same standard. This is accomplished by a set of guidelines describing in detail what should and should not be annotated and other important considerations such as how to decide if two entities are related and how to deal with co-reference. This guidelines also provide a sequence of steps, a recipe, which annotators should follow when working on a document. The guidelines were developed through a rigorous iterative process by a small team of computational linguistics and clinicians. They were tested against a significant number of documents before the use on the final gold standard.

The resulting representation model, the MedAlert Discharge Letters Representation Model (MDLRM) was implemented within Knowtator [8], a Protégé [9] plugin. The gold standard set is currently being manually annotated by a linguistic and a clinician. The annotators manually fill in the attributes and relations in the classes of the representation model with information from the reports using the Knowtator tool.

Table 2 presents the entities defined in the MDLRM which were used in the entity recognition task presented in this paper.

Table 2: MedAlert Entities

Classes	Description
Condition	Complications, conditions and other problems manifested by a patient;
Anatomical Site	Anatomical structure or location, normally the locus of a *Condition*;
Evolution	The clinical evolution of the patient or *Condition* after a given *Therapeutics*;
Examination	Interaction between doctor and patient or *Anatomical Site* with the purpose of measuring or studying some aspect of a *Condition*;
Finding	The numeric or qualitative finding of an *Examination*, excluding *Condition*;
Location	Geographically defined location, normally where an *Examination* or *Therapeutic* is performed;
Therapeutic	Action performed by a clinician targeted at a patient, *Anatomical Site* or *Condition* with the purpose of changing or treating a *Condition*;
DateTime	Temporal expressions, including dates and times (absolute or relative), duration and frequencies;
Value	Absolute or relative quantifications or classifications;

2.1.1. *Development and evaluation sets*

For the purpose of development we used the set of documents not used in the manual annotation task, ie, the 825 documents of the corpus which do not belong to the gold standard set were used to develop the system. The entity recognition task described in this paper is evaluated on regards to precision against a randomly selected set of 10 discharge letters belonging to the set gold standard set.

2.2. MedAlert ontologies

In this project we used two codified terminologies and a drugs ontology as the underlying terminologies for *Conditions*, *Examinations*, *Anatomical Sites* and *Therapeutics* entities recognition. Namely, we used the International Classification of Diseases - Ninth Revision, Clinical Modification (ICD-9-CM) and the Unified Medical Language System (UMLS) [10] as codified terminologies. The drugs ontology was created in a semi-automatic way using the information provided by INFARMED, the National Authority of Medicines and Health Products and by the Hospital regarding the medication used and available in the institution. It has a total of 5 228 instances of Medication and Active Substances (and the corresponding relationships between these).

These terminologies, particularly the UMLS semantic network, allows the use of specified vocabularies in entity recognition avoiding the manual creation of lists and gazetteers. This resource and its use in MedAlert is described in more detail in Section 2.2.1. The ICD9 terminology is intended to be used in a future task of automatically assigning codes to the health conditions described in the discharge summaries.

2.2.1. UMLS

The UMLS [10] is a compendium of many controlled vocabularies in biomedical sciences. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems. It may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts. UMLS main purpose is to facilitate the development of computer systems that behave as if they 'understand' the meaning of the language of biomedicine and health.

In order to develop the system presented in this paper we used the 2009 UMLS version and particularly the Portuguese translation of MeSH (the National Library of Medicine's controlled vocabulary thesaurus, a sets of terms naming descriptors in a hierarchical structure that allows search at various levels of specificity.), the DeCS.

3. System Architecture

MedAlert, the medical language processing system, is based on natural language processing principles and contains both rule-based and machine-learning based components and runs within UIMA framework. An application within such framework consists of a set of programs (annotators), each having a configuration file in XML format being the execution sequence, or pipeline, of annotators also described in a configuration file. Annotators mark up an unstructured textual document, inserting 'annotations' that can be associated with a particular piece of text or which can contain objects for other annotations. A subsequent annotator can read and process all previously created annotations. MedALERT, provides a mechanism to use external resources, such as terminologies and ontologies.

The system pipeline can be broken into several components:

1. Ingestion - a component which reads the patient discharge letters XML files and converts them into plain text while keeping information about the document's structure. Also this component is responsible for reading the knowledge sources and converting them into the MedAlert type system.
2. General natural language processing - component for sentence discovery, tokenization, part-of-speech tagging and shallow parsing. This component contains an abbreviation sensitive sentence splitter.
3. Named entity recognition - this component identifies the concepts defined in the MDLRM based on specified terminology. It also determines negation, lateralization and other modifiers of an entity.

3.1. Ingestion

The document ingestion annotator converts embedded tags of an input document or set of documents into annotations and simultaneously adds information as the sections of the document and header information containing the episode number and the codes assign to the episode. The terminologies used are also read in this component and added to UIMA type systems in order to be available in the annotation as the documents are analyzed.

3.2. General Natural Language processing

This component starts by identifying the abbreviations contained in the document. This step is critical for a correct determination of sentence boundaries as for abbreviations containing a period, the process of identifying them involves solving the sentence boundary problem as well (e.g. Dr. should not be considered the end of sentence). We developed a method for detecting abbreviations in clinical notes which is a heuristic rule-based program developed by observing several discharge letters. It utilizes information concerning word formation, such as capital letters, numeric and alphabetic characters and their combinations, together with the a list of Portuguese words created from a general corpus of Portuguese, the corpus developed to be used in the second evaluation contest for named entity recognition in Portuguese, the HAREM II [13] (with 784 119 words). In particular, if a word contains special characters such as "-" and ".", has less than 6 characters, is lower case and is not in the Portuguese list it is considered an abbreviation.

Part of this component are also a tokenizer and a part-of-speech tagger based on the popular Tree Tagger [11]. The component also contains a context tokenizer which is a regular expression annotator that in conjunction with short lists discovers textual mentions describing dimensions and sizes, dates, number, etc.

3.3. Named entity recognition

The named entity recognition (NER) component is one of the most important components of the system. This component makes use of the terminologies to recognize the entities belonging to the knowledge model. As exact string match is not sufficient to recognize entities in unstructured text, the NER annotators have several characteristics that help the specification of the tokens used for lookup. The possibility of ignoring case, skipping terms for lookup if they appear in a stop word list, use the context of the lookup string (sentence, paragraph) and word order independence are the main features of the component.

Part of the NER component are also the negation and lateralization annotators. The first is a generalized algorithm, based on the popular NegEx [12]. Negation trigger words such as *'sem' (without), 'nunca' (never), 'não' (not))* are specified in a user modifiable dictionary. The trigger words become the anchors around which negated sentences are discovered. When such word is found, the following semantic entities in the sentence are marked as negated. The lateralization annotator uses also trigger words like *'esquerda' (left), 'superior' (superior)* to characterize entities of anatomical site.

4. Results

In this section we report on the precision of MedAlert annotations. As the gold standard set is not yet fully annotated and verified we evaluate the system by analysing its output and determining if it corresponds to true or false positives. This allowed us to evaluate MedAlert's precision, which we consider to be the most important characteristic of systems intended to be used in clinical domains.

Table 3 shows the precision for the recognized entities. The system achieved a precision of 1 for the entities of anatomical site, evolution, location, datetime and value. The first three of these classes are also the ones with less representation on the discharge letters. The dateTime and value entities are mainly recognized with help of the context tokenizer which is based on regular expressions.

The entities of therapeutic were recognized with 0.99 precision. This result is a consequence of the use of the drugs ontology which contains all the active substances and medication used in Portugal. The system did not perform with 100% precision due to the use of the same acronym when refering to *Ácido Clavulânico (Clavulanic Acid)*(AC) and *Auscultação Cardíaca (Cardiac Auscultation)*(AC) in one of the discharge letters.

The entities of finding and condition are recognized with a precision of 0.93. Most of the false positives are originated because of the presence of anatomical site entities near keywords such as *mal* (this word can be interpreted as 'sickness' but is mostly meant as 'poorly' as, for example, in *'poorly controlled Diabetes'*).

The precision for the recognition of examination entities is 0.69, which is significantly lower than for the other entities. The recognition of some active substances present in the discharge letters as laboratory examinations is the main reason for this lower precision value. One example of is Insulin, recognized by the system as an hormone (and consequently as a laboratory examination) and as a therapeutic.

Table 3: MedAlert Results

Entities	True Positives	False Positives	Precision
Condition	75	6	0.93
Anatomical Site	4	0	1
Evolution	7	0	1
Examination	24	11	0.69
Finding	28	2	0.93
Location	1	0	1
Therapeutic	146	1	0.99
DateTime	55	0	1
Value	59	0	1
TOTAL	**399**	**20**	**0.95**

5. Conclusions

In this paper we describe an information extraction system MedAlert which automatically extracts information from free text discharge summaries written in Portuguese. We describe the MedAlert corpus of 915 hypertension related discharge letters and the method used to create the discharge letters representation model. Based on this model, detailed annotation guidelines were created and a gold standard set of approximately 10% of the corpus is currently being manually annotated by a linguistic and a clinician. The system runs within the UIMA framework and contains a set of components capable of using external resources, as ontologies and terminologies, to extract information. The precision of the system was evaluated with values ranging from 0.93-1 except for the entities of the class examination. This is mainly due to the recognition of active substances such such as insulin as laboratory examinations.

5.1. Future work

In order to improve the entity recognition results it is essential the use more context information, namely the information concerning the section in which the entity is referred. Another task of great interest in the medical domain is the automatic code assignment to the discharge letters, which is currently being developed. The main objective of the MedAlert system is to act as a medical decision support system capable of inferring doubts/irregularities in the decisions made by the health professionals. The development of components capable of interact between the structured representation of the discharge letters and the information extraction task are essential to reach this objective.

6. Acknowledgements

Our thanks to César Telmo Oliveira from Hospital Infante D. Pedro, Aveiro and Catarina Oliveira, University of Aveiro for helping developing the knowledge model and gold standard set. Our thanks also to Science and Technology Foundation (PhD grant SFRH/BD/27301/2006).

7. References

[1] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, P. C. de Groen, "Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model", Journal of Biomedical Inform, 42(5): 937-949 (2009).

[2] D. Ferrucci, A. Lally, "UIMA an architectural approach to unstructured information processing in the corporate research environment", in Natural Language Engineering (2004), 10:327-348.

[3] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications", in 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Lisbon, July 2002.

[4] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research", in Yearbook of Medical Informatics, 128-144, 2008.

[5] S. Ananiadou, J. Mcnaught, "Text Mining for Biology And Biomedicine", Artech House, Inc., 2005.

[6] <http://caties.cabig.upmc.edu/>

[7] C. Friedman, S. B. Johnson, B. Forman, J. Starren, "Architectural requirements for a multipurpose natural language processor in the clinical environment", in Proceedings Annual Symposium on Computer Applications in Medical Care, 347-351, 1995.

[8] Philip Ogren, "knowtator: A plug-in for creating training and evaluation data sets for Biomedical Natural Language systems", in Proceedings of the 9th International Protégé Conference, 73–76, Stanford, California, 2006.

[9] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubzy, H. Eriksson, N. F. Noy, Samson W. Tu, "The Evolution of Protégé: An Environment for Knowledge-Based Systems Development, in International Journal of Human-Computer Studies, 58:89-123, 2002.

[10] UMLS Knowledge Sources, United Stated National Library of Medicine, 2008

[11] H. Schmid, "Tree Tagger, a language independent part-of-speech tagger", Institut fur Maschinelle Sprachverarbeitung, Stuttgart University, 1995.

[12] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan, "A simple algorithm for identifying negated findings and diseases in discharge summaries", in Journal of Biomedical Informatics, 34:301-310, 2001.

[13] "Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM", Cristina Mota and Diana Santos (eds), Linguateca, 2008 http://www.linguateca.pt/LivroSegundoHAREM/, ISBN: 978-989-20-1656-6.

New ASR Technique to Enhance the Performance of Spoken Dialogue Systems

R. López-Cózar[1], *Gonzalo Espejo*[1], *Nieves Ábalos*[1], *David Griol*[2], *José F. Quesada*[3]

[1] Dept. of Languages and Computer Systems, CITIC-UGR, University of Granada, Spain
[2] Dept. of Computer Science, Carlos III University of Madrid, Spain
[3] Dept. of Artificial Intelligence and Computer Science, Seville, Spain
rlopezc@ugr.es, {gonzaep,nayade}@correo.ugr.es, dgriol@inf.uc3m.es, Jose.Quesada@infinity.es

Abstract

In this paper we present a new technique to enhance the performance of spoken dialogue system, which employs contextual models and grammatical rules to optimise the automatic correction of some errors made by the ASR component of these systems. Different experiments have been carried out to evaluate this technique employing a previously developed spoken dialogue system designed for the fast food domain. The results of this experimentation show that our technique enhances the performance of the system, by notably incrementing the rates of word accuracy, speech understanding and task completion.

Index Terms: Spoken Dialogue Systems, Automatic Speech Recognition, Language Modelling.

1. Introduction

Spoken dialogue systems (SDSs) are employed nowadays by many companies to provide automatic services such as travel information [1, 2], weather forecasts [3, 4], fast food ordering [5, 6], call routing [7] or directory assistance [8]. These systems can be very convenient for companies as they enable important economic savings in providing services available 24 hours a day, 365 days a year. In addition, these systems can be very handy for users, as they can get easily information (in theory) by means of spontaneous speech using a telephone. However, many users reject using these systems because the interaction many times is not very natural and friendly.

The clearly observable differences with respect to human-to-human interaction are caused by several reasons. One is the current limitations of state-of-the-art automatic speech recognition (ASR) for real-world applications. Hence, to make these systems more widely accepted by all potential users, it is very important to develop methods to increase the robustness of the speech recogniser. One method to do this is by automatically correcting some errors made by this system's component.

Many techniques can be found in the literature addressing this task. For example, [9] proposed to use a channel model and a language model, in which the former takes into account errors made by a speech recogniser whereas the latter provides information about sequences of uttered words. Following a different approach, [10] proposed a technique that carries out ASR correction at two levels of analysis. The former uses a classifier to decide whether the outcome of the ASR is incorrect; if it is, the outcome is passed on to the second level of analysis, where another classifier is used to decide the incorrect words.

One problem with the techniques described above is that they rely on statistical information only, and thus need vast amounts of training data. To overcome this drawback a number of authors have proposed to combine lexical, syntactic or semantic information, and some of them have employed knowledge concerned with dialogue management [11]. Following this approach, the technique that we propose considers statistical, lexical, syntactic, semantic and dialogue-related information. The main novelty is that it takes into account prompt-dependent models to correct the errors, being the optimal model selected by the computation of a similarity score between the pattern obtained from the uttered sentence and patterns learnt during training. In addition, our technique considers grammatical rules to correct errors that cannot be detected using these models.

After this short introduction, the remainder of the paper is organised as follows. Section 2 presents the proposed technique, discussing the information required and the algorithms for implementation. Section 3 addresses the experiments, in which we compare results on word accuracy, sentence understanding and task completion, with and without using the proposed technique in a SDS that we developed in a previous study.

2. The proposed ASR technique

The proposed ASR technique to enhance the performance of SDSs is based on the use of semantic, grammatical and lexical information at the ASR level, as described in the following sections.

2.1. Semantic information

The semantic information is represented by means of what we call *semantic classes*. A semantic class is a set of keywords of a given type which are necessary to extract the semantic content of sentences within an application domain. For example, in our experiments in the fast food domain, we consider, among others, the following semantic classes: DESIRE = {want, need, …}, FOOD = {sandwich, cake, salad, …}, DRINK = {water, beer, wine, …} and AMOUNT = {one, two, three, …}.

2.2. Grammatical information

The grammatical information is represented by means of rules of the following form: *ssp* → *restriction*, where *ssp* denotes a syntactic-semantic pattern and *restriction* is a condition that

This research has been funded by Spanish project HADA TIN2007-64718.

must be satisfied by all the semantic classes in the pattern. For example, one rule used in our experiments is as follows:

NUMBER DRINK SIZE →
number(NUMBER) = *number*(DRINK) **and**
number(DRINK) = *number*(SIZE) **and**
number(NUMBER) = *number*(SIZE)

where *number* is a function that returns either 'singular' or 'plural' for each word in the semantic classes NUMBER, DRINK and SIZE. The goal of this rule is to check number correspondences of drink orders uttered in Spanish. For example, the sentence "dos cervezas grandes" (two large beers) holds this correspondence.

We consider that a dialogue state T represents a prompt type of a SDS by means of which the system expects to obtain a particular type of data from the user, for example, a telephone number. We consider as well that the sentences uttered by users in a dialogue state T can be represented by what we call a *syntactic-semantic* model. To create such a model, we transform each sentence into what we call a *syntactic-semantic pattern* (*ssp*). This pattern is a sequence of semantic classes obtained by replacing each word in the sentence with the semantic class(s) the word belongs to. From the analysis of all the sentences uttered in response to each prompt type we create a set of *ssp*'s, in which we remove those that are redundant and associate to each *ssp* its relative frequency within the set. The outcome of this process is a syntactic-semantic model associated with the prompt type T (SSM_T). We call *α* model the set of SSM_T's created considering the *m* prompt types of a SDS:

$$\alpha = \{SSM_{Ti}\},\ i = 1 \ldots m.$$

2.3. Lexical information

The lexical information takes into account the performance of the speech recogniser of a SDS. In accordance with our technique, we must create a lexical model for each dialogue state T, which we call LM_T. To do so, we consider the sentences uttered in the dialogue state and their corresponding recognition results. The format of this model is: $LM_T = \{w_a, w_b, p_{ab}\}$, where w_a is a word uttered by a user, w_b is the recognised word and p_{ab} is the posterior probability of obtaining w_b given w_a. To create LM_T we align each uttered sentence with the recognised sentence using the method described in [12], and compute the probabilities p_{ab} for each word pair (w_a, w_b). We call *β model* the set of LM_T's created considering the *m* prompt types of a SDS:

$$\beta = \{LM_{Ti}\},\ i = 1 \ldots m.$$

2.4. Algorithms to implement the technique

The correction of ASR errors is performed at two levels (statistical and linguistic) as explained in the following sections.

2.4.1. Correction at statistical level

The goal of this correction level is to find words w_I's in the recognised sentence which belong to incorrect semantic classes K_I's. For each word, we must decide the correct semantic class K_C and select the most appropriate word $w_C \in K_C$ to substitute w_I in the recognised sentence. We can implement this procedure in two steps:

Step 1. Pattern matching. This step employs what we call an *enriched syntactic-semantic pattern* ($essp_{INPUT}$) obtained from the recognised sentence. This pattern is a sequence of what we call *containers*. Each container stores a word of the sentence and has a name if the word is a keyword, which is the name of the semantic class the word belongs to (e.g., DESIRE). The goal of this step is to transform $essp_{INPUT}$ into another pattern called $essp_{BEST}$, which is initially empty. To create this new pattern, we firstly create a syntactic-semantic pattern called ssp_{INPUT}, which only contains the semantic classes in $essp_{INPUT}$, for example: ssp_{INPUT} = DESIRE AMOUNT INGREDIENT FOOD.

Next, we decide whether ssp_{INPUT} matches any pattern in the syntactic-semantic model associated with the dialogue state T (SSM_T). If so, we make $essp_{BEST} = essp_{INPUT}$ and proceed with the correction at the linguistic level (section 2.4.2). Otherwise, we look for patterns similar to ssp_{INPUT} in SSM_T. To do this we compare ssp_{INPUT} with every pattern *p* in the model, and compute a similarity score as follows: $similarity(ssp_{INPUT}, p) = (n - m_{ed}) / n$, where *n* is the number of semantic classes in ssp_{INPUT} and m_{ed} is the minimum edit distance between both patterns, computed using the method described in [13]. We call $ssp_{SIMILAR}$ any pattern *p* in SSM_T such that similarity(ssp_{INPUT}, *p*) > *t*, where $t \in [0.0, 1.0]$ is a similarity threshold, the optimal value of which must be experimentally determined. We consider 3 cases depending on the number of $ssp_{SIMILAR}$'s in SSM_T:

Case 1. There is just one $ssp_{SIMILAR}$ in SSM_T. Thus, we create a new pattern called ssp_{BEST}, make $ssp_{BEST} = ssp_{SIMILAR}$ and proceed with Step 2 (Pattern alignment).

Case 2. There are no $ssp_{SIMILAR}$'s in SSM_T. Thus, we try to find $ssp_{SIMILAR}$'s in the α model (discussed in section 2.2). If no $ssp_{SIMILAR}$'s are found, we do not make any correction at the statistical level; if there is just one, we proceed as in Case 1; if there are several, we proceed as in Case 3.

Case 3. There are several $ssp_{SIMILAR}$'s in SSM_T (or in α). The question then is to decide the best $ssp_{SIMILAR}$. To make this selection we search for the $ssp_{SIMILAR}$ that has the greatest similarity with ssp_{INPUT}. If there is just one $ssp_{SIMILAR}$ satisfying this condition, we make $ssp_{BEST} = ssp_{SIMILAR}$ and proceed with Step 2. If there are several patterns, we select those with the highest frequency in SSM_T (or in α): if there is just one, we make $ssp_{BEST} = ssp_{SIMILAR}$ and proceed with Step 2; if there are several we do not make any correction at the statistical level.

Step 2. Pattern alignment. The goal of this step is to build $essp_{BEST}$ in case it is still empty. To do this, we take into account each container C_a in ssp_{INPUT} and consider three cases:

Case A. The word w_a in C_a does not affect the semantics of the sentence, i.e., it is not a keyword (e.g. '*please*'). Thus, we create a new container *D*, make $D = C_a$ and add *D* to $essp_{BEST}$.

Case B. The word w_a in C_a affects the semantics of the sentence, i.e., it is a keyword (e.g. '*sandwich*'). Thus, we study whether the word must be corrected. To do this, we try to align the container C_a with a container C_b in ssp_{BEST} using the method described in [12] and consider three cases:

Case B.1. C_a can be aligned. In this case we assume that the container C_a is correct and do not make any correction at the statistical level. We create a new container D, make $D = C_a$ and add D to $essp_{BEST}$.

Case B.2. It is not possible to align C_a. This case may happen in the two following situations:

Case B.2.1. The container is a result of an insertion recognition error. In this case we discard C_a, i.e. it is not added to $essp_{BEST}$.

Case B.2.2. The container is a result of a substitution recognition error. Therefore, we must find a correction word from a different semantic class, $w_C \in C_b$, store it in a new container D, and add this container to $essp_{BEST}$. To find w_C we consider the lexical model associated with the dialogue state T (LM_T) and create the set U of words $u \in C_b$ with which the word w_I is confused. If there is only one word u in U, we create a new container D that we name C_b, store it in u, and add D to $essp_{BEST}$. If there are several words, we carry out the same procedure but using the word that has the highest confusion probability with w_I if it is unique; if it is not unique, or there are no words in U, we do not make any correction at the statistical level.

2.4.2. Correction at the linguistic level

The goal of this correction level is to repair errors that are not detected at the statistical level and which affect the semantics of the sentences. To carry out the correction we use the grammatical rules described in section 2.2. For each rule we carry out the following procedure. The syntactic-semantic pattern *ssp* of the rule is inserted in a *window* that slides from left to right over $essp_{BEST}$. If the sequence of semantic classes in the window is found in $essp_{BEST}$, then we apply the *restriction* of the rule to the words in the containers of $essp_{BEST}$. If the words satisfy the restriction, we do not make any correction. Otherwise, we try to find out the reason for the insatisfaction by searching for an incorrect word w_I. To decide the word w_C to correct the incorrect word, we consider the lexical model LM_T and take into account the set $U = \{u_1, u_2, ..., u_p\}$ comprised of words of the same semantic class than the word w_I. Next, we proceed similarly as discussed in Case B.2.2 but considering that the goal now is to replace one word in one semantic class with other word in the same semantic class.

3. Experiments

The goal of the experiments is to test the proposed technique using the Saplen system, which we developed in a previous study to answer fast food queries and orders made in Spanish [6]. The evaluation has been carried out in terms of word accuracy (WA), speech understanding (SU) and task completion (TC), considering two front-ends for ASR: i) *baseline ASR*, comprised of the standard HTK-based speech recogniser of the Saplen system, and ii) *enhanced ASR*, comprised of the same speech recogniser plus an additional module that implements the proposed technique.

We have employed a dialogue corpus collected in our University from students interacting with the Saplen system, which contains around 5,500 utterances and roughly 2,000 different words. The utterance corpus has been divided into two separate corpora, each containing around 50% of the utterances. Using the training corpus we have compiled a word bigram that allows recognising sentences of the 18 different types in the corpus. The remaining 50% of the utterances have been used for testing.

The experiments have been carried out employing a user simulator developed in a previous study [15]. The interaction between the Saplen system and the simulator is decided considering a set of scenarios that represent user goals. We have created two scenario sets: *ScenariosA* (300 scenarios) and *ScenariosB* (100 scenarios). Each dialogue generated by the interaction between the Saplen system and the user simulator is stored in a log file for analysis and evaluation purposes.

Given that the construction of the syntactic-semantic and lexical models described in sections 2.2 and 2.3 has been carried out employing simulated dialogues, we have made additional experiments to decide the necessary number of dialogues to obtain the maximum amount of syntactic-semantic and lexical knowledge. The results indicate that 900 dialogues is the optimal trade-off.

3.1. Experiments with the baseline ASR

Employing the user simulator, the Saplen system and *ScenariosA,* we have generated a corpus of 900 dialogues, which we have called $DialoguesA_1$. Table 1 sets out the average results obtained from the analysis of this corpus. The results show the problems of the system in correctly recognising and understanding some utterances. Analysis of the log files reveals that in some cases the misrecognised sentences are similar to the uttered sentences. For example, “dos fantas grandes de limón” (two large lemon fantas) is recognised as “uno fantas grandes de limón” (one large lemon fantas) because of the acoustic similarly between ‘dos’ and ‘uno’ when uttered by users with strong Southern Spanish accents.

Table 1. *Results using the baseline ASR (in %).*

WA	SU	TC
76,12	54,71	24,51

We have also observed problems with confirmations, which happen because the speech recogniser usually substitutes the word ‘sí’ (yes) by the word ‘seis’ (six), when the former word is uttered by strongly accented speakers. In other cases, the recognised sentences are very distorted by ASR errors. For example, the sentence “quiero una fanta de naranja grande” (I want one big orange Fanta) is sometimes recognised as “queso de manzana tercera” (cheese of apple third).

3.2. Experiments with the enhanced ASR

As the semantic classes required for the technique (discussed in section 2.1), we have employed a set of 21 semantic classes that we created in a previous study [14]. Following section 2.2 we have created a set of grammatical rules to check the number correspondences for food and drink orders. To create the syntactic-semantic and lexical models, discussed in sections 2.2 and 2.3, we have analysed $DialoguesA_1$ thus obtaining $\alpha = \{SSM_{Ti}\}$ and $\beta = \{LM_{Ti}\}$, with i = 1 ... 43 given that the Saplen system can be in 43 different dialogue states.

To decide the optimal value for the similarity threshold t (discussed in section 2.4.1) we have carried out experiments considering values in the range [0.1, 0.9]. Employing the user simulator and *ScenariosB*, we have generated a corpus comprised of 300 dialogues for each value, using in all cases the proposed technique. Analysis of the outcomes of these experiments reveals that the best results are obtained when t =

0.5. Using this optimal value, we have employed again *ScenariosA* to generate another corpus of 900 dialogues, which we call *DialoguesA*$_2$. Table 2 shows the average results obtained from the analysis of this corpus.

Table 2. *Results using the enhanced ASR (in %).*

WA	SU	TC
84,62	71,25	68,32

Analysis of the log files shows that the technique is successful in correcting some incorrectly recognised sentences. For example, the incorrectly recognised drink order "one large lemon fantas" is corrected by doing no changes at the syntactic-semantic level, and replacing 'one' with 'two' at the lexical level. In other product orders the correction is carried out at the semantic-syntactic level. For example, "one curry salad" is sometimes recognised as "one error curry salad". In this case the correction is carried out removing the ERROR semantic class at the syntactic-semantic level.

The technique is useful in correcting the errors with confirmations discussed in the previous section. To do this, it replaces the semantic class NUMBER with the semantic class CONFIRMATION, and then selects the most likely word in CONFIRMATION.

The enhanced ASR enables as well correction of some misrecognised telephone numbers. For example, "nine five eight twenty-one fourteen eighteen" is sometimes recognised as "gimme five eight twenty-one fourteen eighteen" because of acoustic similarity between 'nine' and 'gimme' in Spanish. The technique corrects the error by replacing the semantic class DESIRE with the semantic class NUMBER and selecting the most likely word in NUMBER given the word 'gimme' at the lexical level.

The technique is also useful to correct some misrecognised postal codes. For example, "eighteen zero zero one" is sometimes recognised as "eighteen zero zero turkey". This error is corrected by replacing the semantic class INGREDIENT with the semantic class NUMBER and selecting the most likely word in NUMBER given the word 'turkey'.

Our proposal is also successful in correcting some incorrectly recognised addresses (in the Spanish format). For example, "almona del boquerón street number five second floor letter h" is sometimes recognised as "almona del boquerón street error five second floor letter zero". This error is corrected by making a double correction. First, replacement of the semantic class ERROR with the semantic class NUMBER_ID and selection of the most likely word in NUMBER_ID given the word 'error'. Second, replacement of the semantic class NUMBER concept with the semantic class LETTER and selection of the most likely word in LETTER given the word 'zero'.

There are cases where the technique fails in detecting errors, and thus in correcting them. This happens when words in the uttered sentence are substituted by other words and the result is valid in the application domain. For example, this occurs when the sentence "two green salads" is recognised as "twelve green salads", given that there is no conflict in terms of semantic classes and there is agreement in number between the words.

4. Conclusions and future work

Comparing the results set out in Tables 1 and 2 we observe that the proposed technique allows enhancing the performance of the Saplen system in terms of WA, SU and TC by 8.5%, 16.54% and 44.17% absolute, respectively. These enhancements are mostly achieved because considering the proposed threshold for similarity scores between patterns, the technique decides whether to use correction models associated with the current dialogue state, or general correction models for the application domain. In particular, we have observed that the benefit of the proposed method is particularly noticeable in the correction of misrecognised confirmations.

Future work includes considering additional information sources to correct errors that in the current implementation cannot be detected, such as domain-dependent knowledge. For example, in our application domain we could use this kind of information to consider that the sentence "twelve green salads", although syntactically correct, is likely to be incorrectly recognised, given that it is not usual that the users order such a large amount of a product. We also plan to study the performance of the technique considering prompt-dependent similarity thresholds.

5. References

[1] Seneff S., Polifroni J. "Dialogue management in the Mercury flight reservation system", Proc. of ANLP-NAACL Satellite Workshop, pp. 1-6, 2000.

[2] Billi, R., Castagneri, G., Danielli, M. "Field trial evaluations of two different information inquiry systems", *Speech Communication*, 23, pp. 83–93, 1997.

[3] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L. "Jupiter: A telephone-based conversational interface for weather information",. *IEEE Trans. on Speech and Audio Proc*., 8(1), pp. 85-96, 2000.

[4] Nakano, N., Minami, Y., Seneff, S., Hazen, T. J., Cyphers, D. S., Glass, J., Polifroni, J., Zue, V. "Mokusei: A telephone-based Japanese conversational system in the weather domain", Proc. of Eurospeech, pp. 1331-1334, 2001.

[5] Seto, S., Kanazawa, H., Shinchi, H., Takebayashi, Y. "Spontaneous speech dialogue system TOSBURG II and its evaluation", *Speech Communication*, 15, pp. 341-353, 1994.

[6] López-Cózar, R., García, P., Díaz, J., Rubio, A. J. "A voice activated dialogue system for fast-food restaurant applications", Proc. of Eurospeech, pp. 1783-1786, 1997.

[7] Lee, C.-H., Carpenter, B., Chou, W., Chu-Carroll, J., Reichl, W., Saad, A., Zhou, Q. "On natural language call routing", *Speech Communication*, 31, pp. 309-320, 2000.

[8] Kellner, A., Rueber, B., Seide, F., Tran, B.-H. "PADIS – An automatic telephone switchboard and directory information system", *Speech Communication*, 23, pp. 95–111, 1997.

[9] Ringger, E. K., Allen, J. F., "A fertility model for post correction of continuous speech recognition", Proc. of ANLP-NAACL Satellite Workshop, pp. 1-6, 2000.

[10] Zhou, Z., Meng, W. K., "A multi-pass error detection and correction framework for Mandarin LVCSR", Proc. of ICSLP, pp. 1646-1649, 2006.

[11] Jeong, M., Jung, S., Lee, G. G. "Speech recognition error correction using maximum entropy language model", Proc. of Interspeech, pp. 2137-2140, 2004.

[12] Fisher, W. M., Fiscus, J. G. "Better alignment procedures for speech recognition evaluation", Proc. ICASSP, pp. 59-62, 1993.

[13] Crestani, F. "Word recognition errors and relevance feedback in spoken query processing", Proc. of Conf. on Flexible Query Answering Systems, pp. 267-281, 2000.

[14] López-Cózar, R., Callejas, Z. "Combining language models in the input interface of a spoken dialogue system", Computer Speech and Language, 20, pp. 420-440, 2006.

[15] López-Cózar, R., Callejas, Z., McTear, M. "Testing the performance of a spoken dialogue system by means of a new artificially simulated user", Artificial Intelligence Review, 26, pp. 291-323, 2006

Seamless Tree Binarization for Interactive Predictive Parsing

Ricardo Sánchez-Sáez[†], *Luis A. Leiva*[‡], *Joan-Andreu Sánchez*[†], *José-Miguel Benedí*[†]

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, Spain
{rsanchez,luileito,jandreu,jbenedi}@{[†]dsic,[‡]iti}.upv.es

Abstract

This paper introduces a seamless method for tree binarization/debinarization that is employed within the Interactive Predictive Parsing framework for tree annotation. This novel method allows that, while the human annotator verifies and corrects standard non-binary trees, the parse engine can work with parsing algorithms that process and produce binary trees, such as a CYK-Viterbi based parser.

Within the Interactive Predictive Parsing framework the user is tightly integrated into the interactive parsing system, in contrast with the traditional post-editing approach. User feedback for tree correction and validation is provided by means of natural mouse gestures and keyboard strokes.

Index Terms: parsing, interactive predictive parsing, syntactic tree annotation, tree binarization

1. Introduction

Probabilistic parsing is a fundamental problem in Computational Linguistics. Probabilistic parsing has been greatly benefited in the past from the availability of annotated corpora. Perfectly annotated parsing trees are used in the training of new automatic parsing systems, and are needed for parser validation and experimentation. Therefore, there is a pressing need in efficiently constructing new perfectly annotated corpora.

The trees contained in these corpora must be manually annotated: either creating them from scratch, or based on automatically obtained error-prone parse trees. This results in a laborious and time-consuming task.

Several tools exist that can aid in easing the work of human annotators. Some examples are the TreeBanker [2], TrEd[1] or eBonsai [3], for structural annotation; or DepAnn [4], for dependency style annotation. The well-known Penn Treebank itself was annotated using automatically obtained basic skeletal parses, and an aid tool was used to finish the annotation of the parse trees [6].

The problem with using these tools for treebank creation is that they all typically introduce a two-step workflow: first, a chosen system generates the best tree for the sentence being annotated, and then the human annotator has to verify it and amend the errors within the proposed parse tree. This paradigm is rather inefficient and uncomfortable for the human annotator.

Recent work has introduced a new type of Web-based interactive predictive annotation tool, the Interactive Predictive Parsing Tree Annotator (or IPP-Ann) [10]. This tool follows the Interactive Predictive Parsing (IPP) paradigm, whose novelty is that it fully integrates the human annotator into the parsing loop, making him part of the system. The annotator interacts in real time with the IPP engine, and the system uses the readily available user feedback to make predictions about the parts of the tree that have not been validated by the corrector.

Experiments carried out to simulate user interaction with the IPP framework suggest figures ranging from 42% to 46% of effort saving compared to manually post-editing the trees without an interactive system, both for English [11] and Spanish [9] sentence annotation. Additionally, this kind of man-machine integration presents yet unexplored opportunities, such as the scenario in which the parsing system adapts its models, incorporating the new ground-truth data provided by the user.

IPP-Ann is a implementation of the IPP framework, and takes the form of a decoupled annotation system consisting in a parse engine and a Web client working together. IPP-Ann can be used online at `http://cat.iti.upv.es/ipp/`.

The parsing subsystem of IPP-Ann currently uses a CYK-Viterbi based parsing algorithm which works with a Probabilistic Context Free Grammar (PCFG) in Chomsky Normal Form (CNF) as its model. Algorithms that use grammars in CNF can only produce and process binary trees. An automatic process for binarization/debinarization of the trees going through the parse engine is needed, so they are presented to the human annotator in an usable non-binary form.

In this paper, after reviewing the IPP theoretical framework and the IPP-Ann system, we present a novel method for seamless tree binarization/debinarization. This method allows that, while the human annotator uses IPP-Ann to modify and annotate standard non-binary trees, the parsing subsystem of the annotation tool is able to internally work with parsing algorithms that process and generate binary trees, such as the CYK-Viterbi parsing algorithm. Parsing algorithms that use binary trees are widespread in the parsing world, as they are simpler to understand and more efficient.

2. Interactive Predictive Parsing Framework

In this section we review the IPP framework [11]. Interactive predictive methods have been successfully demonstrated to ease the work of transcriptors and translators in fields like Handwriting Text Transcription [8, 12] and Statistical Machine Translation [7, 13].

A tree t, associated to a string $x_{1|x|}$, is composed by substructures that are usually referred as constituents. A constituent c_{ij}^{A} is defined by the non-terminal symbol A (either a *syntactic label* or a *POS tag*) and its span ij (the starting and ending indexes which delimit the part of the input sentence encompassed by the constituent).

Here follows a general formulation for the non-interactive parsing scenario. Using a grammatical model G, the parser analyzes the input sentence $\boldsymbol{x} = \{x_1, \ldots, x_{|x|}\}$ and produces the

[1] `http://ufal.mff.cuni.cz/~pajas/tred/`

parse tree $\hat{t}$

$$\hat{t} = \arg\max_{t \in \mathcal{T}} p_G(t|\boldsymbol{x}), \tag{1}$$

where $p_G(t|\boldsymbol{x})$ is the probability of parse tree t given the input string $\boldsymbol{x}$ using model G, and $\mathcal{T}$ is the set of all possible parse trees for $\boldsymbol{x}$.

In the interactive predictive scenario, after obtaining the (probably incorrect) best tree $\hat{t}$, the user is able to individually correct any of its constituents c_{ij}^{A}. The system reacts to each of the corrections introduced by the human, proposing a new $\hat{t}'$ that takes into account the afore-mentioned corrections.

Within the IPP framework, the user reviews the constituents contained in the tree to assess their correctness. The action of modifying an incorrect constituent (either setting the correct span or the correct label) implicitly validates a subtree that is composed by the partially corrected constituent, all of its ancestor constituents, and all constituents whose end span is lower than the start span of the corrected constituent. We will name this subtree the validated prefix tree t_p (for analogy with the prefix sentence in the above mentioned interactive machine translation and interactive text transcription scenarios). When the user replaces the constituent c_{ij}^{A} with the correct one $c_{ij}'^{A}$, the validated prefix tree is:

$$\begin{aligned} t_p(c_{ij}'^{A}) = \{ & c_{mn}^{B} \ : \ m \leq i, \ n \geq j, \\ & d(c_{mn}^{B}) \leq d(c_{ij}'^{A}) \} \ \cup \\ \{ & c_{pq}^{D} \ : \ p >= 1 \ , \ q < i \} \end{aligned} \tag{2}$$

with $d(c_{mn}^{B})$ being the depth of constituent c_{mn}^{B}.

When a constituent correction is performed, the prefix tree $t_p(c_{ij}'^{A})$ is fixed and a new tree $\hat{t}'$ that takes into account the prefix is proposed:

$$\hat{t}' = \arg\max_{t \in \mathcal{T}} p_G(t|\boldsymbol{x}, t_p(c_{ij}'^{A})). \tag{3}$$

Given that we are working with context-free grammars, the only subtree that effectively needs to be recalculated is the one starting from the parent of the corrected constituent.

3. The IPP Tree Annotator

IPP-Ann has been previously adapted to parse both English (with a Penntreebank based model) [10] and Spanish text (with a UAM Treebank based model) [9], and can be accessed at `http://cat.iti.upv.es/ipp/`.

IPP-Ann can help users to efficiently annotate correct syntactic trees. For this, the user feedback (provided by means of keyboard and mouse operations) allows the system to predict new subtrees for unvalidated parts of the annotated sentence, which in turn reduces the human effort and improves annotation efficiency.

The IPP Tree Annotator uses the CAT-API library [1] as a communication backend between the server and client modules. This library allows for a clean application design, in which both the server side (the parsing engine) and the client interface (which draws the trees, captures and interprets the user feedback, and requests parsed subtrees to the server) are independent of each other. One of the features that steams from the CAT-API library is the ability for several annotators to work concurrently on the same problem-set, each in a different client computer sharing the same parsing server.

When working with IPP-Ann, the user is presented with the sentences from the selected corpus, and starts parsing them one by one. They can then make corrections in the trees and the user feedback is decoded on the client side which in turn requests subtrees to the parse engine.

Two kind of operations can be performed over constituents when using IPP-Ann: span modification (done by dragging a line from the constituent to the word that corresponds to the span's upper index), and label substitution (done by typing the correct one on its text field). Modifying the span of a constituent invalidates its label, so the server recalculates it as part of the suffix. Modifying the label of a constituent validates its span. Constituents can be deleted or inserted by adequately modifying the span of the left-neighbouring constituent. Figure 1 shows a span modification performed by a human annotator using.

Additionally, two unary-related operations can be performed over constituents: unary production insertion (done by drawing a line from the constituent node to the floating ball that appears below itself), and unary production removal (done by resetting the span of the constituent parenting the unary production).

As visual aid, when the user is about to perform an operation, the affected constituent and the prefix that will be validated are highlighted. The target span of the modified constituent is visually shown as well. When the user obtains the correctly annotated tree, they can accept it by by clicking on a new sentence.

3.1. Server side implementation

The server side of the system is a parsing engine based on a customized CYK-Viterbi parser, which uses a Probabilistic Context-Free Grammar in Chomsky Normal Form. The English grammar was obtained from sections 2 to 21 of the UPenn Treebank as a model [11], and the Spanish grammar was obtained from the first 1400 sentences of the UAM Treebank [9].

The client can send requests to the parsing server in order to obtain the best subtree for any given span of the input string. For each requested subtree, the client can either provide the root label or not. If the subtree root label is not provided, the server calculates the most probable label. The server also performs transparent tree debinarization/binarization, as presented in this work, and unary-rule expansion when communicating with the client.

3.2. Client side implementation

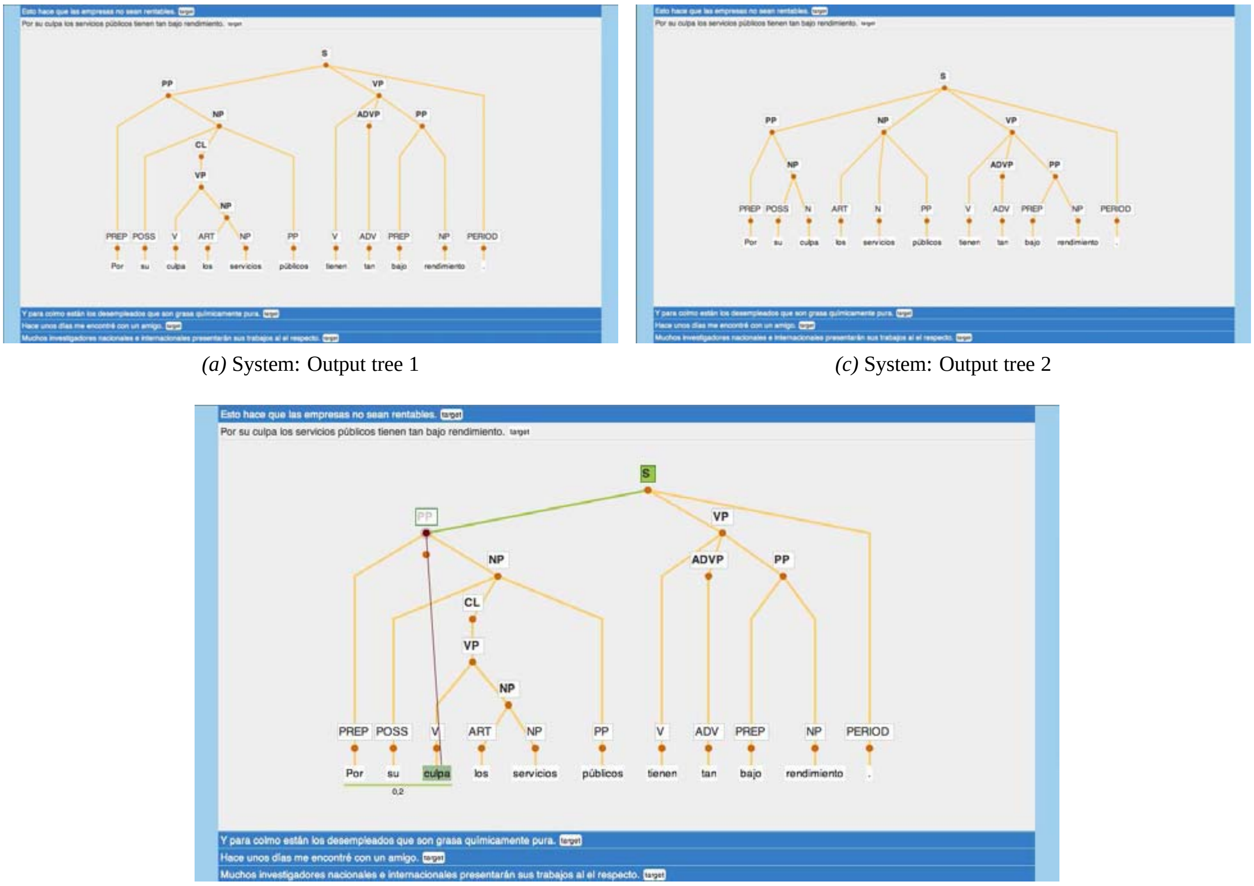
The client part is implemented in a combination of HTML, PHP and Flash ActionScript. As such, it is accessed through a Web browser, being the Flash plugin the only requisite. The hardware requirements are very low on the client side, as the parsing process is performed remotely on the server side: any computer (including netbooks) capable of running a modern Web browser is enough.

The Web client side of IPP-Ann communicates with the IPP engine through binary TCP sockets, resulting in very low response times.

4. Transparent binarization

In this section we introduce a transparent debinarization/binarization process which is consistent with our Interactive Predictive strategy and with our parsing model.

On the one hand, the CYK-Viterbi algorithm used within the parsing server works with Probabilistic Context-Free Grammars in Chomsky Normal Form. By the usage of CNF-PCFGs



(a) System: Output tree 1

(c) System: Output tree 2

(b) User: Span modification

Figure 1: Interaction example on the IPP Tree Annotation tool.

as the model to obtain syntactic trees, only binary trees can be obtained. In our case, the grammars used by the server were obtained using the Chomsky Normal Form transformation method from the Open Source Natural Language Toolkit[2] (NLTK) to calculate minimal right-factored binary grammars [5] from the corresponding regular vanilla reebank grammars.

On the other hand, manual tree annotation is generally performed using non-binary trees, as the underlying syntactic structures present in natural language sentences do not conform to binary trees. To sum up, the human annotator needs to work with non binary trees but the interactive predictive parsing algorithm uses and produces binary trees.

This fundamental necessity led us to devise and implement a seamless and automatic debinarization/binarization process for the trees produced by the parsing server. The debinarization/binarization process is performed in real time and in a completely oblivious form to the user. The human annotator working with regular non-binary trees never notices that the parsing server is internally working with binary trees.

Our seamless binarization/debinarization method employs the aforementioned Chomsky Normal Form transformation, which has been natively implemented within the parsing server. The debinarization process is performed when the algorithm calculates a new subtree, before sending it to the client. The tree binarization process is carried out when the server receives a new tree with the user corrections, before sending it to the interactive predictive parsing algorithm.

[2] `http://www.nltk.org/`

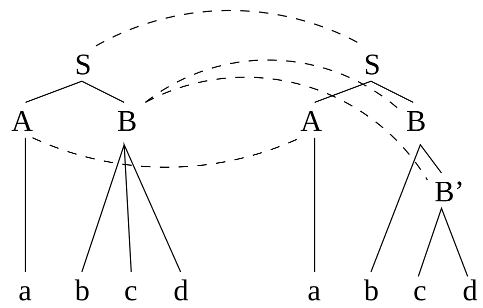
4.1. Linked tree structure

In a regular, non-binary tree, the non-terminal of each node (excluding POS tag nodes) corresponds to a syntactic label. Each represents a syntactic structure that relates to a sequence of words from the sentence being parsed. When such a tree is binarized by the CNF transformation, some new *dummy* nodes are introduced when one node has more than two descendants. These newly *introduced nodes* have non-terminals that do not carry new syntactic information by themselves. Instead, their only function is to propagate the syntactic label of the original non-binary ancestor through the structure of the binary tree.

When IPP-Ann is being employed by an annotator, the IPP algorithm needs to be informed about which tree constituent — the node with its syntactic label and span — was modified after each user interaction. Given that the user performs changes on regular trees via the client interface but the parsing algorithm works with their binarized counterparts, the binarization process must keep information about the correspondency between the introduced nodes in the binary tree and their original ancestor in the non-binary tree.

At binarization/debinarization time, for each introduced node in the binary tree, we note its corresponding ancestor node in the regular tree. Each *non-introduced node* in the binary tree is also linked to the matching node in the regular tree. This generates a one-to-many relationship, in which each node of the regular tree relates to one or more nodes of the binary tree (either just one *non-introduced node*, or one *non-introduced* node plus several *introduced nodes*).

In order to keep the node correspondency information consistent over the binarization/debinarization process, we constructed a new tree structure which we call a *linked tree*. A linked tree consists on the binary and non-binary versions of the same tree, and the one-to-many relationships between the nodes of both trees. See Figure 2 for an example of a linked tree structure.



(a) Non−binary tree (b) Binary tree

Figure 2: Linked tree structure. Note how the *introduced node* B' in the binary tree is related to its original ancestor in the non-binary tree.

Within a linked tree, when either the binary tree or the regular tree is modified, the debinarization/binarization process operates automatically over the linked tree structure, recalculating the associated tree and updating the correspondency relationships. This action takes place either when the user modifies a constituent in the regular tree, or when the parsing algorithm recalculates a new binary subtree based on the user feedback. This method allows for the regular tree, the binary tree and their node correspondency information to remain synchronized and up-to-date at all times.

5. Conclusions and future work

We have introduced a seamless method for tree binarization/debinarization whithin the Interactive Predictive Parsing framework. This novel method allows that while the human annotator verifies and corrects standard non-binary trees, the parse engine can work with parsing algorithms that process and produce binary trees, such as the CYK-Viterbi parser.

The reviewed tool, by using a parse engine in an integrated manner, aids the user in creating correctly annotated syntactic trees. IPP-Ann greatly reduces the human effort required for this task compared to using a non-interactive automatic system.

Future work includes improvements to the client side (e.g., confidence measures as a visual aid, multimodality), as well as exploring other kinds of parsing algorithms for the server side (e.g., adaptative parsing).

6. Acknowledgements

Work partially supported by the Spanish MICINN under the MIPRCV “Consolider Ingenio 2010” (CSD2007-00018), MITTRAL (TIN2009-14633-C03-01), Prometeo (PROMETEO/2009/014) research projects, and the FPU fellowship AP2006-01363. The authors wish to thank Vicent Alabau for his invaluable help with the CAT-API library.

7. References

[1] V. Alabau, D. Ortiz, V. Romero, and J. Ocampo. A multimodal predictive-interactive application for computer assisted transcription and translation. In *Proc. of ICMI-MLMI*, pages 227–228, New York, USA, 2009. ACM.

[2] D. Carter. The TreeBanker. A tool for supervised training of parsed corpora. In *Proc. of ENVGRAM Workshop*, pages 9–15, 1997.

[3] I. Hiroshi, N. Masaki, H. Taiichi, T. Takenobu, and T. Hozumi. eBonsai: An integrated environment for annotating treebanks. In *Proc. of IJCNLP*, pages 108–113, 2005.

[4] T. Kakkonen. Depann - an annotation tool for dependency treebanks. In *Proc. of ESSLLI Student Session*, pages 214–225, Malaga, Spain, June 2006.

[5] D. Klein and C.D. Manning. Accurate unlexicalized parsing. In *Proc. of ACL*, volume 1, pages 423–430, Morristown, USA, 2003. ACL.

[6] M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[7] D. Ortiz, L.A. Leiva, V. Alabau, and F. Casacuberta. Interactive machine translation using a web-based architecture. In *Proc. of IUI*, pages 423–425. Hong Kong, China, February 2010.

[8] V. Romero, L.A. Leiva, A.H. Toselli, and E. Vidal. Interactive multimodal transcription of text imagse using a web-based demo system. In *Proc. of IUI*, pages 477–478. Sanibel Island, Florida, February 2009.

[9] R. Sánchez-Sáez, L.A. Leiva, J.A. Sánchez, and J.M. Benedí. Interactive predictive parsing framework for the spanish language. In *Proc. of SEPLN*, Valencia, Spain, September 2010.

[10] R. Sánchez-Sáez, L.A. Leiva, J.A. Sánchez, and J.M. Benedí. Interactive predictive parsing using a web-based architecture. In *Proc. of NAACL-HLT*, Los Angeles, United States of America, June 2010.

[11] R. Sánchez-Sáez, J.A. Sánchez, and J.M. Benedí. Interactive predictive parsing. In *Proc. of IWPT'09*, pages 222–225, Paris, France, October 2009. ACL.

[12] A.H. Toselli, V. Romero, and E. Vidal. Computer assisted transcription of text images and multimodal interaction. In *Proc. MLMI*, volume 5237, pages 296–308. Springer, 2008.

[13] E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera, and C. Martínez. Computer-assisted translation using speech recognition. *IEEE TASLP*, 14(3):941–951, 2006.

Knowledge Extraction from Minutes of Portuguese Municipalities Meetings

Mário Rodrigues[1], *Gonçalo Paiva Dias*[2], *António Teixeira*[3]

[1]ESTGA/IEETA, [2]ESTGA/GOVCOPP, [3]DETI/IEETA
University of Aveiro, Aveiro, Portugal
mjfr@ua.pt, gpd@ua.pt, ajst@ua.pt

Abstract

A very relevant problem in e-government is that a great amount of knowledge is in natural language unstructured documents. If that knowledge was stored using a computer-processable representation it would be more easily accessed. In this paper we present the architecture, modules and initial results of a prototype under development for extracting information from government documents. The prototype stores the information using a formal representation of the set of concepts and the relationships between those concepts - an ontology. The system was tested using minutes of Portuguese Municipal Boards meetings. Initial results are presented for an important and frequent topic of the minutes: the subsidies granted by municipalities.

Index Terms: entities and relations extraction, e-government, semantic query.

1. Introduction

E-government relates to the use of information and communication technologies (ICT) by government, including the online provision of government services. These technologies have the potential to improve the government service delivery, can empower citizens through access to information, or make government management more efficient. A great amount of information in local and central government agencies is registered in unstructured formats, making the access/query/search to it not readily available. Although many of these documents are stored in computers, their format prevents the information they contain to be computer-processable and thus they cannot be manipulated to meet user's specific needs. E-government would benefit from systems able to integrate several sources of information and able to understand unstructured documents [1].

The minutes of Municipal Board meetings contain information that would benefit from being made available in searchable knowledge bases. These documents are important because municipalities are often the closest point of service for citizens and enterprises, and the minutes record the decisions of the Municipal Board.

In this paper we present a system able to create semantic information from this type of documents - natural language, unstructured - using natural language processing algorithms, integrating open source software, and using external sources of information as Google Maps and Geo-Net-PT01. The remaining of the paper starts by discussing the related work. Section 2 starts with an overview of the developed system and continues by elaborating on each of the three parts that compose it. In Section 3 the initial results are presented and discussed. The paper ends with the conclusions in Section 4 and the acknowledgments in Section 5.

1.1. Related Work

The research activity done so far in e-government is usually centered in solving problems as interoperability and service integration, which are very important problems and should be further addressed. In such projects it is usually considered that the information is already in the system, whether placed by human operators or using existing databases (e.g. OneStopGov and Access-eGov). To our knowledge, no project was dedicated to the relevant problem of automatic acquisition of information from natural language government documents [1].

Several projects were dedicated to the task of scalable, domain independent information extraction (IE). Some are more focused in building semantic knowledge bases from Wikipedia - DBpedia, Kylin, YAGO/NAGA and others. They generally extract information from the (natural language) documents texts and use the Wikipedia's structure to infer the semantics. Other state of the art systems work at the web scale - KnowItAll, TextRunner. Finally, some works focus on the knowledge evolution over time - TARSQI, Timely YAGO.

2. Developed System

The proposed system reuses open source software - adapted to work with Portuguese - to take advantage of the state of the art approaches and software. Specific software was developed to integrate the reused software in a coherent system (Figure 1).

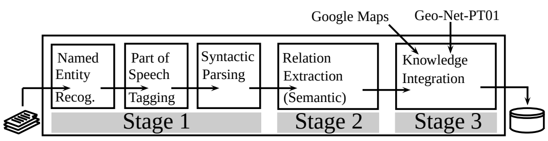


Figure 1: *The three stages of processing and their components. Input: natural language; Output: structured information.*

This information extraction (IE) system is composed of three parts, organized in pipeline architecture. First, the text is extracted and enriched with the inclusion of named entity (NE), part-of-speech (POS) tags and syntactic information. Second, the system uses this enriched text to train a classifier that looks for patterns of information in order to detect relationships between entities. Third, the system integrates the knowledge extracted in the previous part with information from other sources (e.g. geocode information) and stores it in a knowledge base that conforms to a defined ontology. This part also includes the capabilities to semantically query the knowledge base.

The following sections explain how each module was adapted and integrated. The explanations are illustrated using the fragment of a minute presented in the first row of Table 1.

Fragment:	"Seguidamente, a Câmara deliberou, por unanimidade, atribuir os seguintes apoios financeiros: ... À ARCEL - Associação Recreativa e Cultural de Espinhel, um subsídio no valor de 8.640,00€, destinado a apoiar a execução do Plano Anual e a Escola Artística"									
Translation:	"Subsequently, the Board decided unanimously to award the following financial aid: ... To ARCEL - Associação Recreativa e Cultural de Espinhel, a subsidy amounting to €8,640.00, to support the implementation of the Annual Plan and the Art School".									
NER out:	<EM C1="EM">À ARCEL</EM>- <EM C1="ORGANIZATION" C2="INSTITUTION">Associação Recreativa e Cultural de Espinhel</EM>, <EM C1="NUMBER" C2="TEXTUAL">um </EM>subsídio no valor de <EM C1="VALUE" C2="MONEY">8.640,00€</EM>, destinado a apoiar a execução do <EM C1="OBRA" C2="PLAN">Plano Anual</EM> a <EM C1="EM">Escola Artística</EM>.									
Maltparser:	1	À_Arcel	À_Arcel	prop	prop	_	0	UTT	_	_
	2	–	-	punc	punc	_	1	PUNC	_	_
	3	Associação_Recreativa...	Associação...DE_Espinhel	prop	prop	_	1	N<PRED	_	_
	4	,	,	punc	punc	_	1	PUNC	_	_
	5	um	um	art	art	_	6	>N	_	_
	6	subsídio	subsídio	n	n	_	1	N<PRED	_	_
	7	em	em	prep	prep	_	6	N<	_	_
	8	o	o	art	art	_	9	>N	_	_
	9	valor	valor	n	n	_	1	N<PRED	_	_
	10	de	de	prep	prep	_	9	N<	_	_
	11	8.640,00€	8.640,00€	num	num	_	9	N<	_	_
		(...)								
LEILA out:	503.0689326929 a_arcel 8.64000eur # 20091112171306990926.leilaout, Bridge: #3: <first> -unknown-> @dummy@ -unknown-> <second> EXAMPLE									
KB entry:	<owl:NamedIndividual rdf:about="http://mri.ieeta.pt/2010/04/07/municipality.rdf#s_8.64000eur"> <rdf:type rdf:resource="http://mri.ieeta.pt/2010/04/07/municipality.rdf#Subsidy"/> <moneyAmount>8.64000eur</moneyAmount> <assignedTo rdf:resource="http://mri.ieeta.pt/2010/04/07/municipality.rdf#a_arcel"/> <terms:isReferencedBy rdf:resource="http://mri.ieeta.pt/2010/04/07/municipality.rdf#acta_20091112171306990926"/> </owl:NamedIndividual>									

Table 1: Example fragment and output of several system modules.

2.1. Text Annotation/Enrichment

This section describes the modules that compose the first part of the system: named entity recognition (NER), POS tagging, and syntactic parsing.

2.1.1. Named Entity Recognition

The current version of NER is based on a system developed for Portuguese named Rembrandt. It uses Wikipedia as a raw knowledge resource and its document structure to classify all kinds of NEs in the text [2]. An early version of this module, designed to detect locations, was based on a small set of rules and the geographical ontology Geo-Net-PT01 [3, 4].

Rembrandt tries to classify each NE according to the Second HAREM directives [5]. Unclassified NEs are collected to be classified using other strategies. For now, the strategy is to query the Google Maps API to have the location of the NE and, if a location is retrieved, to classify it as an entity with a fixed physical location. In this case, the entity is marked as having latitude and longitude which can be an organization (enterprise or institution headquarters), a place (physical or human), or an event that happens always in the same place.

Regarding our example, the output of Rembrandt can be found in the third row of Table 1. The example shows that Rembrandt has identified six NEs and was able to classify four of them. The class is the value of C1 in the tag EM. When that value is EM (C1="EM") it means that the NE was not classified.

2.1.2. Part-Of-Speech Tagging

The POS tagging is performed by TreeTagger. It annotates text with POS and lemma information and has been successfully used to tag several natural languages including Brazilian Portuguese. TreeTagger implements a decision tree to obtain reliable estimates of context transition probabilities in order to avoid sparse-data problems, a relevant problem in statistical models training. The decision tree automatically determines the appropriate size of the context - number of surrounding words - which is used to estimate the transition probabilities [6].

The tagger was trained with a European Portuguese lexicon in order to be integrated in the system. The tagger provides tools to train a language model given three files: a corpus with tagged training data; a full form lexicon; and an open class file with the list of possible tags of unknown word forms.

The corpus used to train it - and the syntactic parser - was Bosque v7.3, the only Portuguese corpus usable to train the chosen syntactic parser. Bosque is a subset of Floresta (a publicly available treebank for Portuguese), fully revised by a linguistic team, that contains about 185,000 words [7]. The full form lexicon used in the training process was based on the computational lexicon LABEL-LEX-sw that comprises more than 1,500,000 inflected word forms, automatically generated from a lexicon of about 120,000 lemmas [8]. The output of the tagger is in the fourth (and fifth) column of the fourth row of Table 1.

2.1.3. Syntactic Parsing

The syntactic parsing is done with a data-driven dependency parser named MaltParser [9]. MaltParser was selected because there are no parsers freely available for Portuguese. It can be used to induce a parsing model from treebank data - and to parse new data using that model - and was already successfully used to parse several natural languages as English, French, Greek, Swedish, and Turkish. The parsing algorithm used was the same of the Single Malt system [9].

The parsing model was induced with the version 7.3 of

Bosque used in the Tenth Conference on Computational Natural Language Learning (CoNLL-X) shared task: multi-lingual dependency parsing. This particular version was selected because, as far as we know, is the only version - and treebank - that is available for Portuguese in the CoNLL-X format, the format accepted by MaltParser. The format defines a sentence as one or more tokens, each one starting in a new line and consisting of ten fields.

The output of the parser is presented in the fourth row of the Table 1. The value of the seventh field - HEAD - is assigned by the parser and indicates the head of the current token. For instance, the HEAD field of the 11th token (8.640,00€) is 9, which means that "8.640,00€" depends syntactically of "valor" - the 9th word which means value.

After the parsing step, the POS tag assigned to each NE is replaced by the class assigned by Rembrandt to that NE. This allows taking advantage of the information given by Rembrandt about the class of the NEs.

2.2. Relation Extraction

The general problem of interpreting text involves the determination of the semantic relations among the entities and the events they participate in. Informally, this task aims to detect elements as "who" did "what" to "whom", "when" and "where" [10].

The approach followed was to use LEILA. LEILA is a system able to extract instances of arbitrary binary relations [11]. It was chosen because it uses deep syntactic analysis to detect the relations in natural language sentences. The syntactic analysis of the original configuration is performed by a link grammar parser. The linguistic structures constructed by this parser are connected planar undirected graphs called linkages. The words of the sentence are the nodes of the graph and the edges are called links and have labels. An adapter was build for LEILA to be able to use the output of MaltParser. The adaptation was mostly straightforward and a linkage of the example sentence is represented in Figure 2.

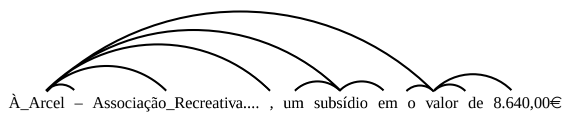


Figure 2: *The linkage of the syntactic structure of the fragment presented in Table 1.*

Each target relation of LEILA requires a function to decide into which category a pair of words (entities) falls. The pair can be: an *example* if it belongs to a list of examples; a *counterexample*, deducted from the examples if it is incompatible with the examples; a *candidate* if it obeys to some criteria and is neither an *example* nor a *counterexample*; or can be none of the previous and should be ignored. Using the output of the function and one classifier - it was used k-nearest neighbor - the core algorithm has three phases: in the *discovery phase* seeks linkages where the example pairs appear to produce positive patterns, and collects as negative patterns all linkages that match a positive pattern but produce a counterexample; in the *training phase* it uses statistical learning to produce a pattern classifier based on the patterns acquired in the discovery phase; in the *testing phase* the classifier evaluates all sentences: if a pair of entities is classified as *candidate* and the pattern connecting the entities is classified as positive, the pair is considered a new element of the target relation.

For the proof of concept and first application of the system a function was developed to detect subsidies granted by municipalities. This subject of was selected because the amount of subsidies (granted to whom) is a relevant issue in local government. The output of LEILA is at the fifth row of Table 1.

2.3. Information Integration, Management and Access

An ontology was created to define the semantics of the knowledge base. To be as standard as possible the ontology results from the usage of well known ontologies with a minimum amount of entities and properties added. It combines the ontologies Friend of a Friend (FOAF), Dublin Core, World Geodetic System (1984 revision), and GeoNames (full version), with a new class (Subsidy), a new object property (assignedTo), and a new data type property (moneyAmount).

A reasoner checks the coherence between new information and the information already in the knowledge base. The new information is added to the knowledge base when it is coherent with the existing one. Otherwise is discarded (for the moment). The reasoning is performed by an open source reasoner for OWL-DL named Pellet. It supports reasoning with individuals and user defined data types [12].

2.3.1. Geo Location

The information about entities with a fixed location (as streets, organizations headquarters, and some events) is enriched with its geocoding information. The geocodes are obtained via queries using the Google Maps API. The political organization of the spaces - street $\subset$ neighborhood $\subset$ city $\subset$ municipality ... - is obtained using a free geographic ontology of Portugal with about 418,000 features named Geo-Net-PT01 [3]. This allows the system to display the information spatially on a map and to search and relate information by its location.

2.3.2. Knowledge Base

The knowledge base is defined with the web ontology language (OWL). The storage and management is performed by Virtuoso Universal Server which features an endpoint for SPARQL.

The last row for Table 1 shows an entry of the knowledge base relative to the example. It is visible an *owl:NamedIndividual* of type *Subsidy* with some *moneyAmout*, with property *assignedTo a_arcel* and *isReferencedBy acta_2009...* (in Portuguese "acta" means minute). This entry defines a subsidy. Other existing entries (not showed here) define the minute *acta_2009...* and the NE *a_arcel*.

3. Initial Results

First Experiments were performed to extract information (for now just subsidies) from several documents belonging to a Portuguese municipality, Águeda town municipality. The experiments were conducted using 23 minutes of the Municipal Board meetings. Information regarding subsidies was extracted and subject to query and evaluation against manual annotations. Each subsidy annotation identifies the entity which received the subsidy and the amount of money involved.

3.1. Examples

It is possible to make complex queries to retrieve information not explicit (or not present) in the documents. For example, it is possible to query which Institution had the highest number of subsidies or the highest total value, or the distribution of

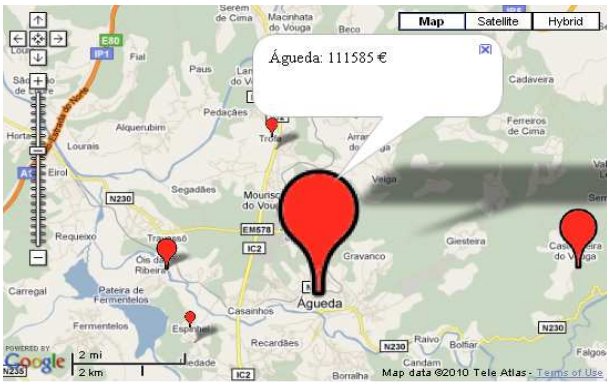


Figure 3: *Total amount of subsidies per "freguesia". The area occupied by the marker is proportional to the amount of money received. Five "freguesias" of Águeda municipality had subsidies for associations/companies/schools in the selected minutes: Águeda; Castanheira do Vouga; Espinhel; Óis da Ribeira; and Trofa.*

the total amount of subsidies granted by administrative subdivisions of the Municipality. The later was queried to the knowledge base. The query involves retrieving all subsidies of all entities which are located in one *freguesia* (neighborhood) of Águeda, group them by *freguesia* and compute the total amount per *freguesia*. After this query, the geographical location of the *freguesias* was added, and a web interface renders the result in a map as showed in Figure 3.

3.2. Performance Evaluation

A total of 107 subsidies were manually annotated in the test set. In the same set the system detected 62 subsidies - 53 well detected. The performance of the system was measured against the manual annotations. It was also measured discounting the subsidies found in enumerations (36) because enumerations like "the following subsidies were granted: *entity1 - amount1*; *entity2 - amount2*;..." were not considered for now. To detect what kind of thing is being enumerated it is necessary to have the context because enumerations (can) span across several sentences and the current algorithm processes one sentence at a time. It is being currently studied the way context tracking can be seamlessly integrated in the system. Table 2 summarizes the results.

Table 2: *The performance of the system measured for 23 documents containing 107 subsidies. Results were also compared against all subsidies that were not in enumerations (71).*

	detections				
	true	false	precision	recall	F_1
all (107)	53	9	0.85	0.50	0.63
not enum. (71)	53	9	0.85	0.75	0.80

The system achieved a precision (prec.) of 0.85 and a recall of 0.50 or 0.75 when enumerations are not considered. This performance is comparable to state of the art systems: DBpedia (prec. 0.86 to 0.99; recall 0.41 to 0.77), Kylin (prec. 0.74 to 0.97; recall 0.61 to 0.96), and YAGO/NAGA (prec. 0.91 to 0.99; recall not reported).

4. Conclusions

This article presented a first version of an IE system for natural language (government) documents in Portuguese. In this early stage of development the goal was to make a proof of concept and to perform a first evaluation. IE for e-government and in Portuguese is still a challenge because it requires research and adaptation to the specific area and language. Such systems are important because e-government own success can depend on how easy it is to maintain and use its services.

This design can be seen as a framework where different modules can be plugged in and/or switched. The current version is capable of acquiring information about subsidies and to integrate it with geographical information. The knowledge base complies with an ontology resulting form the merge of 4 public, mature, and broadly supported ontologies thus increasing interoperability. Initial results seem promising. They showed a system performance comparable with state of the art systems.

To conclude, the system works for Portuguese and was built reusing state of the art third party software, mostly developed aiming the English language. This shows that it's possible (and should be further tempted) to integrate high performance software tools designed for other natural languages.

5. Acknowledgements

The authors would like to thank Câmara Municipal de Águeda for making their documents available in digital format, and to thank Carlos Pereira for the careful annotation of them.

6. References

[1] M. Rodrigues, G. P. Dias, and A. Teixeira, "Human Language Technologies for E-Gov," in *Proc. of the 6th WEBIST*, 2010.

[2] N. Cardoso, "REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto," in *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[3] M. Chaves, M. Silva, and B. Martins, "A Geographic Knowledge Base for Semantic Web Applications," in *Proc. of SBBD*, 2005.

[4] M. Rodrigues, G. P. Dias, and A. Teixeira, "Automatic Extraction and Representation of Geographic Entities in eGovernment," in *Proc. of the 5th CISTI*, 2010.

[5] C. Mota and D. Santos, Eds., *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[6] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of International Conference on New Methods in Language Processing*, vol. 12. Manchester, UK, 1994.

[7] C. Freitas, P. Rocha, and E. Bick, "Floresta Sintá (c) tica: Bigger, Thicker and Easier," *Computational Processing of the Portuguese Language*, 2008.

[8] E. Ranchhod, C. Mota, and J. Baptista, "A Computational Lexicon of Portuguese for Automatic Text Parsing," in *Proc. of SIGLEX99: Standardizing Lexical Resources - ACL*, 1999.

[9] J. Hall, J. Nilsson, J. Nivre, G. Eryiǧit, B. Megyesi, M. Nilsson, and M. Saers, "Single Malt or Blended? A Study in Multilingual Parser Optimization," in *Proc. of the EMNLP-CoNLL*, 2007.

[10] L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson, "Semantic Role Labeling: An Introduction to the Special Issue," *Computational Linguistics*, vol. 34, no. 2, 2008.

[11] F. Suchanek, G. Ifrim, and G. Weikum, "LEILA: Learning to Extract Information by Linguistic Analysis," in *Proc. of the ACL Workshop OLP*, 2006.

[12] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A Practical OWL-DL Reasoner," *Web Semantics: science, services and agents on the World Wide Web*, vol. 5, no. 2, 2007.

Evaluation of the incremental dialogue annotation using N-gram Transducers

Carlos-D. Martínez-Hinarejos, Vicent Tamarit, José-Miguel Benedí

Instituto Tecnológico de Informática, Universidad Politénica de Valencia
Camino de Vera, s/n, 46022, Valencia, Spain
{cmartine,vtamarit,jbenedi}@dsic.upv.es

Abstract

The annotation of dialogues in terms of Dialogue Acts (DA) is an important task in the development of dialogue systems. Recently, the N-gram Transducers (NGT) technique showed a better performance than other techniques in the annotation of unsegmented dialogue transcriptions. However, this technique has not been employed in an incremental fashion, which is closer to the annotation framework. In this work, we checked the performance of NGT in this incremental framework and the influence of the size of the partitions in the effort of annotating the whole SwitchBoard corpus.

1. Introduction

One interesting application of natural language processing is dialogue systems [1]. A dialogue system is a computer system that interacts with a human user to fulfil a task whose completion requires several interactions. The behaviour of the dialogue system is defined by the dialogue strategy, which defines the reactions of the system to the user input. The user input is generally interpreted in terms of Dialogue Acts (DA) [2], which are labels that define the intention and the involved data in a subsequence of the input (usually known as segment). DA can be extended to system interactions, to reflect the actions that the system carries out.

The dialogue strategy can be based on statistical models [3], whose parameters are estimated from dialogues annotated with DA. This statistical approach is more flexible than a classic rule-based approach, but requires a large amount of annotated data to accurately estimate the models. Consequently, the annotation of a training dialogue corpus is one of the biggest efforts in the construction of a dialogue system. In the last decade, some works presented statistical models [4, 5] to speed-up this annotation process: the automatic annotation models are used to obtain a draft annotation that is corrected by the human annotator, which supposes a lower effort than annotating the dialogue from scratch. One of the most powerful annotation techniques is the NGT (N-gram Transducers) model [7], which uses an N-gram derived from the joining of words and DA and another N-gram derived from sequences of DA to obtain the DA annotation of unlabelled dialogue turns.

However, experiments reported in those previous works use a large training set and a small test set [4, 6]. In this work we present results using the NGT model in an incremental fashion, i.e., a small set of dialogues is used to train the models, another set is annotated with these models, the annotation is corrected and the corrected dialogues are added to the training set for the next step. This process is closer to the usual annotation framework. The results will show that, although there is a degradation in performance, the NGT model is still a reasonable tool to speed-up the complete annotation of a dialogue corpus. Moreover, the results demonstrate that a small amount of data annotated from scratch is more convenient, in terms of effort, than using a larger amount.

This paper is organised as follows: Section 2 provides an overview of the NGT model; Section 3 describes the corpus used for the experiments (SwitchBoard); Section 4 defines the experimental framework and shows the results of the experiments; Section 5 provides some conclusions and reveals possible work lines.

2. The NGT dialogue annotation model

The N-gram Transducers (NGT) model [7] is based on the inference of an N-gram from a set of extended symbol sequences. These extended symbols are build from the alignment of the symbols of a parallel corpus of input-output sequences. In the case of a dialogue corpus, the input symbols are the words and the output symbols are the DA labels, which are usually aligned to the last word of the segment they label. From the extended sequences an N-gram can be inferred. This N-gram can be used to process an unlabelled input sequence (sequence of words) and associate the corresponding DA labels to each possible segment. The Viterbi decoding process is shown in Figure 1.

In this labelling process each word is taken to build its corresponding tree level. Each node is expanded into o nodes, where o is the number of different outputs (DA labels) that were associated to the word in the training samples (including the empty output). The probability of each node is recalculated according to the probability of the parent node, the probability given by the N-gram of extended symbols and the probability of the associated sequence of DA labels (given by another N-gram model of DA sequences, see [7] for a detailed description of the probability computation). The output of the decoding process is the sequence of words with the corresponding attached DA labels. This provides an annotation of the input sequence (dialogue) along with its segmentation. The NGT implementation used in this work is publicly available in [8].

3. The SwitchBoard corpus

The SwitchBoard corpus [9] is a corpus of human-to-human conversations by telephone in English. It includes spontaneous speech conversations about general topics, without a clear task to complete, with frequent interruptions, background noises, hesitations and non-linguistic sounds (such like laughter). The final corpus consists of 1,155 dialogues, with approximately 115,000 turns, and a vocabulary size about 42,000 words.

This corpus was manually transcribed (including special annotation for the previously described phenomena) and annotated at the dialogue level using the SWBD-DAMSL scheme [10], a simplified version of the standard DAMSL (Di-

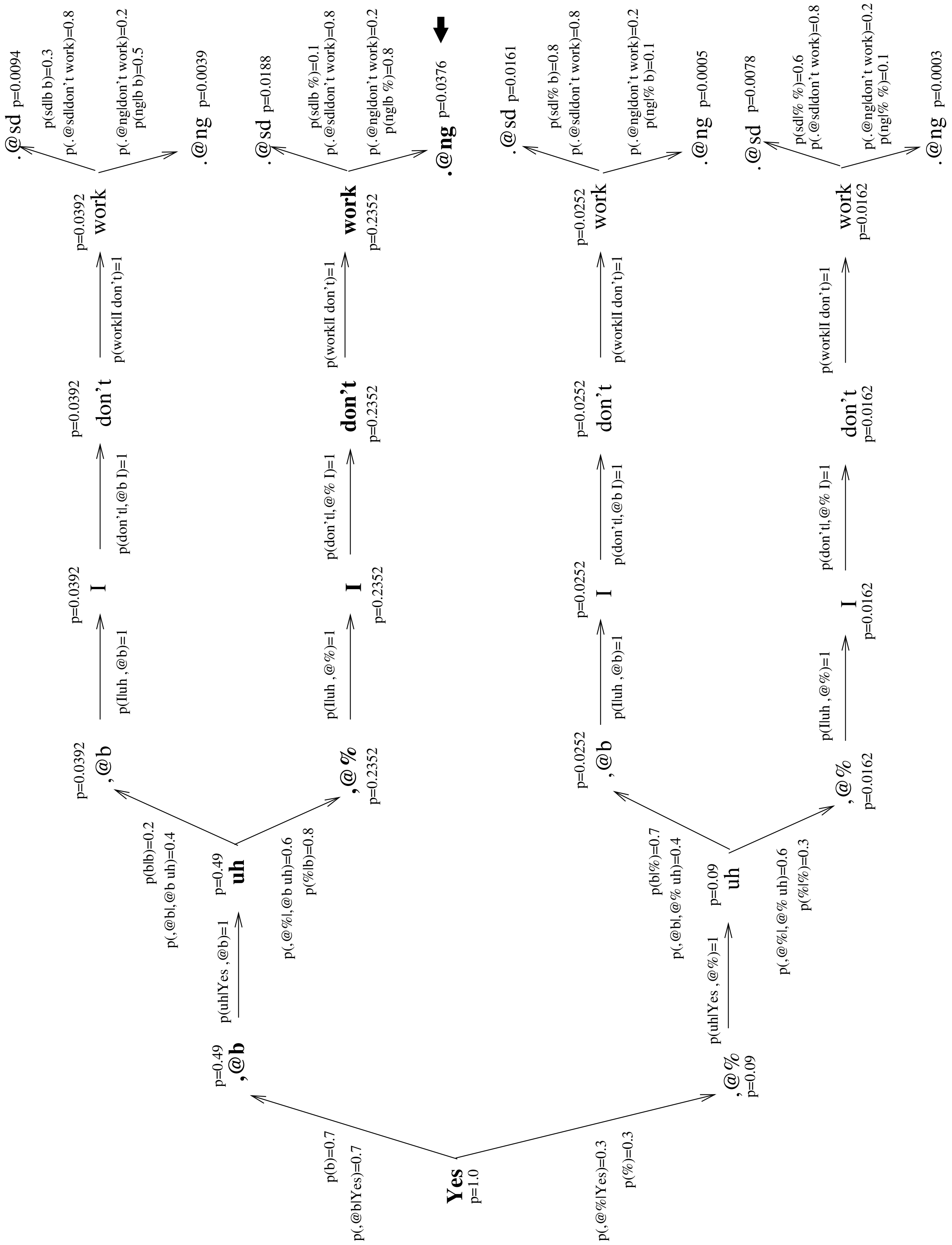


Figure 1: An example of the Viterbi tree search for the NGT model for the sentence “Yes, uh, I don’t work.”. Symbols before @ are words and symbols after @ are DA labels (in this case, *b*-backchannel, %-uninterpretable, *sd*-statement-non-opinion, and *ng*-negative-non-no-answer). Best hypothesis is in boldface and marked by the dark arrow. In this example, trigrams are used in all models.

alogue Act Mark-up in Several Layers) annotation set [11]). SWBD-DAMSL includes 42 different DA labels that represent several communicative functions, such as statement, question, backchannels, etc., and subcategories of these functions (e.g., statement opinion/non-opinion).

4. Experiments and results

The objective of our experiments is to verify the appropriateness of the NGT model for the annotation of the SwitchBoard corpus in an incremental fashion. This analysis is useful to check whether NGT is convenient to be used in an actual annotation task and to adapt the technique to an active learning [14] (the selection of the most informative samples for training) or interactive-predictive framework [13] (the use of information given by the user to obtain a better search in the model).

The annotation task usually starts from a set of transcribed dialogues on which the human annotators must place the DA labels according to a set of predefined rules. To use a statistical annotation model to obtain draft annotations, an initial set of dialogues must be annotated from scratch. The parameters of the statistical model are inferred from this initial set, and the model is applied to a new set of unlabelled dialogues. These dialogues are revised by the human annotators to correct the possible errors. The correctly annotated dialogues are added to the previously annotated set, and this new set is used to improve the estimation of the parameters of the models. This cycle is repeated until the entire set of dialogues is correctly annotated.

Consequently, the annotation framework employs an incremental training set, whose size is initially much smaller than the complete corpus and becomes larger in each cycle. This is in contrast with the approach taken by many previous works [4, 7], where training sets are usually composed of a large number of dialogues from the corpus.

In our experiments we used this incremental approach to verify the appropriateness of the NGT model. We initially compared the incremental approach with the standard cross-validation approach. To simplify the framework, we used incremental partitions of regular size on the SwitchBoard corpus. The complexity of the transcription of the SwitchBoard corpus was lowered down by removing the interruptions and overlaps (and joining the corresponding interrupted turns), transcribing the words to lowercase and separating punctuation marks.

The partitions were based on those used in the cross-validation approach. The first comparison we made was in terms of annotation error rates. In the annotation task both the correct label and the correct position are important. Consequently, we adopted the SegDAER (Segmentation and Dialogue Act Error Rate) measure. In this measure, the sequences to be compared are formed by the DA labels joined with their position in the turn. SegDAER is the average edit distance between the correct sequence and the sequence obtained by the annotation model, and was used in previous works on the measure of the quality of annotation errors [7].

In this case, we compared the results for each partition using the cross-validation approach and the incremental approach, using 11 partitions of 105 dialogues each one. This allows us to avoid the possible differences given by the specific difficulty of each partition. In the cross-validation approach, each training set is formed by 10 partitions, whereas in the incremental approach, the training set for annotating the n-th partition in the sequence is composed of the $n-1$ previous partitions.

Following the best results reported by the cross-validation approach, the experiment was performed using trigrams as NGT

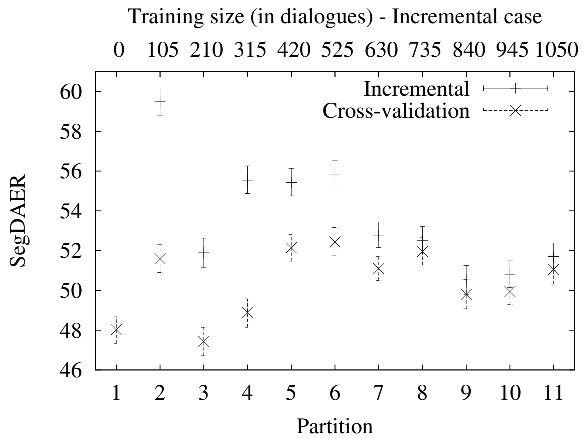


Figure 2: SegDAER comparison between the cross-validation approach and the incremental approach using 11 partitions.

model and DA language model. The results, along with the 90% confidence intervals (obtained by bootstrapping with 1000 repetitions, following the method in [12]), are presented in the graph in Figure 2. The first partition is not included in the incremental approach as its SegDAER is 100%.

From these results we can see that the absolute difference in SegDAER in each partition is lower than the 8% in all cases (and lower than 5% in most partitions). These differences become insignificant from the partition 8 (735 dialogues in the training set). Consequently, we can conclude that the NGT technique presents a moderate degradation (which disappears when using approximately 60% of the whole data) in performance when using the incremental approach, and that it is still useful in the dialogue annotation framework.

Another interesting experiment is related to the effect on the global annotation effort dependence on the partition size. In this case, we introduce a global annotation effort measure based on the SegDAER of each partition: the Relative Histogram Error Area (RHEA). This measure is based on measuring the percent of the area of the error histogram for each partition and dividing it by the area of the worst-case error histogram (all partitions present a SegDAER of 100%), i.e., RHEA=$\frac{\sum_i E_i S_i}{S_T}$, where E_i is the SegDAER of partition i, S_i is the size of partition i (the size of the individual partition, not the cummulative sum of the sizes of the previously annotated partitions) and S_T is the size of the whole corpus. The lower bound for this measure is defined by the ratio between the size of the partition annotated from scratch and the total size of the corpus (RHEA=$\frac{100 S_1}{S_T}$). The size of the partitions and the corpus can be measured in different terms, such as number of dialogues, number of turns and number of words, among others. In any case, RHEA reductions can be considered as proportional reductions in the number of errors that must be corrected by human annotators.

We computed the RHEA measure for five different sets of partitions (of 3, 5, 7, 11, and 21 partitions each set). We computed RHEA using dialogue and turn as the basic size unit for each partition. The results using 3-grams and 4-grams for each model are presented in Table 1.

From these results, a clear conclusion is that the smaller the size of the partition, the lower the annotation effort. This is congruent with the intuitive idea, as the main effort is related to the

Table 1: RHEA results for the different sets of partitions using 3 and 4-grams as NGT and DA N-grams. The size of the partitions is defined in terms of number of dialogues and number of turns.

NGT N-gram		3				4			
DA N-gram		3		4		3		4	
Num. part.	Dial/Part	Dial	Turn	Dial	Turn	Dial	Turn	Dial	Turn
3	385	69.4	69.5	69.3	69.5	71.0	71.1	71.0	71.1
5	231	62.5	63.0	63.0	63.5	64.9	65.4	64.9	65.4
7	165	60.2	61.0	60.4	61.2	62.0	62.8	62.2	62.9
11	105	57.9	58.5	58.1	58.7	60.2	60.8	59.9	60.5
21	55	56.4	56.8	56.3	56.6	58.4	58.8	58.0	58.3

annotation from scratch, and the correction of a draft annotation requires, in general, less effort. From the results we can see that the optimal combination of N-grams for the NGT and the DA N-gram depends on the size of the partitions, but differences are really small for the same NGT N-gram degree. We can see that measuring the size of the partitions in dialogues or turns does not change the optimal combination of models and the conclusions on the best partition size.

5. Conclusions and future work

In this work we presented the use of a dialogue annotation technique (NGT) in a more realistic incremental framework, in order to compare its behaviour with respect to that in the cross-validation approach. Results showed that, although there is a statistically significant increment of annotation error, it is not a dramatical increment that disregards the use of the annotation technique. We evaluated the effort of the annotation from the SegDAER of each partition, and results demonstrate that the smaller the size of the partitions, the lower the effort.

Future work is directed to studying these elements in other corpora and applying the technique in an interactive-predictive framework to make a more user-oriented evaluation [13]. Another direction is related to the use of active learning [14] to obtain an appropriate selection of the partitions to be annotated from scratch and the annotation order of the rest of the dialogues, in order to reduce the annotation effort even more.

6. Acknowledgements

This work was partially supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), the MITTRAL (TIN2009-14633-C03-01) project, and the FPI BES-2007-16834 grant. Also supported by the EC (FEDER) and the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grants Prometeo/2009/014, GV/2010/067, and ACOMP/2010/051.

7. References

[1] L. Dybkjær and W. Minker, Eds., *Recent Trends in Discourse and Dialogue*, ser. Text, Speech and Language Technology. Dordrecht: Springer, 2008, vol. 39.

[2] H. Bunt, "Context and dialogue control," *THINK Quarterly*, vol. 3, 1994.

[3] J. D. Williams and S. Young, "Partially observable markov decision processes for spoken dialog systems," *Comput. Speech Lang.*, vol. 21, no. 2, pp. 393–422, 2007.

[4] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modelling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 1–34, 2000.

[5] N. Webb and Y. Wilks, "Error analysis of dialogue act classification," in *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, 2005, pp. 451–458.

[6] C.-D. Martínez-Hinarejos, J.-M. Benedí, and R. Granell, "Statistical framework for a spanish spoken dialogue corpus," *Speech Communication*, vol. 50, pp. 992–1008, 2008.

[7] C.-D. Martínez-Hinarejos, V. Tamarit, and J.-M. Benedí, "Improving unsegmented dialogue turns annotation with n-gram transducers," in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23)*, vol. 1. Hong Kong: City University of Hong Kong Press, Dec. 2009, pp. 345–354.

[8] C.-D. Martínez-Hinarejos, "The ngt dialogue annotation software," 2010, http://users.dsic.upv.es/~cmartine/research/resources/ngt.tgz.

[9] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP-92*, 1992, pp. 517–520.

[10] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard swbd-damsl shallow- discourse-function annotation coders manual - draft 13," University of Colorado Institute of Cognitive Science, Tech. Rep. 97-01, 1997.

[11] M. G. Core and J. F. Allen, "Coding dialogues with the DAMSL annotation scheme," in *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, D. Traum, Ed. Menlo Park, California: AAAI, 1997, pp. 28–35.

[12] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *Proceedings of ICASSP'04*, vol. 1, May 2004, pp. 409–412.

[13] E. Macklovitch, "Transtype2: The last word," in *Proceedings of the 5th International Conference on Languages Resources and Evaluation (LREC 06)*, Genoa, 2006, p. 167172.

[14] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, 2009, pp. 49–56.

An Automatic Dialog Simulation Technique to Learn a Dialog Strategy for a Spoken Dialog System

David Griol[1], *Zoraida Callejas*[2], *Ramón López-Cózar*[2], *Nieves Ábalos*[2], *Gonzalo Espejo*[2]

[1]Dept. of Computer Science, Carlos III University of Madrid, Leganés (Spain)
[2]Dept. of Languages and Computer Systems, University of Granada, CITIC-UGR, Granada (Spain)
dgriol@inf.uc3m.es, {zoraida,rlopezc}@ugr.es, {nayade,gonzaep}@correo.ugr.es

Abstract

In this paper, we present a technique for learning new dialog strategies by using a statistical dialog manager that is trained from a dialog corpus. A dialog simulation technique has been developed to acquire data required to train the dialog model and then explore new dialog strategies. A set of measures has also been defined to evaluate the dialog strategy that is automatically learned. We have applied this technique to explore the space of possible dialog strategies for a dialog system that collects monitored data from patients suffering from diabetes.

Index Terms: Dialog Strategy, Dialog Simulation, Dialog Management, Dialog Systems

1. Introduction

The application of statistical approaches to dialog management has attracted increasing interest during the last decade [1]. Statistical models can be trained from real dialogs, modeling the variability in user behaviors. The final objective is to develop dialog systems that have a more robust behavior and are easier to adapt to different user profiles or tasks.

The success of these approaches depends on the quality of the data used to develop the dialog model. Considerable effort is necessary to acquire and label a corpus with the data necessary to train a good model. A technique that has currently attracted an increasing interest is based on the automatic generation of dialogs between the dialog manager (DM) and an additional module, called the user simulator, which represents user interactions with the dialog system [2].

A very important application of the simulated dialogs is to support the automatic learning of optimal dialog strategies. Large amounts of data are required for a systematic exploration of the dialog state space and corpora of simulated data are extremely valuable for this purpose, given the costs of collecting data from real users.

In this paper, we present a technique for learning optimal dialog strategies. Our technique is based on the use of a dialog simulation technique to automatically generate the data required to learn a new dialog model. We have applied our technique to explore dialog strategies for the DI@L-log dialog system, designed to collect monitored data from patients suffering from diabetes. In addition, a set of specific measures has been defined to evaluate the new strategy once new simulated data is used to re-train the dialog manager. The results of the comparison of these measures for an initial corpus and a corpus acquired using the dialog simulation technique show how the quality of the dialog is improved and the variability of the dialog model is increased by detecting new dialog situations that are not present in an initial model and new valid answers for the situations that were already present.

The remainder of the paper is organized as follows. Section 2 describes the DI@L-log system and the acquisition of a initial corpus for this task. Section 3 describes the automatic dialog generation technique that has been developed and the measures used to evaluate the quality of the dialogs with different dialog strategies. Section 4 shows the results of the comparison of these measures for the two corpora acquired for the DI@L-log task. Finally, some conclusions and future work lines are described in Section 5.

2. The DI@L-log System

DI@L-log is a spoken dialog system which acts as a voice logbook to collect home monitored data from patients suffering from Type-2 diabetes [3]. The data collected by the system are the patient's weight, blood pressure (systolic and diastolic values) and sugar levels. The system validates and analyses the data, providing some immediate feedback to the patients regarding their current progress as well as communicating the results to doctors at the hospital who are able to review the patient's progress graphically and deal with any alerts generated by the system concerning abnormal developments.

To apply our dialog simulation technique to this task, the definition of the semantics was carried out considering the information that is required to monitor the patients and inform them about their condition. Taking into account the requirements of the task, an initial strategy was designed for the DM. Figure 1 shows an example of a dialog acquired using this strategy.

As can be observed, three different phases are present in every dialog. Firstly, there is an identification phase in which the system asks the user about his login and password and then waits until the user says that he is ready to provide the control data (S1 and S2 system turns). Secondly, the system analyzes which data is required for the current user, taking into account that the weight and sugar values are mandatory and the blood control is only carried out for specific patients (S3 to S10 system turns). In this phase, the system requires the user to provide this data. Every item is confirmed after the user has provided its value. The user can only provide one item at a time. In the last phase, the system consults the information that the patient has provided during the current dialog and compares it with the data that is present in a database that contains the values that he provided in previous dialogs. By means of this comparison, the system is able to inform the user about his condition and provide him with instructions that take this into account (S11 system turn).

A corpus of 100 dialogs was acquired using this strategy.

S1: Welcome to the DI@L-log System. Please say or enter your four digit pin.
U1: 1356.
S2: Are you ready to give your measurements?
U2: Yes.
S3: Let's begin with your weight. How many stones?
U3: 12.
S4: I have understood 12 stones. Is it correct?
U4: Yes.
S5: What is your blood sugar?
U5: 11.
S6: I have understood 11. Is it correct?
U6: Yes.
S7: And finally, for your blood pressure you will be asked for the top number and then the lower number. What is the top number of your blood pressure?
U7: 160.
S8: I have understood 160. Is it correct?
U8: Yes.
S9: And what is the lower number?
U9: 86.
S10: I have understood 86. Is it correct?
U10: Yes.
S11: Your readings have been successfully updated. According to them, your weight is higher than last week, be very careful. Thanks for contacting DI@L-log. Goodbye.

Figure 1: An example of a dialog for the DI@L-log task

In order to learn statistical models, the dialogs of the corpus were labeled in terms of dialog acts. In the case of user turns, the dialog acts correspond to the classical frame representation of the meaning of the utterance. For the DI@L-log task, we defined three task-independent concepts (*Affirmation*, *Negation*, and *Not-Understood*) and four attributes (*Weight*, *Sugar*, *Systolic-Pressure*, and *Diastolic-Pressure*).

The labeling of the system turns is similar to the labeling defined for the user turns. A total of 12 task-dependent concepts was defined, corresponding to the set of concepts used by the system to acquire each of the user variables (*Weight*, *Sugar*, *Systolic-Pressure*, and *Diastolic-Pressure*), concepts used to confirm the values provided by the user (*Confirmation-Weight*, *Confirmation-Sugar*, *Confirmation-Systolic*, and *Confirmation-Diastolic*), concepts used to inform the patient about his condition (*Inform*), and three task-independent concepts (*Not-Understood*, *Opening*, and *Closing*).

3. Our Dialog Simulation Technique

Our approach for acquiring a dialog corpus is based on the interaction of a user simulator and a DM simulator [4]. Both modules use a random selection of one of the possible answers defined for the semantics of the task (user and system dialog acts). At the beginning of the simulation, the set of system answers is defined as equiprobable. When a successful dialog is simulated, the probabilities of the answers selected by the dialog manager during that dialog are incremented before beginning a new simulation.

An error simulation module has been implemented to include semantic errors in the generation of dialogs. This module modifies the frames created by the user simulator once it has selected the information to be provided to the user. In addition, the error simulation module adds a confidence score to each concept and attribute in the semantic representation obtained from the user turn. For the study presented in this paper, we have improved this module using a model for introducing errors based on the method presented in [5]. The generation of confidence scores is carried out separately from the model employed for error generation. This model is represented as a communication channel by means of a generative probabilistic model $P(c, a_u|\tilde{a}_u)$, where a_u is the true incoming user dialog act $\tilde{a}_u$ is the recognized hypothesis, and c is the confidence score associated with this hypothesis.

The probability $P(\tilde{a}_u|a_u)$ is obtained by Maximum-Likelihood using the initial labeled corpus acquired with real users and considers the recognized sequence of words w_u and the actual sequence uttered by the user $\tilde{w}_u$. This probability is decomposed into a component that generates a word-level utterance from a given user dialog act, a model that simulates ASR confusions (learned from the reference transcriptions and the ASR outputs), and a component that models the semantic decoding process.

$$P(\tilde{a}_u|a_u) = \sum_{w_u} P(a_u|\tilde{w}_u) \sum_{w_u} P(\tilde{w}_u|w_u) P(w_u|a_u)$$

Confidence score generation is carried out by approximating $P(c|\tilde{a}_u, a_u)$ assuming that there are two distributions for c. These two distributions are handcrafted, generating confidence scores for correct and incorrect hypotheses by sampling from the distributions found in the training data corresponding to our initial corpus.

$$P(c|a_w, \tilde{a}_u) = \begin{cases} P_{corr}(c) & if \quad \tilde{a}_u = a_u \\ P_{incorr}(c) & if \quad \tilde{a}_u \neq a_u \end{cases}$$

The DM simulator considers that the dialog is unsuccessful when one of the following conditions takes place: i) The dialog exceeds a maximum number of system turns slightly higher than the average number of turns of the dialogs acquired with real users; ii) the answer selected by the DM corresponds to a query not made by the user simulator; iii) the database query module generates an error because the user simulator has not provided the mandatory data needed to carry out the query; iv) the answer generator generates an error when the selected answer involves the use of a data item not provided by the user simulator. A user request for closing the dialog is selected once the system has provided the information defined in its objective(s). The dialogs that fulfill this condition before the maximum number of turns are considered successful.

3.1. Measures defined for the Evaluation

For the evaluation of the quality of the dialogs and services provided by a dialog system, we have defined a set of quantitative evaluation measures based on prior work in the dialog literature [6, 7]. This set of proposed measures can be divided into two types:

- High-level dialog features: These features evaluate how long the dialogs last, how much information is transmitted in individual turns, and how active the dialog participants are.
- Dialog style/cooperativeness measures: These measures analyze the frequency of different speech acts and study what proportion of actions is goal-directed, what part is taken up by dialog formalities, etc.

Six high-level dialog features have been defined for the evaluation of the dialogs: the average number of turns per dialog, the percentage of different dialogs without considering the attribute values, the number of repetitions of the most seen dialog, the number of turns of the most seen dialog, the number of turns of the shortest dialog, and the number of turns of the longest dialog. Using these measures, we tried to evaluate the success of the simulated dialogs as well as its efficiency and variability with regard to the different services.

For dialog style features, we define and count a set of system/user dialog acts. On the system side, we have measured the confirmation of concepts and attributes, questions to require information, and system answers generated after a database query. On the user side, we have measured the percentage of turns in which the user carries out a request to the system, provide information, confirms a concept or attribute, Yes/No answers, and other answers not included in the previous categories.

4. Evaluation Results

By employing the methodology proposed in this paper for dialog simulation, a set of 100,000 dialogs was acquired for the DI@-log task. Table 1 summarizes the statistics of the acquisition of this simulated corpus. A set of 11 different scenarios was defined to specify the objectives of the simulation, taking into account if the pressure values are necessary and different possibilities for the generation of errors and confidence measures. Given that the first and third phases of the dialog are always mandatory and have always the same structure, only the second phase in which the system collects the different values to monitor patients was taken into account for the simulation.

Simulated dialogs	$100,000$
Successful dialogs	27,521
Different dialogs	1,573

Table 1: Statistics of the corpus acquisition for the DI@L-log system

Figure 2 shows an example of a dialog from the acquired corpus. The objective defined for the dialog was to collect the weight, sugar and pressure values. The values defined in the scenario are 12, 11, 160 and 80 respectively. Confidence scores generated by the error simulator are shown between brackets. A sentence in natural language, that is equivalent to the labeled system turn selected by the DM, has been incorporated to clarify the explanation.

In this dialog, the system begins asking the user about his weight. As a low confidence measure is introduced for the value provided by the user simulator in U1, the system decides to confirm this value in S2. Then, the system asks for the sugar value. The user simulator provides this value in U3 and a high confidence measure is assigned. Therefore, this value does not need to be confirmed by the system.

The system asks for the diastolic pressure in S4. An error is introduced in the value provided by the error simulator for this parameter (it changes 160 to 150) and a low confidence measure is assigned to this value. Then, the system asks the user to confirm this value. The user simulation rejects this value in U5 and the system decides to ask for it again. Finally, the system asks for the systolic pressure. This value is correctly introduced by the user simulator and the user simulator also assigns a high confidence level. Then, the system has the data required from the patient and the third phase of the dialog carries out the analysis of the condition of the patient and informs him.

LOGIN PHASE

S1: (*Weight*)
Tell me your weight in stones?
U1: *Weight*: 12 [0.1]
S2: (*Confirmation-Weight*)
I have understood 12 stones. Is it correct?
U2: (*Acceptance*) [0.9]
S3: (*Sugar*)
What is your blood sugar?
U3: *Sugar*: 11 [0.9]
S4: (*Systolic*)
Tell me your blood systolic pressure.
U4: (*Systolic*): 150 [0.1]
S5: (*Confirmation-Systolic*)
I have understood 150. Is it correct?
U5: (*Rejection*) [0.9]
S6: (*Systolic*)
Tell me your systolic pressure.
U6: *Systolic*: 160 [0.9]
S7: (*Diastolic*)
And what is the lower number?
U7: *Diastolic*: 80 [0.9]

DATA ANALYSIS - FINAL ANSWER

Figure 2: A dialog extracted from the simulated corpus of the DI@L-log task

4.1. High-level dialog features

The first group of experiments covers the following statistical properties to evaluate the quality of the dialogs obtained using different dialog strategies: i) Dialog length, measured as the number of turns per task; number of turns of the shortest dialog; number of turns of the longest dialog; and number of turns of the most seen dialog; ii) Different dialogs in each corpus, measured as the percentage of different dialogs and the number of repetitions of the most observed dialog; iii) Turn length, measured as the number of actions per turn; iv) Participant activity, measured as the ratio between system and user actions per dialog. Table 2 shows the comparison of the different high-level measures for the initial corpus and the corpus acquired incorporating the successfully simulated dialogs.

	Initial Strategy	Final Strategy
Average number of turns per dialog	12.9±2.3	7.4±1.6
Number of different dialogs	62.9%	78.3%
Repetitions of the most seen dialog	18	3
User turns of the most seen dialog	9	7
User turns of the shortest dialog	7	5
User turns of the longest dialog	13	9

Table 2: Results of the high-level dialog features defined for the comparison of the dialogs for the initial and final strategy

The first improvement that can be observed is the reduction in the number of turns. This reduction can also be observed in the number of turns of the longest, shortest and most seen dialogs. These results show that improving the dialog strategy makes it possible to reduce the number of necessary system actions. This reduction can also be observed in the number of turns of the longest, shortest and most seen dialogs. The greater variability of the resulting dialogs can be observed in the higher percentage of different dialogs and less repetitions of the most seen dialog obtained with the final dialog strategy. We have observed that there is also a slight increment in the mean values of the turn length for the dialogs acquired with the final strategy. These dialogs are statistically longer, as they show 1.6 actions per user turn instead of the 1.3 actions observed in the initial dialogs. This is also due to the better selection of the system actions Regarding the dialog participant activity, dialogs in the final corpus have a higher proportion of system actions because the systems needs to make a smaller number of confirmations.

4.2. Dialog style and cooperativeness

The experiments described in this section cover the following statistical properties: frequency of different user and system actions (dialog acts), and proportion of goal-directed actions (request and provide information) versus grounding actions (confirmations). We consider as well the remaining possible actions. The histograms in Figures 3 and 4 show the frequency of the most dominant user and system dialog acts, respectively, in the initial and final strategy. In both cases, significant differences in the dialog acts distribution can be observed.

With regard to user actions, it can be observed that users need to employ less confirmation turns in the final strategy, which explains the higher proportion for the rest of user actions in this strategy. It also explains the lower proportion of yes/no actions in the final strategy, which are mainly used to confirm that the system's services have been correctly provided. With regard to the system actions, it can be observed a reduction in the number of system requests for data items. This explains a higher proportion of turns to inform and confirm data items in the dialogs of the final strategy. Finally, we have grouped user and system actions into categories in order to compare turns to request and provide information (goal directed actions) versus turns to confirm data items and make other actions (grounding actions). This study also shows the better quality of the dialogs and services in the final strategy, given that the proportion of goal-directed actions is higher in these dialogs.

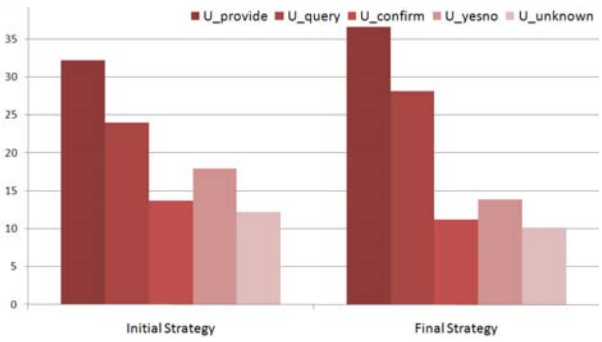


Figure 3: Histogram of user dialog acts

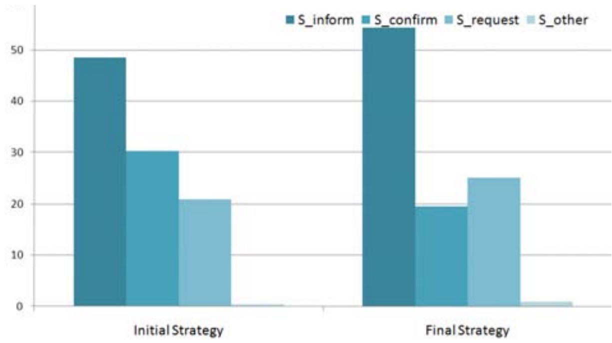


Figure 4: Histogram of system dialog acts

5. Conclusions

In this paper, we have described a technique for exploring dialog strategies in dialog systems. Our technique is based on an automatic dialog simulation technique to generate the data that is required to re-train a dialog model. The results of applying our technique to the DI@L-log system, which follows a very strict initial interaction flow, show that the proposed methodology can be used to automatically explore new enhanced strategies. Carrying out these tasks with a non-automatic approach would require a very high cost that sometimes is not affordable. As a future work, we are adapting a previously developed dialog management technique to learn a dialog manager for this task by employing the dialog corpus described in this paper and evaluate it with real users.

6. Acknowledgments

This research has been funded by the Spanish Ministry of Science and Technology, under project HADA TIN2007-64718.

7. References

[1] S. Young, "The Statistical Approach to the Design of Spoken Dialogue Systems," CUED/F-INFENG/TR.433, Cambridge University Engineering Department, Cambridge (UK), Tech. Rep., 2002.

[2] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young, "A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies," *Knowledge Engineering Review*, vol. 21(2), pp. 97–126, 2006.

[3] L. Black, M. F. McTear, N. D. Black, R. Harper, and M. Lemon, "Appraisal of a conversational artefact and its utility in remote patient monitoring," in *Proc. of the 18th IEEE Symposium CBMS'05*, Dublin, Ireland, 2005, pp. 506–508.

[4] D. Griol, L. F. Hurtado, E. Sanchis, and E. Segarra, "Acquiring and Evaluating a Dialog Corpus through a Dialog Simulation Technique," in *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, 2007, pp. 39–42.

[5] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, "Agenda-Based User Simulation for Bootstrapping a POMDP Dialogue System," in *Proc. of Human Language Technologies HLT/NAACL'07 Conference*, Rochester, USA, 2007, pp. 149–152.

[6] J. Schatzmann, K. Georgila, and S. Young, "Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems," in *Proc. of the 6th SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal, 2005, pp. 45–54.

[7] H. Ai, A. Raux, D. Bohus, M. Eskenazi, and D. Litman, "Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users," in *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, 2007, pp. 124–131.

Acquisition of synthetic dialog corpora using a task-independent dialog system

Francisco Torres Goterris

Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València / Universidad Politécnica de Valencia, Spain
ftgoterr@dsic.upv.es

Abstract

In this paper, we present the acquisition of synthetic dialog corpora through a dialog system that integrates a stochastic dialog manager and a rule-oriented user simulator. These modules are task-independent, and can be adapted to different semantic-restricted domains. Our stochastic dialog manager can interact with real or simulated users, storing automatically the acquired dialogs. In addition, the simulation mode allows us to acquire series of dialogs, verifying automatically their successful endings. These dialogs are used to adapt the stochastic dialog models and, therefore, to enhance the system in new acquisitions. This methodology has been applied to develop two dialog systems in different domains: a train services information system, and a sport booking system.

Index Terms: stochastic dialog management, user simulation, task independence, synthetic acquisition

1. Introduction

In the development of spoken dialog systems, statistical techniques have provided good results, as described in [1], [2], and [3]. However, there are some drawbacks using these techniques, as the high cost of the acquisition of the corpora and the evaluation interacting with real users. In order to overcome them, the user simulation techniques have been considered, as described in [4], [5], and [6]. In addition, significant work has been made to design dialog systems that can be easily adapted to different domains, i.e., to design task-independent dialog systems, as described in [7], and [8].

The adaptation to new tasks is one of the aims in the EDECAN project [9]. In this research frame, a dialog system has been developed to attend different semantic-restricted tasks: the BASURDE and DIHANA tasks [10], which access a train information system; and the EDECAN-SPORT task, which provides access to a sport courts booking system. In this paper, we call them BASURDE and EDECAN tasks.

In this platform, the dialog manager [11] is based on a stochastic dialog model, which is a bigram model (BM) of dialog acts, and includes a historic register (HR), which stores all the data provided in previous turns. This dialog manager follows a hybrid strategy, half stochastic (due to the use of BM), and half heuristic (due to the query of HR). It must be remarked that it can attend both tasks, just reading their corresponding bigram models, and other configuration files.

In addition, a user simulator [12] has been developed. This module allows us to acquire synthetic dialogs, learn dialog models, and evaluate the dialog system. The user simulator selects states of the same BM, and applies some heuristic rules that implement a collaborative dialog strategy. These rules are task-independent, and they serve to generate consistent dialogs, which are useful for learning dialog models, in both tasks.

During a synthetic acquisition, on the one hand, the dialog manager automatically verifies the success of the dialogs and can modify the BM, readjusting the probabilities of the transitions. On the other hand, the user simulator just provides an appropriate flow of user turns to easily generate consistent dialogs. The validity of this simulation technique has been demonstrated by testing and enhancing the dialog manager.

2. Tasks and models description

The definition of the semantics of the tasks is based on the concepts of dialog act and frame. The dialog act is the semantic unit for describing the dialog turns, and dialog acts (DAs) are used in the definition of the dialog models. The frame is the unit that structures the concepts and attributes supplied in dialog turns. Thus, our two tasks have been semantically characterized by identifying the concepts and attributes involved in dialogs with real users. Each user turn consists of utterances in which some intentions are transmitted (i.e., involved concepts) and some items of information are supplied (i.e., attributes and their values).

In the case of the BASURDE task, there was a labeled corpus from which the BM of DAs was extracted. However, at the moment of designing our system, there was not a labeled corpus of the EDECAN task. Therefore, we had to design by hand an initial BM, defining the labels of the states and setting all the transitions with the same probability. Then, applying the simulation technique, this BM was modified appropriately.

The BASURDE dialog corpus was labeled applying the concept of dialog act and a hierarchy of three levels. In this hierarchy, the first level (L1) identifies the generic dialog act; the second level (L2), the semantic of the task; and the third level (L3), the instantiated attributes. Once the dialog turns are labeled, each dialog consists of a sequence of DAs. Thus, the dialog models are structured by sequences of DAs. In the EDECAN task, although the acquired corpus had not been labeled, we have used the same methodology for defining their DAs, which could be used for its labeling.

Table 1 shows the concepts and attributes defined for coding the user turns, and Table 2 shows the labels defined for coding the DAs. There are task-independent labels and concepts, which are common in both tasks, and there are other labels, concepts and attributes that are specific to each task.

Task-independent concepts	
ACCEPTANCE, REJECTION, NOT-UNDERSTOOD	
Task-dependent concepts	**BASURDE task**
DEPARTURE-HOUR, ARRIVAL-HOUR, PRICE, TRAIN-TYPE, SERVICES	
Task-dependent concepts	**EDECAN task**
AVAILABILITY, BOOKING, BOOKED, CANCELLATION	
Attributes	**BASURDE task**
ORIGIN, DESTINATION, DEPARTURE-DATE, ARRIVAL-DATE, DEPARTURE-HOUR, ARRIVAL-HOUR, PRICE, TRAIN-TYPE, SERVICES, NUMBER-OF-TRAINS, ORDER-NUMBER	
Attributes	**EDECAN task**
SPORT, DATE, HOUR, COURT-TYPE, NUMBER-OF-COURTS, COURT-ID	

Table 1. *Concepts and attributes*

First level labeling	
OPENING, CLOSING, WAITING, NEW-QUERY, QUESTION, CONFIRMATION, ANSWER, CHOICE, ACCEPTANCE, REJECTION, NOT-UNDERSTOOD, UNDEFINED	
Second and third levels labeling	**BASURDE task**
ORIGIN, DESTINATION, DEPARTURE-DATE, ARRIVAL-DATE, DEPARTURE-HOUR, ARRIVAL-HOUR, PRICE, TRAIN-TYPE, SERVICES, NUMBER-OF-TRAINS, ORDER-NUMBER, NIL	
Second and third levels labeling	**EDECAN task**
AVAILABILITY, BOOKING, BOOKED, CANCELLATION, SPORT, DATE, HOUR, COURT-TYPE, NUMBER-OF-COURTS, COURT-ID, NIL	

Table 2. *Labels of the dialog acts*

Task-dependent concepts are the goals of the user queries. In the BASURDE task, they are DEPARTURE-HOUR, and ARRIVAL-HOUR (involved in queries about timetables), PRICE (queries about prices), TRAIN-TYPE, and SERVICES (queries about services). In the EDECAN task, they are AVAILABILITY (queries about availability of courts), BOOKING (bookings of courts), BOOKED (queries about booked courts), and CANCELLATION (cancellations of the bookings). In both tasks, the attributes are the items that the users must or can provide to specify their goals, and the system must or can supply in order to answer the queries. The attributes are specific of each task and their names are self-explanatory of their meaning.

In order to acquire dialogs, it was also necessary to define a set of task scenarios. We have defined 15 scenarios for each task, with different levels of complexity. The first and the last of them have been coded as it is shown in Table 3.

Scenario-0	BASURDE task
<DEPARTURE-HOUR> <PRICE> ORIGIN DESTINATION DEPARTURE-DATE [TRAIN-TYPE] [DEPARTURE-HOUR]	
Scenario-14	**BASURDE task**
<DEPARTURE-HOUR> <ARRIVAL-HOUR> <PRICE> <TRAIN-TYPE> ORIGIN DESTINATION DEPARTURE-DATE [DEPARTURE-HOUR] [ARRIVAL-HOUR]	
Scenario-0	**EDECAN task**
<AVAILABILITY> SPORT [COURT-TYPE] [DATE] [HOUR]	
Scenario-14	**EDECAN task**
<BOOKED> <CANCELLATION> [SPORT] [DATE] [HOUR] [<AVAILABILITY>] <BOOKING> SPORT [COURT-TYPE] DATE HOUR	

Table 3. *Codification of some scenarios*

In the case of the BASURDE task, Scenario-0 consists of a query about departure timetables and prices on a journey, in which the user must specify origin, destination, and departure date, and s/he can provide the train-type or the departure time slot. Scenario-14 is a complex query, with four user goals (arrival and departure timetables, prices and train-types), three mandatory attributes, and two optional attributes.

In the case of the EDECAN task, Scenario-0 consists of a query about availability on a certain sport, allowing the user to specify the date, the hour, and the court-type. Scenario-14 can be decomposed into three phases: (1) the user has to obtain his/her booked courts; (2) s/he has to cancel some court of the previous list, and s/he can specify the court providing the sport, the date, or the hour; and (3) s/he has to book some court providing the sport, the date, and the hour, and s/he can supply the court-type, or can make an availability query.

Thus, dialogs of complex scenarios are composed by sequences of sub-dialogs, which can share data among them. In the EDECAN task, it occurs between the BOOKED and CANCELLATION sub-dialogs, and between the AVAILABILITY and BOOKING sub-dialogs. In the BASURDE task, it occurs when there are several user goals related to the same journey.

Regarding the dialog models, we have to consider the DAs labeling of each task. Examples of labeling of segments of dialogs, one for each task, are shown in Figures 1 and 2. As it can be observed, each turn is labeled with one or more DAs.

U0: *I may want to know the train timetables from Zaragoza to Ciudad-Real on next Friday evening, or on Saturday morning.*
(U:QUESTION:DEPARTURE-HOUR:DESTINATION,DEPARTURE-DATE, DEPARTURE-HOUR,ORIGIN)
S1: *I'm sorry, but I haven't understood you. Do you want to travel from Zaragoza to Ciudad-Real on Friday the 29th of October?*
(S:NOT-UNDERSTOOD:NIL:NIL) (S:CONFIRMATION:DESTINATION, DEPARTURE-DATE,ORIGIN:DESTINATION,DEPARTURE-DATE,ORIGIN)

Figure 1. *Segment of a BASURDE dialog*

U0: *Could I book a tennis-court on next Friday?*
(U:QUESTION:BOOKING:DATE,SPORT)
S1: *Do you want to play on Tuesday?*
(S:CONFIRMATION:DATE:DATE)
U1: *No. I want to play on Friday.*
(U:REJECTION:DATE:NIL) (U:ANSWER:DATE:DATE)

Figure 2. *Segment of an EDECAN dialog*

In Figure 1, the descriptor of the U0 turn identifies a state in which the user asks for (L1: QUESTION) timetables (L2: DEPARTURE-HOUR), providing the values of four attributes (those included in L3). In the S1 turn, the system replies with two DAs: a misunderstanding (L1: NOT-UNDERSTOOD), and a confirmation of attributes (L1: CONFIRMATION), giving their values (L2 and L3: DESTINATION, DEPARTURE-DATE, ORIGIN).

In Figure 2, the user asks for (L1: QUESTION) bookings (L2: BOOKING), giving the values of two attributes (L3: DATE, SPORT). In the S1 turn, the system needs to confirm (L1: CONFIRMATION) the date, providing its value (L2 and L3: DATE). Then, the user carries out two DAs in the same turn: s/he rejects the date (L1: REJECTION), and provides other value of this attribute (L1: ANSWER).

Using these label sets, we have defined the descriptors of the dialog states. However, in the case of the BASURDE task, the reduced size of the corpus (215 dialogs) and the great number of different DAs leads to a poor estimation of these states. We found a solution by dismissing the L3 labels in defining the states. In such a case, the number of states is reduced (155 identifiers), and each one is better estimated.

The dialog states of the BASURDE model are defined by one or more descriptors that match the (US-ID:L1-ID:L2-ID) pattern, where US-ID specifies whether a user or system turn, L1-ID is one of the L1 labels, and L2-ID are one or more of the L2 labels. For instance, the U0 turn of Figure 1 is assigned to the (U:QUESTION:DEPARTURE-HOUR) dialog state. Therefore, the same state describes all the turns characterized by certain L1 and L2 labels, without considering the instantiated attributes.

In the EDECAN task, we have defined two dialog models: L2-BM and L3-BM, excluding or including the L3 labeling in the states description. For instance, the (U:QUESTION:BOOKING) descriptor identifies a L2-BM state in which the user asks for booking a court, no mattering the provided attributes; and the (U:QUESTION:BOOKING:DATE) descriptor identifies a L3-BM state of booking question specifying the date. The L2-BM has 66 states, and the L3-BM has 494 states. In both models, initially, all the transitions had the same probability.

3. Task-independent dialog platform

Our platform integrates the user and system dialog managers, the user and system language generators, and the database manager. It can also integrate understanding modules.

In a BASURDE synthetic acquisition, the understanding module receives the sentences generated by the user simulator (user dialog manager, UDM), and extracts its meaning, providing a set of user frames. Currently, none understanding module has been designed for the EDECAN task. Thus, in dialogs of this task, the UDM frames are supplied to the system dialog manager (SDM). In both tasks, there is the possibility of introducing error simulation in the user frames.

The database manager attends the queries of the SDM. The user/system language generators translate the user/system frames into sentences in natural language (currently, in Spanish and English). Both language generators work using a set of templates and rules for instantiating the templates.

In each dialog turn, the SDM reads the user frames, decides the system dialog strategy, and builds the system frames. The SDM determines its action selecting a new state in its BM, by taking into account the last user turn, the probabilities of the transitions in the BM, and the consistence of these transitions given the content of its system HR. The UDM reads the system frames, decides its action (according to its BM, its user HR, and the rules that establish a collaborative strategy for satisfying the scenarios), and builds the user frames. The SDM and UDM algorithms are task-independent. All the task information has been encapsulated into the models, the scenarios, and other configuration files. Thus, the data-structures are initialized reading these files, and the methods have been appropriately parameterized.

We have developed a JAVA dialog platform [13], according to this design of the system. By means of this platform, we can acquire dialogs for both tasks, selecting real or simulated users. In the interactive mode, any human user can give the user frames through a graphical interface, and s/he can read the system answers, carrying out whole dialogs. In the simulation mode, the dialog is completely done by the platform. This application allows us to simulate dialogs turn by turn, or whole dialogs, or series of any number of dialogs, and to specify which scenarios are simulated. In addition, the user frames can be modified by including errors in the attributes whose values are critical to the success of the dialog. Moreover, there are the training and test modes, which are used for learning and evaluating the BM.

Using this platform, we have carried out several training sets for the EDECAN task, starting from the BM described in Section 2. Different trainings have been made by enabling or disabling the error simulation (each training set contains 4,000 dialogs for each scenario, i.e., a total of 60,000 dialogs). Several test sets have been made to evaluate the learnt models (15,000 dialogs per test set). In addition, several test sets for the BASURDE task have been carried out. The platform successfully works with both tasks (achieving success rates of 0.90, in the simulation mode). Although the success rates would be lower interacting with real users, the performance seems acceptable to use it for a real acquisition.

It must be remarked that the SDM applies a hybrid dialog strategy. However, the EDECAN training starts from an initial BM, applying a heuristic strategy. To measure the quality of the learnt model, the initial BM and the learnt BM have been tested disabling the heuristic rules. In such a situation, the initial BM does not work (its success rate is 0.05), whereas the success rate is 0.43 using the learnt BM. This result is coherent with a similar experiment done for BASURDE, and confirms the utility of this technique for training the models.

4. Dialog corpora description

Our platform allows us to quickly acquire a great number of synthetic dialogs. A sample of these acquisitions can be found in [13]. Figure 3 shows a simplified segment of an acquired dialog (because of the platform stores more information of each turn). Three turns are shown: a user turn, asking for courts availability; a system turn, confirming the sport and the date; and a new user turn, in which these items are accepted. After reading the availability frame, the system transits in both models, updates its SHR, and builds a confirmation frame, according to the (S:CONFIRMATION:SPORT:DATE,SPORT) state. The SLG translates this frame into a sentence. Once the system frame is read by the UDM, this module selects transitions in its BM, and generates an acceptance frame, given that the data to confirm are right.

U0: (AVAILABILITY) 1.00
SPORT: tennis 0.32
DATE: 12-08-2010 0.77

*** **System L2-BM** transits to (U:Question:Availability)
*** **System L3-BM** transits to (U:Question:Availability:Date,Sport)
*** **System HR**:
[AVAILABILITY] confidence = 1.00 value = ???
[SPORT] confidence = 0.32 value = tennis
[DATE] confidence = 0.77 value = 12-08-2010
*** **System L2-BM** transits to (S:Confirmation:Sport)
*** **System L3-BM** transits to (S:Confirmation:Sport:Date,Sport)
S1: (CONFIRMATION) 1.00
SPORT: tennis 0.45
CURRENT-DATE: 12-08-2010 0.77
S1: On August the twelfth, do you want to play tennis?

*** **User L2-BM** transits to (S:Confirmation:Sport)
*** **User L2-BM** transits to (U:Acceptance:Sport)
U1: (ACCEPTANCE) 1.00

Figure 3. *Segment of an EDECAN synthetic dialog*

Statistics about the synthetic acquisitions, one for each task, are shown in Tables 4 and 5. The first columns on the right side (labeled *All*) show the results for all the dialogs (10,500 dialogs of each task). The other columns (labeled by number of scenario) show the results for each scenario.

In the BASURDE acquisition, an average success rate of 0.88 has been achieved introducing 1.02 errors per dialog. The average length of the dialogs is 6.07 turns. In the BASURDE corpus acquired with real users, the average length was 6.79 turns. Therefore, our platform develops dialogs of similar length. The system does an average of 2.25 answers, 2.94 confirmations, 0.46 questions, and 0.37 both question and confirmation turns, per dialog. The user simulator does an average of 2.35 questions, 0.59 answers (in which providing items of information), 1.28 acceptances, 0.68 both acceptance and answer turns (in which confirming some items, and providing other items), and 1.15 both rejection and answer turns (in which rejecting some items, and providing the right values), per dialog. There are clear correlations between the user questions and the system answers, and also between the user acceptances & rejections and the system confirmations. The transitions to other states (like the not-understood turns) are scarce (0.02 turns per dialog) due to the collaborative strategy of the user simulator. In an acquisition with real users, these states are more frequent, and the mentioned correlations between system and user turns are not as strong as here.

In the EDECAN acquisition, an average success rate of 0.92 has been achieved introducing 1.50 errors per dialog. The average length of the dialogs is 7.12 turns. The system does an average of 1.72 answers (providing courts according to the specified goals), 1.18 choices (confirming the booking or the cancellation of selected courts), 1.54 confirmations, and 2.65 questions, per dialog. The user simulator does an average of 1.72 questions, 3.84 answers (providing items of information, or selecting courts from supplied lists), 1.01 acceptances, and 0.52 both rejection and answer turns, per dialog. Again, the transitions to other states are infrequent. In the detailed information by scenarios, it can be observed more differences in the EDECAN dialogs than in the BASURDE dialogs because of the former includes more complex scenarios.

BASURDE Scenario	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**All**
Success Rate	**0.92**	**0.88**	**0.91**	**0.90**	**0.90**	**0.88**	**0.81**	**0.89**	**0.87**	**0.93**	**0.82**	**0.88**	**0.86**	**0.89**	**0.84**	**0.88**
Number of Turns	**5.80**	**5.75**	**5.83**	**6.98**	**6.86**	**6.91**	**5.14**	**6.98**	**5.70**	**4.16**	**5.09**	**6.30**	**5.69**	**6.95**	**6.91**	**6.07**
System Turns																
Answer	1.92	1.84	1.90	2.89	2.43	2.88	1.82	2.89	1.80	1.00	1.83	2.87	1.81	2.89	3.05	**2.25**
Confirmation	3.01	2.83	3.05	3.18	3.78	3.06	2.48	3.14	2.83	2.53	2.37	2.77	2.77	3.14	3.18	**2.94**
Question	0.51	0.62	0.55	0.51	0.20	0.55	0.42	0.54	0.62	0.35	0.44	0.32	0.66	0.54	0.11	**0.46**
Quest. & Confirm.	0.33	0.43	0.31	0.38	0.29	0.39	0.40	0.38	0.43	0.27	0.42	0.32	0.42	0.36	0.49	**0.37**
Others	0.00	0.01	0.00	0.00	0.14	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.06	**0.02**
User Turns																
Answer	0.64	0.74	0.64	0.64	0.29	0.67	0.56	0.67	0.74	0.41	0.54	0.39	0.78	0.66	0.41	**0.59**
Acceptance	0.95	1.17	1.01	1.35	2.25	1.26	1.27	1.33	1.17	0.87	1.30	1.38	1.17	1.35	1.40	**1.28**
Accept. & Answer	0.69	0.65	0.65	0.64	0.84	0.65	0.71	0.65	0.64	0.70	0.70	0.75	0.61	0.62	0.65	**0.68**
Reject. & Answer	1.39	1.33	1.43	1.43	0.47	1.41	0.77	1.41	1.32	0.60	0.70	0.87	1.28	1.40	1.40	**1.15**
Question	2.11	1.84	2.08	2.90	2.86	2.89	1.82	2.91	1.80	1.56	1.83	2.88	1.81	2.90	3.03	**2.35**
Others	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	**0.01**

Table 4. *Statistics of a synthetic acquisition in the BASURDE task*

EDECAN Scenario	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	**All**
Success Rate	**0.90**	**0.91**	**0.87**	**0.91**	**0.91**	**0.93**	**0.91**	**0.96**	**0.95**	**0.90**	**0.91**	**0.93**	**0.96**	**0.89**	**0.89**	**0.92**
Number of Turns	**3.31**	**4.51**	**10.3**	**3.27**	**5.87**	**10.4**	**4.29**	**4.54**	**8.20**	**7.14**	**9.10**	**12.5**	**7.02**	**8.07**	**7.94**	**7.12**
System Turns																
Answer	1.00	1.00	1.98	1.00	1.00	1.98	1.00	2.00	2.00	1.99	1.99	2.99	2.00	1.99	1.98	**1.72**
Confirmation	1.06	1.18	1.94	1.05	0.96	1.56	1.10	1.21	1.65	1.94	1.67	2.04	1.87	1.91	1.92	**1.54**
Question	1.25	1.33	4.47	1.21	2.90	4.93	1.19	1.32	3.54	2.20	3.49	5.51	2.14	2.17	2.03	**2.65**
Choice	0.00	0.99	1.96	0.00	0.99	1.96	0.99	0.00	0.98	0.99	1.96	1.98	0.99	1.97	1.96	**1.18**
Others	0.00	0.00	0.02	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.02	0.01	0.00	0.02	0.02	**0.01**
User Turns																
Answer	1.25	2.33	6.45	1.21	3.91	6.90	2.18	1.32	4.53	3.20	5.47	7.50	3.14	4.15	4.01	**3.84**
Acceptance	0.58	0.69	1.35	0.65	0.65	1.16	0.65	0.67	1.13	1.27	1.15	1.50	1.21	1.26	1.32	**1.01**
Reject. & Answer	0.48	0.49	0.59	0.40	0.30	0.40	0.44	0.54	0.51	0.67	0.52	0.54	0.65	0.65	0.60	**0.52**
Question	1.00	1.00	1.98	1.00	1.00	1.98	1.00	2.00	2.00	1.99	1.99	2.99	2.00	1.99	1.98	**1.72**
Others	0.00	0.00	0.01	0.00	0.00	0.02	0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.01	0.02	**0.01**

Table 5. *Statistics of a synthetic acquisition in the EDECAN task*

5. Conclusions

In this paper, the acquisition of synthetic dialog corpora by means of a task-independent dialog platform has been discussed. This dialog platform allows us to carry out real and simulated dialogs, to acquire synthetic corpora, to learn dialog models, and to evaluate the system using these models.

We have integrated different techniques: stochastic dialog management, rule-oriented user simulation, template-based language generation, and task parameterization. Thus, we have faced the known problems in the corpora acquisition and the systems evaluation. Following the proposed methodology, we have developed a dialog platform that facilitates the evaluation for different tasks, whether initial dialog corpora exist or not.

The results are enough satisfactory as to consider using the platform in an acquisition with real users. Future work will be oriented to acquire real user dialog corpora for the considered tasks, and to extend its use to other semantic domains.

6. Acknowledgements

This work has been partially funded by CICYT under project TIN2008-06856-C05-02/TIN, Spain.

7. References

[1] Levin, E., Pieraccini, R., Eckert, W., "A stochastic model of human-machine interaction for learning dialog strategies", IEEE Trans. on Speech and Audio Processing, 8(1):11–23, 2000.

[2] Young, S., "The statistical approach to the design of spoken dialogue systems", Cambridge University, Tech. Report, 2002.

[3] Griol, D., et al., "A statistical approach to spoken dialog systems design and evaluation", Speech Communication, 50, 666–682, 2008.

[4] Eckert, W., Levin, E., Pieraccini, R., "User modeling for spoken dialogue system evaluation", Proc. of ASRU – IEEE Workshop, Santa Barbara, USA, 1997.

[5] López-Cózar, R., et al., "Assessment of dialogue systems by means of a new simulation technique", Speech Communication 40, 387–407, 2003.

[6] Jung, S., et al., "Data-driven user simulation for automated evaluation of spoken dialogue systems", Computer Speech & Language 23, 479–509, 2009.

[7] Lemon, O., et al., "Multi-tasking and collaborative activities in dialogue systems", Proc. of SIGdial Workshop, Philadelphia, USA, 113–124, 2002.

[8] Bohus, D., Rudnicky, A.I., "The RavenClaw dialog management framework: Architecture and systems", Computer Speech & Language, 23, 332–361, 2008.

[9] Lleida, E., et al., "EDECÁN: sistema de diálogo multidominio con adaptación a contexto acústico y de aplicación", Jornadas en Tecnología del Habla (JTH), Zaragoza, Spain, 291–296, 2006.

[10] Benedí, J.M., et al., "Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA", Proc. of 5th LREC, Genove, Italy, 1636–1639, 2006.

[11] Torres, F., et al., "Error handling in a stochastic dialog system through confidence measures", Speech Communication 45, 211–229, 2005.

[12] Torres, F., Sanchis, E., Segarra, E., "User simulation in a stochastic dialog system", Computer Speech & Language, 22, 230–255, 2008.

[13] Torres, F., "Prototype of the dialog system", available in website http://www.laesteladetanit.es/prototype.htm, 2010.

HMM-based Speech Synthesis in Basque Language using HTS

D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, I. Hernáez

AHOLAB Signal Processing Laboratory, University of the Basque Country, Bilbao
derro@aholab.ehu.es

Abstract

This paper shows how an HMM-based speech synthesizer in Basque language has been built using HTS and AhoTTS (the TTS system developed at Aholab). The resulting system, which is being used only for research purposes at present, has a highly satisfactory performance.

Index Terms: statistical parametric speech synthesis, hidden Markov models, Basque language

1. Introduction

Speech synthesis systems based on hidden Markov models (HMMs) [1] are gaining ground over unit selection based systems [2][3], which had been dominant during many years, as confirmed by the results of the last editions of the Blizzard Challenge [4]. A similar conclusion could also be drawn from the last Albayzin Evaluation [5]. Such systems model the acoustic characteristics of the speaker, given by a framewise parametric representation of the spectrum and the excitation, using multi-stream context-dependent HMMs (CD-HMMs) trained on a corpus. During synthesis, given the phonetic/prosodic context of an input text, a single sentence-HMM is built from the trained CD-HMM set, and the system returns the sequence of parameter vectors whose likelihood with respect to the model is maximal. The synthetic utterances are reconstructed by inverse parameterization. The main advantage of HMM-based systems is their flexibility: the trained models can be adapted to generate speech with different voices, speaking styles, emotions, etc.

The level of popularity achieved by this synthesis technology during the last decade is closely linked to the release of the HMM-based Speech Synthesis System (HTS) [6][7]. A number of improvements on the basic system introduced along these years (multi-space distribution for f0 modelling [8], trajectory modeling through explicit relationships between statics and dynamics [9], explicit state duration distributions [10], parameter generation considering global variance [11], strong vocoding techniques [12], etc.) have made the performance of HTS very satisfactory. In view of this great development of statistical parametric speech synthesis, many research groups around the world have built synthesizers based on HTS in more than 30 different languages and dialects. Please refer to [1] for a complete list of languages. Regarding Iberian Languages, HMM-based speech synthesizers have been already built in Castilian Spanish [13][14], Catalan [15], and Portuguese [16].

This paper presents a new language to be incorporated to that list: Basque language, which is spoken by more than 800K speakers in northern Spain and southern France. The system described here results from the combination of HTS and AhoTTS, the text-to-speech (TTS) synthesis system developed at Aholab Signal Processing Laboratory [17][18]. The rest of the paper is structured as follows. Section 2 shows a detailed explanation of the system. Several aspects regarding its performance are discussed in section 3, and some conclusions and future works are listed in section 4.

2. From AhoTTS to Aho-HTS

2.1. Brief description of AhoTTS

AhoTTS is the multiplatform modular TTS synthesis system being developed at Aholab since 1997. Although it was conceived as a multilingual system (up till now, a number of voices have been built in Basque [17], Spanish [18] and English [19]), special emphasis has been placed on Basque language, for which AhoTTS is the reference system in the world. AhoTTS consists of three basic modules: 1) text and linguistic processing, 2) prosody prediction, and 3) waveform generation. Next, each of these modules is briefly described.

The linguistic module reads the input text and generates the corresponding sequence of phonemes. Moreover, it provides information at different linguistic levels. The tasks carried out by this first module are: normalization, sentence delimitation, part-of-speech tagging, syllabification, stress marking, and phonetic transcription.

The prosodic module uses the linguistic and phonetic information provided by the previous module to generate a prosodic contour (at three levels: intonation, durations, and energy) suitable for the sentence to be spoken by the system. Regarding intonation, three different strategies have been implemented until now: a very simple peak-valley model, a more sophisticated model based on trees and Fujisaki curves [20], and corpus-based contour selection [3]. Durations are predicted using classification and regression trees (CARTs) [21].

The waveform generation module takes the information provided by the two previous modules as input and yields the final acoustic signal. The current implementation of AhoTTS applies the unit selection technique [2].

According to the described architecture, extending the system to adopt the statistical parametric synthesis paradigm implies replacing the second and third modules by HTS itself. Note that HTS is capable of generating both the prosody and the spectrum of speech from the trained acoustic models, whereas it does not perform any kind of linguistic analysis. Therefore, in this case, the role of AhoTTS is supplying the context labels required by HTS to generate the synthetic waveforms. In other words, the output of the first module of AhoTTS has to be translated into labels containing phonetic and linguistic information.

2.2. Some comments on Basque language

As well as in other languages, the linguistic information of Basque is allocated at different levels, namely, phonemes, syllables, words, accent groups, phrases, sentences… The accent group can be defined as a set of syllables pronounced around one accented syllable [22]. In several languages, particularly in the Iberian ones [23][24][25][26], the accent group is formed by words, one of them having the accent. Due to the inflectional and agglutinative nature of Standard Basque (the grammatical relations between components within a clause are represented by suffixes, and many words are formed by joining morphemes together [27]), the accent groups in this language are very often constituted by just one word. Only in some cases the accent group includes succeeding words such as short auxiliary verbs, demonstratives and some numerals. However, the possible redundancy at these linguistic levels is not harmful for the performance of the system, as only the most discriminative information is taken into account by HTS when training the CD-HMMs. On the other hand, considering the accent group level eases extending the application domain of the system to other Iberian languages such as Spanish using the same information sources.

Another consequence of inflection and agglutination is the appearance of long words showing more than one accent in natural spoken sentences. Dealing with that kind of words is not straightforward. In addition, the high dialectal fragmentation of Basque (it has seven main dialects and more than 50 varieties according to modern commonly accepted assumptions) increases the intonation variability. Therefore, multiple accents are not considered in this work. Instead, we assume that the system is capable of learning secondary accent patterns from acoustic data and other existing labels (for instance, those related to the position of the syllable in the word).

2.3. Generation of context labels using AhoTTS

Among the features provided by the linguistic module of AhoTTS, the ones that have been encoded into the context labels used by HTS are the following:

- **Phoneme level**:
  - SAMPA label of the current phoneme.
  - Labels of 2 phonemes to the right and 2 phonemes to the left.
  - Position of the current phoneme in the current syllable (from the beginning and from the end).
  - Position of the current phoneme after the previous pause and before the next pause.
- **Syllable level**:
  - Number of phonemes in current, previous and next syllables.
  - Accent in current, previous and next syllables.
  - Stress in current, previous and next syllables.
  - Position of the current syllable in the current word (from the beginning and from the end).
  - Position of the current syllable in the current accent group.
  - Position of the current syllable in the current sentence.
  - Position of the current syllable after the previous pause and before the next pause.
- **Word level**:
  - Simplified part-of-speech tag of the current, previous and next words (content/function).
  - Number of syllables of the current, previous and next words.
  - Position of the current word in the sentence (from the beginning and from the end).
  - Position of the current word after the previous pause and before the next pause.
- **Accent group level**:
  - Type of current, previous and next accent groups, according to the accent position.
  - Number of syllables in current, previous and next accent groups.
  - Position of the current accent group in the sentence (from the beginning and from the end).
  - Position of the current accent group after the previous pause and before the next pause.
- **Pause context level**:
  - Type of previous and next pauses.
  - Number of pauses to the right and to the left.
- **Sentence level**:
  - Type of sentence.
  - Number of phonemes.
  - Number of syllables.
  - Number of words.
  - Number of accent groups.
  - Number of pauses.
  - Emotion of the sentence.

3. Performance Results

In order to evaluate the performance of the system, the naturalness of the synthetic utterances was measured by means of a mean opinion score (MOS) test. The database used for this evaluation consisted of 2K short sentences (around 2 hours of speech) spoken by a Basque female speaker in neutral style. Eighteen volunteer listeners (six among them were familiar with speech synthesizers to some extent) were asked to listen to ten different synthetic utterances (the texts to be spoken by the system were taken from newspapers) and rate their naturalness in a 1-to-5 MOS scale, where 1 point means "very low naturalness" and 5 points means "very high naturalness". The state-of-the-art Straight-based vocoder was used to translate the speech frames into f0, 40 MFCCs, and 5 band-aperiodicity coefficients, which were used to feed the system during training. The spectral envelope and the aperiodic component (together with their first and second derivatives) were modeled using continuous-density CD-HMMs, whereas f0 was modeled by means of multi-space probability distributions, following the specifications of the demo scripts supplied under the current HTS distribution (available at [28]). Natural speech and unit selection based synthetic speech were evaluated together with the described system.

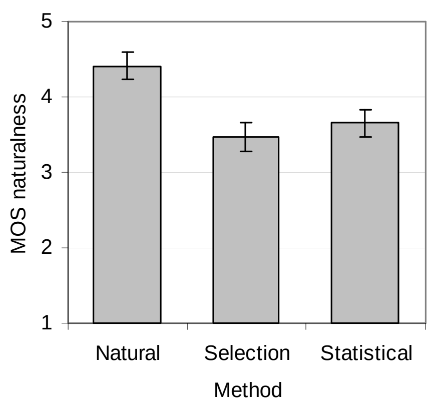


Figure 1: *Naturalness MOS achieved by natural voice, unit selection synthesis and statistical synthesis at 95% confidence intervals.*

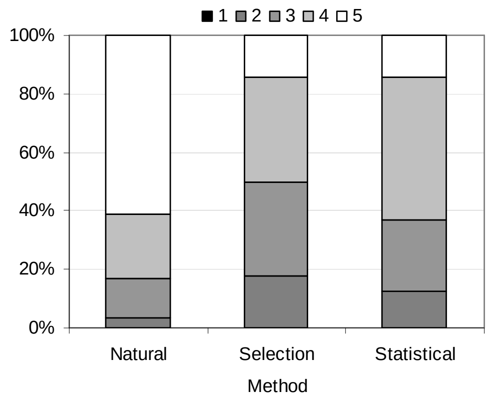


Figure 2: *Distribution of the scores in the MOS test.*

The naturalness MOS achieved by the system was 3.7, slightly higher than that of the unit selection synthesizer, although the confidence intervals overlap (see Figure 1). A deeper analysis considering some individual sentences revealed that, despite vocoding, the former was preferred by listeners when the concatenation artifacts were somewhat audible, which is coherent with the results of the last Albayzin evaluation campaign [5]. Regarding the score distributions in Figure 2, only 12% of the scores given to the statistical system were lower than 3. This confirms that HTS is capable of generating speech in a very stable and pleasant manner, as reported for many other languages [1]. The unit selection based synthesizer was given more “5s” than the statistical one, which is also coherent with previous studies and Blizzard Challenge evaluations.

Table 1 shows to what extent the nodes of the trees capturing the context dependency were related to each of the context levels. It can be seen that phonemes and syllables contain a high percentage of the relevant information. With regard to words and accent groups, they both play an important role in prosody. However, as they appear at earlier nodes of the trees, accent groups are found to carry more important information than Basque words, even if they often coincide, as explained in section 2.

Table 1. *Percentage of tree nodes related to each level of the context labels. Inside the parenthesis: first level of ramification where they appeared.*

	Spectrum	Pitch	Duration
Phoneme	93.87 (0)	43.56 (0)	72.94 (0)
Syllable	3.54 (3)	24.97 (1)	12.94 (2)
Word	0.38 (5)	9.14 (3)	4.51 (3)
Acc. group	0.63 (4)	12.29 (0)	5.49 (0)
Pause ctxt.	0.94 (2)	2.58 (2)	0.98 (1)
Sentence	0.64 (4)	7.47 (2)	3.14 (4)

4. Conclusions

An HMM-based speech synthesis system in Basque language has been built using HTS and the linguistic analysis module of the AhoTTS synthesizer. Although affected by typical limitations of statistical synthesis, the generated speech was considered quite natural by the listeners.

Research is currently being done towards the design of new vocoding techniques that allow generating synthetic speech at a higher quality.

5. Acknowledgements

This work has been partially supported by UPV/EHU (Ayuda de Especialización de Doctores), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Berbatek, IE09-262).

6. References

[1] H. Zen, K. Tokuda, A.W. Black, “Statistical parametric speech synthesis”, Speech Communication, vol.51, no.11, pp.1039-1064, 2009.

[2] A. Hunt, A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, Proc. ICASSP, vol. 1 pp. 373-376, 1996.

[3] A. Raux, A. Black, “A unit selection approach to F0 modeling and its application to emphasis”, Proc. ASRU, pp. 700- 705, 2003.

[4] A.W. Black, K. Tokuda, "The Blizzard Challenge – 2005: evaluating corpus-based speech synthesis on common datasets", Proc. Interspeech, pp.77-80, 2005.

[5] I. Sainz, E. Navas, I. Hernáez, A. Bonafonte, F. Campillo, “TTS Evaluation Campaign with a Common Spanish Database”, Proc. 7th International Language Resources and Evaluation Conference, pp. 2155-2160, 2010.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, Proc. Eurospeech, pp.2347–2350, 1999.

[7] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, “The HMM-based speech synthesis system version 2.0”, Proc. 6th ISCA Speech Synthesis Workshop, 2007.

[8] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, “Multi-space probability distribution HMM”, IEICE Trans. Inf. Syst. E85-D (3), pp.455–464, 2002.

[9] H. Zen, K. Tokuda, T. Kitamura, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences”, Computer, Speech and Language, vol.21(1), pp.153–173, 2006.

[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “A hidden semi-Markov model-based speech synthesis system”, IEICE Trans. Inf. Syst. E90-D (5), pp.825–834, 2007.

[11] T. Toda, K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, IEICE Trans. Inf. Syst. E90-D (5), pp.816–824, 2007.

[12] H. Zen, T. Toda, M. Nakamura, K. Tokuda, “Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005”, IEICE Trans. Inf. Syst. E90-D (1), pp.325–333, 2007.

[13] X. Gonzalvo, J.C. Socoro, I. Iriondo, C. Monzo, E. Martinez, “Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish”, Proc. 6th ISCA Speech Synthesis Workshop, pp. 362–367, 2007.

[14] R. Barra-Chicote, J. Yamagishi, J.M. Montero, S. King, S. Lufti, J. Macías-Guarasa, “Generación de una voz sintética en castellano basada en HSMM para la evaluación Albayzin 2008: conversión texto a voz”, Proc. V Jornadas en Tecnología del Habla, pp.115-118, 2008.

[15] A. Bonafonte, L. Aguilar, I. Esquerra, S. Oller, A. Moreno, “Recent work on the FESTCAT database for speech synthesis”, I Joint SIG-IL / Microsoft Workshop on Speech and Language Technologies for Iberian Languages, 2009.

[16] M. Barros, R. Maia, K. Tokuda, D. Freitas, F. Resende Jr., “HMM-based European Portuguese speech synthesis”, Proc. Interspeech, pp.2581–2584, 2005.

[17] I. Hernaez, E. Navas, J.L. Murugarren, B. Etxebarria, “Description of the AhoTTS system for the Basque language”, Proc. 4th ISCA Speech Synthesis Workshop, 2001.

[18] I. Sainz, I. Hernáez, E. Navas, J. Sanchez, I. Luengo, I. Saratxaga, I. Odriozola, E. de Bilbao, D. Erro, “Descripción del Conversor de Texto a Voz AhoTTS Presentado a la Evaluación

Albayzin TTS 2008", Proc. V Jornadas en Tecnología del Habla, pp.96-99, 2008.
[19] I. Sainz, D. Erro, E. Navas, I. Hernáez, I. Saratxaga, I. Luengo, I. Odriozola, "The AHOLAB Blizzard Challenge 2009 Entry", Blizzard Challenge 2009 Workshop, 2009.
[20] E. Navas, I. Hernaez, J. Sanchez, "Subjective evaluation of synthetic intonation", Proc. IEEE Workshop on Speech Synthesis, pp.23-26, 2002.
[21] E. Navas, I. Hernáez, J. Sánchez, "Predicting Segmental Durations for Basque Using CARTs", Proc. 15th International Congress of Phonetic Sciences, pp.2083-2086, 2003.
[22] B. Möbius, M. Pätzold, W. Hess. "Analysis and synthesis of German F0 contours by means of Fujisaki's model". Speech Communication, vol.13, pp. 53-61, 1993.
[23] D. Escudero, "Modelado Estadístico de Entonación con Funciones de Bézier: Aplicaciones a la Conversión Texto-Voz en Español", PhD thesis, Universidad de Valladolid, 2002.
[24] E. Navas, "Standard Basque Prosodic Modeling for Text to Speech Conversion", PhD thesis, University of the Basque Country, 2003.
[25] P.D. Agüero, J. Adell, A. Bonafonte, "Prosody Generation for Speech-to-Speech Translation", Proc. ICASSP, pp.557-560, 2006.
[26] F. Campillo, E.R. Banga, "A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems", Speech Communication, vol.48, pp.941-956, 2005.
[27] J.I. Hualde, J. Ortiz De Urbina (Eds.), "A Grammar of Basque", Mouton de Gruyter, Berlin, 2003.
[28] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

Cross-lingual ToBI accent tones identification: preliminary results

David Escudero-Mancebo, César González-Ferreras, Carlos Vivaracho-Pascual, Valentín Cardeñoso-Payo

ECA-SIMM Laboratory, University of Valladolid, Spain
{descuder,cesargf,cevp,valen}@infor.uva.es

Abstract

This paper tackles the problem of corpus based ToBI symbol automatic identification. We focus on the binary decision of accent vs. no accent tone presence. We test a cross-lingual alternative to identify accents in a given language using supervised data learning tools trained with data of a different language. A multilayer perceptron and a C4.5 decision tree have been trained with the English Boston Radio News corpus and we test its capabilities on predicting accents in a Spanish corpus. Results are promising leading us to discuss on the application of previous work on ToBI accents multiclass classification.

Index Terms: prosody, ToBI, crosslingual, automatic recognition of prosody

1. Introduction

ToBI is a standard for representing and labelling prosodic events including tones (accent tones and boundary tones) and breaks. The tones level is used to mark the occurrence of phonological tones at appropriate points in the F0 contour. The break level is used to mark break indices, which are numbers representing the strength of the boundary between two orthographic words. The tones codification is based on the combination of two single symbols: H (high) and L (low). One of the most important prosodic features is prominence: a word or part of a word made prominent is perceived as standing out from its environment [1]. This paper focuses on this particular aspect of prosody that is also marked in the ToBI prosodic representation model[2].

ToBI has been implemented for several languages including English, German and Japanese. Despite the intesive research activity for Iberian languages; the need of a reference corpus similar as the ones existing for other languages (e.g. the Boston Radio Corpus for English [3]) is still a need both for Catalan and Spanish. The activity presented in this paper is included in the Glissando project [1], that has the aim to record and label with ToBI marks a bilingual Spanish and Catalan corpus that contains Radio news recordings and spontaneous dialogs.

Labelling a corpus with ToBI tags is an expensive procedure. In [4] it is estimated that the ToBI labelling commonly takes from 100-200 times real time. To speed up the process, automatic or semiautomatic methods seem to be a productive resource. [5] or [6] are good examples of the state of art on automatic labelling of ToBI events. For Catalan [7] presents a procedure to label break indices reducing the set of break indices merging together some of them with the aim to increase the identification results. This merging strategy is common in other studies such the ones already mentioned of [5] or [6] that combine the different type of accent tones transforming the labelling problem into a binary one to decide whether an accent is present or not.

[1] Partially funded by the Ministerio de Ciencia e Innovacion, Spanish Government Glissando project FFI2008-04982-C003-02

Here we explore a cross-lingual approach where a given corpus with ToBI labels will be used to predict the labels of a different corpus in a different language. Despite the ToBI sequences are highly dependent on the language, they codify universal functions of prosody, one of them the prominence. Thus we use the Boston Radio Corpus to train prosodic models that are used then to identify the prominence in a Spanish corpus. This cross-lingual approach is pertinent as the number of linguistic resources with ToBI labels is sparse and the number of languages that lack of this information is large.

In [8] we point out data sparseness, the high inter-symbols similarity and the large number of prosodic features potentially affecting prosodic profiles as the main difficulties for ToBI labelling automatic approximations. Here we add the normalization of the prosodic features as the challenging problem to cope with.

First we present the experimental procedure and then we present the results on crosslingual accent identification. Discussion of the future work to extend the approach to different accent type is then presented.

2. Processing of the corpus

We used the Boston University Radio News Corpus [3]. This corpus includes labels separating phonemes, syllables and words. Accents are marked with a ToBI label and a position. We take into account the 7 more frequent types of accent tones: H*, L+H*, !H*, H+!H*, L+!H*, L*, and L*+H discarding other undetermined marks like * or *?. Inspired in previous works [9, 5] we aligned the accent tones with respect to the prominent syllable and to the word that contains it (words with more than one label are discarded in this work). All the utterances in the corpus with TOBI labels, from all the speakers (f1a, f2b, f3a, m1b, m2b and m3b) have been used, as shown in table 1.

The Spanish corpus used in this paper is ESMA-UPC. It was designed aiming the construction of a unit concatenative TTS system for Catalan and Spanish at the UPC (http://www.gps-tsc.upc.es) [10]. It contains three hours recordings of spoken utterances in both languages. Although it was not specifically designed for prosodic studies, it contains enough data to get significant results. The corpus was acquired under recording studio conditions in two separate channels at 32 kHz. Speech was recorded in one of the channels and the output of a laryngograph in the other. Data were automatically labelled and manually supervised. Labelling included silences, allophonic transcription, and allophonic boundaries. This information was increased by the additional syllable and word boundaries and stress positions. Pitch was estimated by means of glottal pulses closing time points. It eases the automatic segmentation of stress groups and the selection of the corresponding F0 profiles. Figure 2 resumes the figures of this corpus.

Similar features to other experiments reported in the bib-

	word	syllable
# utterances	421	421
H*	7587	8098
L+H*	2383	2501
!H*	2144	2358
H+!H*	586	654
L+!H*	638	666
L*	517	548
L*+H	44	48
none	13868	32450
Total	27767	47323

Table 1: Accent events in the Boston Corpus

# utterances	421
# accent groups	
Accent	7587
No accent	2383
Total	9970

Table 2: Accent events in the UPC-ESMA Corpus

liography [5] have been used. They concern to frequency: within word F0 range, difference between maximum and average within word F0, difference between average and minimum within word F0, difference between within word F0 average and utterance average F0; to energy: within word energy range, difference between maximum and average within word energy, difference between average and minimum within word energy; to duration: maximum normalized vowel nucleus duration from all the vowels of the word (normalization is done for each vowel type); and to grammatical information POS: part of speech.

3. Experimental procedure

3.1. Experimental strategy

We used two different classifiers, a C4.5 Decision Tree (DT) and a Multilayer Perceptron (MLP) Neural Network (NN), applying stratified 10-fold cross-validation. Details on the classifiers are depicted in section 3.3

First, the Accent vs No Accent classification problem (the most classical one in the literature) was approached. The goal is to contrast our systems with the state of the art. Next the more complex multiclass accent type classification problem was approached.

Once shown the trouble of the multiclass problem (high error rates in accent recognition) we focused on the data analysis, previous to continue with the classification problem. A contrast in pair of accent types was performed by applying the classifier to the easier task of binary classifications for every pair of accents. The goal is to identify similar classes as a source of confusion in the multiclass problem. Multidimensional scaling [11] is used to display these inter-class potential similarities.

3.2. Data preprocessing

Some classifiers can not handle qualitative features as the POS ones. We transformed them into quantitative characteristics by using two approaches: binary masks (one bit per POS type); and codification of the 33 values using 6 bits.

Due to the different range of the features, we applied different normalization techniques: the Z-Norm, Min-Max, divide by maximum and euclidean norm 1.

The approaches proposed for dealing with the imbalanced data can be divided into internal and external ones, i.e., at algorithmic and data level, respectively [12]. In the first, new algorithms or modifications of existing ones are proposed. In the second, the data sets are re-sampled **over-sampling** the minority class or **under-sampling** the majority class. Both options can be accomplished randomly or directed. We are interested in general solutions, so only external solutions have been applied, more specifically, re-sampling method based on minority class example repetition has been performed.

3.3. The classifiers

The Weka machine learning toolkit [13] was used to build C4.5 decision trees (J48 in Weka). Different values for the confidence threshold for pruning have been tested, although the best results are obtained with the default value (0.25). The minimum number of instances per leaf is also set to the default value (2). This classifier was trained with un-normalized data and qualitative POS feature.

A Multilayer Perceptron (MLP) is trained per each classification problem, using the Error Backpropagation learning algorithm. Non-linear sigmoid units are used in the hidden and output layers because they showed better performance than $tanh$ ones in our experiments. Several network configurations were tested to define the final MLP configuration: i) single hidden layer, ii) training epochs equal to 100, iii) although Gori [14] has demonstrated that only using more hidden units than inputs the separation surfaces between classes in the pattern space can be closed, the results showed that using more than 16 hidden units is not worth it, iv) as many units as classes are used in the output layer, one per each class to classify.

To train the MLP unsaturated desired outputs [15] were tested. The chosen ones, however, were 1.0 for the output corresponding to the training vector class and 0.0 for the rest, since a better performance was achieved.

Although the assumptions to approximate the MLP output to a posteriori probability are not fulfilled [15], given a test vector x_i, each output of the MLP, trained to distinguish between n classes C_j, can be seen as the estimation of the membership degree, $\Gamma(C_j/x_i)$, of vector x_i to class C_j. Then, the input vector is assigned, in accordance with this probabilistic output interpretation, as follow: $x_i \in C_j$ with $j = \arg\max_j \Gamma(C_j/x_i)$. If all the outputs have the same value, that is very rare, the input is assigned to the most probable class, i.e., the largest.

The codification alternative showed better performance to transform the POS feature (besides, the input vector is smaller). Z-Norm was the chosen to normalize the feature ranges, since it showed the best performance.

4. Results

When the classifiers are applied to the *Accent* vs *No Accent* binary decision, results are close to the expected according to the state of the art: we achieved 84.7% with NN and 82.7% with DT. [6] summaries the state of the art up to date reporting results from 75.0% to 87.7%. These results have been obtained using the Boston Radio Corpus using all the input features mentioned in the last paragraph of section 2.

When we change to cross-lingual scenarios, first limitation is that not all the prosodic features are immediate to be used. For example POS Tags can be highly dependent on the language. We decided in this preliminary approach to make use of the F0

```
f0_minavg_diff <= 17.9909
|   f0_range <= 8.9: none (4224.0/373.0)
|   f0_range > 8.9
|   |   e_maxavg_diff <= 792.058: none (6008.0/1616.0)
|   |   e_maxavg_diff > 792.058
|   |   |   e_maxavg_diff <= 1054.15: none (2186.0/924.0)
|   |   |   e_maxavg_diff > 1054.15: accent (1916.0/874.0)
f0_minavg_diff > 17.9909
|   f0_avgutt_diff <= -28.3332
|   |   f0_minavg_diff <= 52.6954: none (1297.0/389.0)
|   |   f0_minavg_diff > 52.6954
|   |   |   e_maxavg_diff <= 965.426
|   |   |   |   f0_avgutt_diff <= -35.3209: none (166.0/64.0)
|   |   |   |   f0_avgutt_diff > -35.3209
|   |   |   |   |   e_range <= 1123.68: none (25.0/9.0)
|   |   |   |   |   e_range > 1123.68: accent (85.0/23.0)
|   |   |   e_maxavg_diff > 965.426: accent (272.0/52.0)
|   f0_avgutt_diff > -28.3332: accent (11588.0/2388.0)
```

Figure 1: Decision tree C4.5. Simplified version with pruning confidence of 0.001 (default is 0.25)

Classified as –>	Accent	No Accent
Accent	1698	634
No Accent	616	1698

Table A

Classified as –>	Accent	No Accent
Accent	1663	669
No Accent	564	1750

Table B

Table 3: Classification results using the C4.5 decision tree. Table A uses input features of frequency and energy. Table B only uses F0 features.

and energy features.

Second difference with respect to the mono-lingual scenario is the need to normalize the input features. Figure 1 shows the simplified decision tree resulting when the algorithm is feed with not normalized data. Despite the features are relative only the F0 features would support a comparison between corpora and speakers. Energy features are still highly dependent, not only on the speaker, but also on recording conditions. A change on the energy scale between corpora would affect dramatically results. The data in Boston Radio Corpus and in ESMA-UPC have been normalized separately using the same z-norm.

Table 3 shows the confusion matrix using normalized fundamental frequency features and energy when the decision tree is trained with the Boston Radio Corpus and tested with the ESMA-UPC corpus. Despite the most relevant set of features seems to be the one relating to F0, energy seems to have a rol to take into account.

Table 4 compares recognition rates in pairs of corpus.

Modelling-Testing corpora	Accent	No Accent
Boston-Boston	73.4%	76.7%
ESMA-ESMA	80.6%	72.1%
Boston-ESMA	75.6%	71.6%
ESMA-Boston	75.1%	73.4%

Table 4: MLP Neural Network classification rates results. Input features correspond to F0 and energy features.

When the pair modelling-testing correspond to samples of the same corpus, results are better. Nevertheless, the figures corresponding to cross-lingual experiments are not far away from the mono-lingual ones and are encouraging to make use of other features relating to duration or to grammar to increase the identification results.

No Accent classification seems to be more difficult in the ESMA corpus than in the Boston one (72.1% and 71.6% versus 76.7% and 73.4%). This is because the ESMA corpus is divided into stress groups not in words. The first word (in any) of the stress group is un-stressed by definition, so that there is an important number of words that would belong to the No Accent category that would be identified easily by the classifier.

5. Discussion and future work

In [8] we show the importance of the proper selection of the input feature in order to improve result by entering a more expressive parameterization technique based on the use of Bzier functions [16]. When we perform multiclass classification with the data of the Boston Radio Corpus results dramatically decrease (see table 6). Nevertheless, the table 5 shows the classification rates for every pair of classes. The use of the Bézier coefficients outperforms the results in both classifiers. Although in the *Accent* vs *No Accent* the improvement is very low, in the multiclass and in the pairwise classification problem the use of Bézier coefficients permits to improve results. For example !H* increases its rates from 18.7 to 29.3 in multiclass classification, and it also increases its performance with respect to all the other classes in the pairwise classification problem. In this paper we made use of a set of features getting inspiration from other state of the art studies, but a deeper analysis and the test of alternative features seems to be a need.

The significant differences between the two type of classifiers opens the door to an alternative research such as the use of expert fusion. By combining results of different classifiers or by specialising experts in different type of accents we expect to improve the performance.

In this work we focus on the binary accent vs. no accent problem. The questions arising at this moment is weather is possible or not to extend this approach for the recognition of different ToBI accent. The immediate answer is no, and reason is that ToBI sequences are highly language dependent. Furthermore, ToBI identification results in the corpus are still very poor. Finally, the high level of inconsistency in labellers tagging for the case of Sp-ToBI[17] does not encourage to explore this possibility.

6. Conclusions

We have presented a cross-lingual experience on ToBI accent identification. The two corpora used have been presented and the experimental strategy has been described. Results indicate that this is a promising alternative to analyze in deeper detail in future work.

Related work [8] has shown us the difficulties of doing multiclass ToBI accent classification, but we have also learn the future work to be done to cope with data sparseness: using more expressive prosodic features and using more powerful learning tools and strategies.

7. References

[1] J. Terken, "Fundamental frequency and perceived prominence," *Journal of Acoustics of America*, vol. 89, no. 4, pp. 1768–1776,

(a) MLP Neural Networks

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		60,7	59,8	77,8	66,7	85,8	98,6	86,8
L+H*	59,0		71,0	77,7	64,6	83,4	96,7	86,6
!H*	65,4	68,4		71,7	59,5	77,8	96,5	85,5
H+!H*	61,4	73,4	60,2		67,5	66,1	92,4	69,5
L+!H*	51,9	53,0	53,3	78,6		80,3	90,3	71,9
L*	73,3	79,0	66,7	64,6	78,5		91,5	68,3
L*+H	4,0	6,0	12,0	18,0	16,0	32,0		4,0
none	81,1	87,9	82,7	81,9	89,9	85,2	99,4	

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		67,1	64,8	84,2	72,1	93,4	99,0	85,0
L+H*	60,8		72,8	85,5	65,8	92,8	97,8	87,3
!H*	65,5	74,2		77,9	65,9	86,7	97,8	84,0
H+!H*	62,2	72,4	61,4		74,9	78,8	96,8	63,4
L+!H*	48,3	52,0	56,4	78,0		88,8	90,8	73,3
L*	71,9	78,3	73,7	68,1	89,8		92,9	63,7
L*+H	8,0	12,0	20,0	56,0	38,0	36,0		26,0
none	85,6	90,2	84,2	85,7	94,0	90,4	99,6	

(b) C4.5 Decision Trees

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		71,3	68,3	91,8	89,9	93,6	98,3	80,6
L+H*	41,5		67,4	83,7	73,5	88,5	97,2	69,6
!H*	50,8	62,6		79,2	66,6	82,6	97,1	59,8
H+!H*	29,7	54,6	43,2		71,5	60,6	90,3	22,7
L+!H*	14,7	33,7	42,3	74,1		77,3	91,5	47,8
L*	33,8	65,8	49,9	66,2	77,0		91,7	21,7
L*+H	6,8	11,4	9,1	25,0	22,7	38,6		9,1
none	82,2	92,2	90,5	95,8	96,2	96,4	98,7	

	H*	L+H*	!H*	H+!H*	L+!H*	L*	L*+H	none
H*		75,6	75,3	92,8	90,4	95,3	98,2	78,1
L+H*	36,5		71,2	86,4	77,1	93,0	97,0	70,0
!H*	43,6	68,6		81,3	77,1	87,8	97,2	57,6
H+!H*	33,1	63,5	47,3		74,4	71,7	93,5	25,3
L+!H*	16,8	32,6	38,9	74,1		86,8	93,3	48,7
L*	59,2	81,2	70,4	70,8	85,9		91,1	29,2
L*+H	2,3	13,6	15,9	29,5	34,1	43,2		15,9
none	84,2	93,4	91,7	95,6	97,0	96,6	99,0	

Without Bézier parameters   With Bézier parameters

Table 5: Accuracy (in %) of the pairwise classifiers using neural networks (a) and decision trees (b). In both cases, individual class success rate is shown. Tables on the left show results without Bézier coefficients and the ones on the right with Bézier coefficients. Position i, j of the table represents the success rate of the class i in the classifier $i\ vs.\ j$.

	C4.5 DT		**MLP NN**	
Acc Type	**NBez**	**Bez**	**NBez**	**Bez**
H*	44.4	45.5	21.5	22.1
L+H*	22.7	25.6	35.4	41.0
!H*	18.1	21.9	18.7	29.3
H+!H*	9.4	12.5	32.7	42.7
L+!H*	6.6	7.1	28.8	31.1
L*	11.4	17.6	43.5	59.6
L*+H	0.0	2.3	0.0	2.0
none	75.3	75.5	68.3	68.2
Acc-NoAcc	82.6	82.7	83.0	84.7

Table 6: Accuracy (in %) of the Decision Trees (column *C4.5 DT*) and Neural Networks (column *MLP NN*) in the multiclass accent type and accent vs. no accent (last row) recognition tasks, when the Béziers coefficients are used (column *Bez*) and not used (column *NBez*).

1991.

[2] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling English prosody," in *Proceedings of ICSLP-1992*, 1992, pp. 867–870.

[3] M. Ostendorf, P. Price, and S. Shattuck, "The boston university radio news corpus," Boston University, Tech. Rep., 1995.

[4] A. K. Syrdal, J. Hirshberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speech manual labeling of prosody," *Speech Communication*, no. 33, pp. 135–151, 2001.

[5] S. Ananthakrishnan and S. Narayanan, "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 216–228, January 2008.

[6] V. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 797–811, May 2008.

[7] L. Aguilar, A. Bonafonte, F. Campillo, and D. Escudero, "Determining Intonational Boundaries from the Acoustic Signal," in *Proceedings of Interspeech 2009*, 2009, pp. 2447–2450.

[8] C. Gonzalez, C. Vivaracho, D. Escudero, and V. Cardenoso, "On the automatic ToBI accent type identification from data," in *Proceedings of Interspeech 2010*, 2010.

[9] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, October 1994.

[10] A. Ferrer, "Sintesi de la parla per concatenaci basada en la selecci," Ph.D. dissertation, Dpto. de Teora del Senyal i Comunicacions, Universidad Politcnica de Catalua, Espaa, 2001.

[11] R. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley and Sons, 2001.

[12] A. Vivaracho-Pascual, Simon-Hurtado, "Improving ann performance for imbalanced data sets by means of the ntil technique," in *Accepted to the IEEE International Joint Conference on Neural Networks*, 18-23 July 2010.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[14] M. Gori, "Are multilayer perceptrons adequate for pattern recognition and verification?" *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1121–1132, November 1998.

[15] S. Lawrence, I. Burns, A. Back, A. Chung Tsoi, and C. L. Giles, "Neural networks classification and prior class probabilities," *Lecture Notes in Computer Science State-of-the-Art Surveys*, pp. 299–314, 1998.

[16] D. Escudero and V. C. A. Bonafonte, "Corpus based extraction of quantitative prosodic parameters of stress groups in spanish," in *Proceedings of ICASSP 2002*, vol. 1, 2002, pp. 481–484.

[17] D. Escudero and L. Aguilar, "Procedure for assessing the reliability of prosodic judgements using Sp-TOBI labeling system," in *Proceedings of Prosody 2010*, 2010.

Text classification of domain-styled text and sentiment-styled text for expressive speech synthesis

Alexandre Trilla, Francesc Alías, Isaac Lozano

GTM – Grup de Recerca en Tecnologies Mèdia
LA SALLE – UNIVERSITAT RAMON LLULL
Quatre Camins 2, 08022 Barcelona (Spain)

atrilla@salle.url.edu, falias@salle.url.edu, st18187@salle.url.edu

Abstract

In the context of text processing for Text-to-Speech (TTS) synthesis, this work aims to automatically direct the expressiveness in speech through tagging the input text appropriately. Since the nature of text presents different characteristics according to whether it is domain-dependent (related to its topics) or sentiment-dependent, it is studied how these traits influence the identification of expressiveness in text.

To this end, two principal Text Classification (TC) methods are considered: a graph-based approach named the Reduced Associative Relational Network and the Maximum Entropy classifier. Their effectiveness in domain/sentiment dependent environments is evaluated. The results indicate that moving from a domain-dependent environment to a more general sentiment-dependent environment strictly results in poorer effectiveness rates, despite the sensible direct association that sentiment provides for dealing with expressiveness. Additionally, it is also evaluated how sensitive the classifiers are to a small increase of training data, yielding a slight positive influence.

Index Terms: domain classification, sentiment classification, expressive Text-to-Speech synthesis

1. Introduction

Expression is suggested to be a manner of speaking, a way of externalising feelings, attitudes and moods – conveying information about an affective state [1]. Traditionally, the Text-to-Speech (TTS) synthesis community relates expressiveness with *emotion* [2], while the Text Analysis community focuses on *sentiment* [3]. Despite many existing TTS-related publications study these problems, as far as we know a specific study on the implications concerning the expressive information that the text includes, as well as the implications of the acoustic features that convey expressiveness in speech, is still lacking. In general, the focus is on one particular aspect (either speech or text) relying on some other existing system for the missing complementary information, regardless that this already existing system may be conceived for a purpose different from the complete system. For example, [4, 5, 6] focus on the production process of expressive synthetic speech while [7, 8, 9] focus on the extraction of relevant information from text in order to direct expressiveness in speech.

In detail, [4] detects affect in text through the identification of situations that evoke common emotional responses (based on a former psychological study), and then concentrates on synthesis. In [5] and [6] the authors employ different dictionaries of affect (based on different models of affect) to extract emotional information from the lexicon, and again are focused on synthesis. On the contrary, [7] focuses on extracting linguistic features from text in order to predict its sentiment, and then synthesises speech. In [8] the authors produce a content-dependent list of emotional words to match the words in a given text, to then classify its emotion for further speech synthesis. And [9] concentrates on classifying texts pertaining to different domains (topics) and then assigns an expressive style to each domain based on a predefined expert criterion (relating each domain with a pre-defined speaking style).

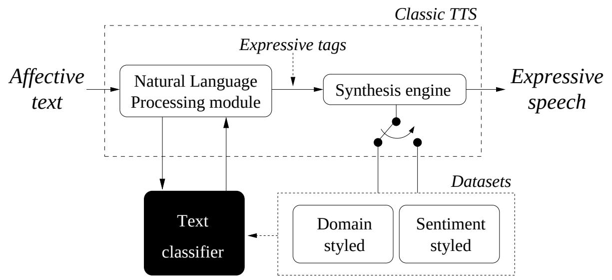


Figure 1: *Framework of a Text-to-Speech (TTS) synthesis system including a text classifier to automatically direct the expressive style in speech. Dealing with a domain-styled dataset or a sentiment-styled dataset determines the expressiveness of the TTS system, for one given language.*

This paper is focused on the automatic extraction and tagging of expressive information from text at sentence level, moving from domain to sentiment dependent characteristics, and considering its implications regarding the expressiveness of the TTS synthesis system, see Figure 1. These tags should then direct the expressive style in speech. In this study two expressiveness-enabled environments are presented: (1) a domain-styled dataset, i.e. a multidomain dataset where the domains of the texts are heuristically assigned to – and recorded with – particular expressive styles [9], and (2) a sentiment-styled dataset, i.e. a dataset where the sentiments evoked by the texts directly determine the expressive styles [10]. In domain-styled data, the lexicon pertains to a given topic, whereas in sentiment-styled data, this association may not be direct or even it may not exist at all. For example, while it is plausibly clear that words like "teacher" and "school" pertain to the same domain, namely "education", it is unclear that these same words by themselves alone may be related with a particular sentiment. The relevance of this domain dependency to attain a good effectiveness rate in text classification is studied in environment 1. Nevertheless, dealing only with such domain-styled data sensibly limits the generalisation of expressive TTS synthesis system as it only succeeds in identifying texts pertaining to the training domains, and thus delivering speech in only the predefined domain-related expressive speaking styles. Therefore, environment 2 investigates how the text classification method performs in a more general environment, that is the sentiment-styled text, where the sentiment labels directly determine the expressive styles. Also in this work it is studied if training the classifiers with an additional small amount of text may improve the text classification effectiveness in both expressive environments. In expressive Unit Selection TTS (US-TTS), most (if not all) of the texts need to be recorded with expression [9], and the creation and labelling of a speech database is a cost to minimise. Thus, there is an interest in maintaining a restriction on the small size of the corpora because the final goal is the production of synthetic speech. In addition, real-time speech synthesis is pursued, hence, the Text Classification (TC) computer cost should also be minimised.

Section 2 describes the TC methods and corpora considered to conduct the experiments, which are described in Section 3. The obtained results are discussed in Section 4 and the paper is concluded in Section 5.

2. Text classification method

This section explores how TC performs in domain/sentiment dependent environments. To this end, two domain and sentiment corpora labelled in expression are used to to train three different principles of classification given the task at hand.

2.1. Expressiveness-styled corpora

On the one hand, the Advertising Database compiles 1350 advertisements pertaining to three different domains (topics), namely education, technology and cosmetics. Each of these domains was assigned to an expressive speaking style, happy, neutral and sensual respectively, based on an expert criterion [9].

On the other hand, the Semeval 2007 training dataset [10] compiles 1000 headlines, labelled in emotion after conducting a subjective survey, and adapted to the sentiment label set (positive/negative/neutral) through a mapping on Russell's model of affect, see [11] for further details.

Table 1: *Properties of the corpora. 5-lexicon represents the size of the lexicon of words appearing at least 5 times.*

Property	Advert. Database	Semeval 2007
Instances	1350	1000
Vocabulary	2643	3145
5-lexicon	368	212
Class-balance	0.39 (HAP) 0.38 (SEN) 0.23 (NEU)	0.55 (NEU) 0.33 (NEG) 0.12 (POS)

Note that the datasets are of comparable size and are also labelled with the same amount of categories, see Table 1. Also, for both datasets the instances correspond to sentences, i.e. one sentence per document, that is the hardest context to attain a good classification effectiveness [9]. Besides, observe the different nature of the data: while the Semeval 2007 dataset is slightly smaller, its vocabulary size is slightly greater, denoting a richer lexicon, less bound to a domain (topic) in particular. Nevertheless, the TC performance considering the nature of the

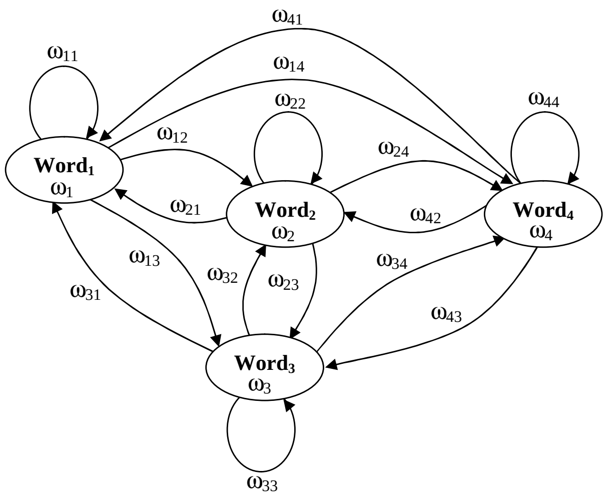


Figure 2: *Graphical representation of the ARN [9].*

data might still be influenced by the balance of the classes, but the overall results should be faithfully comparable.

2.2. Classification methods description

This section first describes the TC approaches for domain-styled text. To this purpose, the Reduced Associative Relational Network (ARN-R) is used given its conception for multidomain US-TTS synthesis and its promising effectiveness in domain-dependent TC at sentence level [9]. This method showed to perform better than Nearest Neighbour (example-based classifier), Independent Component Analysis (unsupervised text classifier) and n-grams at character-level (inductive generative classifier) in [9]. Support Vector Machines (SVM), which are some of the most successful techniques for conducting TC, were determined to fail at classifying the Advertising Database due to its small size [9]. Thus, SVM are excluded from the experiments given the size of the datasets.

However, in order to extend the performance comparison of the ARN-R, it is compared against a Maximum Entropy (MaxEnt) classifier. On the one hand, due to its functional form, that is different from the rest of the already compared classifiers (inductive discriminative classifier). And on the other hand, due to its use with success in many sentiment analysis tasks [12], with a performance comparable to the SVM. Hence, its consideration is found necessary to obtain a good baseline of the current state of the art.

Then the work continues on describing the TC approaches for sentiment-styled text. The ARN-R and MaxEnt classifiers are compared, now on the sentiment environment, and additionally considering a Nearest Emotional Centroid (NEC) classifier (centroid-based classifier with expert knowledge) [11]. The NEC is regarded to be a good baseline because it intuitively captures the expert knowledge related to the field of emotion.

2.2.1. *Reduced Associative Relational Network*

The ARN-R builds a graph of words for each training category by associating each word to a node and each collocation (ordered co-occurrence of two adjacent words) to a directed link, like in Figure 2. The weights associated to each term (node or link) are computed from training corpora and weighted according to a term weighting method.

On testing, the ARN-R first builds a similar graph with the text to test. Afterwards, it vectorises its weighted terms defining the dimensions of a Vector Space Model (VSM). In this space, a vector for each category is projected, maintaining the weights of each term coinciding with the dimensions of the VSM. In the end, the ARN-R makes use of a similarity measure (e.g., the cosine distance) and assigns the label corresponding to the most similar categorical representation (hard classification).

Regarding the term weighting methods, the ARN-R considers an unsupervised method named the Inverse Term Frequency (ITF), which yields better results than simple term frequencies, see [9]. Moreover, a supervised weighting method, named the Relevance Factor (RF), is also considered. The RF yields better results than traditional unsupervised term weighting methods in TC, see [13]. Equations (1) and (2) show the ITF and the RF respectively.

$$ITF_t = \log\left(\frac{\sum_{t' \in T} tf_{t'}}{tf_t}\right) \tag{1}$$

$$RF_{t,C} = \log\left(1 + tf_t\right)\ \log_2\left(2 + \frac{tf_{t,C}}{\max(1, \sum tf_{t,\bar{C}})}\right) \tag{2}$$

where t represents the term, tf represents its Term Frequency, T represents the vocabulary (total number of different terms) and C represents the positive category (likewise $\bar{C}$ represents the negative category).

While the ITF intends to weight the local contribution of a given term, the RF intends to weight the contribution of a term considered to pertain to a given category regarding the rest of categories.

2.2.2. *Maximum Entropy*

Maximum entropy modelling is a framework for integrating information from many heterogeneous information sources for classification. MaxEnt models are first given a set of constraints that are justified by the available data, and then compute the model with maximum entropy of all the models that satisfy the constraints. The MaxEnt model is motivated by the desire to preserve as much uncertainty as possible, avoiding to infer anything beyond the data, see [14] for further details. The categories are modelled with the exponential form shown in Equation (3).

$$P(C|t) = \frac{1}{Z}\ \exp\left(\sum_i \lambda_{i,C}\ F_{i,C}(t, C)\right) \tag{3}$$

where the Z above is a normalisation factor in order to define a probability distribution. The feature/category functions $F_{i,C}$ deal directly with the term binary features of presence or absence without assuming any relationship among them. The parameters $\lambda_{i,C}$ are set to maximise the entropy of the induced distribution. The expected values of the feature/class functions have to be equal to the evidence shown in the training data [12].

2.2.3. *Nearest Emotional Centroid*

The NEC first represents text in a space defined by emotional dimensions (basic properties of affective states according to expert knowledge: valence, activation and control), named the circumplex. Then, it computes the sentiment centroids on training data (i.e. their vector sum). Finally, it performs comparisons with a minimum-distance (to the centroids) criterion, see [11] for further information. The NEC approach is considered to be a suitable baseline to deal with such texts of subjective nature (the sentiment of affective states) for incorporating offline expert knowledge.

3. Experiments

Firstly, the classifiers are submitted to experimentation on both datasets, the Advertising Database and the Semeval 2007 dataset, in order to study how the considered classification methods perform on domain versus sentiment-styled texts respectively. To that effect, several configurations are considered: in the term weighting method (ITF or RF) and in the consideration of collocations.

Moreover, it is studied if slightly augmenting the size of the training data could be of help to improve the effectiveness rates. According to the premise of small-sized datasets, a 10% data increase is appended to the Advertising Database (the largest dataset), and a 25% data increase is appended to the Semeval 2007 dataset (the smallest dataset). The different size in these extensions intends to make the corpora more similar with regard to their size.

The experiments are compared using the macroaveraged F_1^M effectiveness measure [15], estimated with a 10-fold Cross Validation method. ANOVA tests at the 0.05 confidence level are used for evaluating statistical significance.

4. Results and discussion

The results are shown in Table 2. A preliminary experiment with the ARN-R with plain term frequencies (no term weighting method applied) yielded an average effectiveness rate of 58.69% for the Advertising Database, and 46.73% for the Semeval 2007 dataset. This experiment intended to highlight the importance of the term weighting method in the TC task, given that any of the presented methods with weighted terms has yielded better results than plain term frequencies (8.54% better on average).

Regarding the relevance of domain dependence to attain a good effectiveness rate in TC, as suggested in Section 1, the obtained results indicate that domain dependence contributes positively toward a good effectiveness rate. It can be observed that the ARN-R and MaxEnt methods perform equivalently (with no significant difference) in the two environments despite their different learning paradigm. They yield almost a 21% of average superior performance for the domain-styled environment (the Advertising Database) regarding the non domain dependent, but generalisable, sentiment-styled environment (the Semeval 2007 dataset). These high effectiveness rates are attributed to the relation between the lexicon and the domain, i.e. the typical case in TC [15]. When the TC method intends to match an unknown text to any of the learnt domain-dependent words, it performs better when there are many words alike (note that the Advertising Database contains 7.18% more frequent words than the Semeval 2007 dataset, see "5-lexicon" field in Table 1).

Nevertheless, a significant difference is observed among the features used for classification. In the domain dependent environment, the ITF has performed significantly better than RF, about 20% better without collocations, and about 11% better with collocations. The fact that the RF performs better with collocations responds to the theory that the RF raises the presence of singularities in the data [9], and such singularities are more easily found among collocations than among words alone. In the sentiment dependent environment, though, all classification methods have performed similarly regardless of their principle

Table 2: *Results evaluating the F_1^M effectiveness rate of several text classification methods on different environments (mean ± std).*

Classification method	**Advertising Database**		**Semeval 2007 corpus**	
	Training	Extension	Training	Extension
ARN-R ITF	79.73% ± 3.69	79.81% ± 3.13	51.45% ± 4.71	51.81% ± 4.94
ARN-R ITF + Col.	79.15% ± 4.36	79.40% ± 3.08	52.27% ± 6.02	54.09% ± 5.45
ARN-R RF	59.36% ± 4.30	60.06% ± 4.28	53.96% ± 3.71	55.93% ± 4.09
ARN-R RF + Col.	68.56% ± 3.75	70.45% ± 3.73	55.78% ± 2.86	56.18% ± 3.35
MaxEnt	79.95% ± 3.47	80.36% ± 2.59	51.74% ± 7.95	53.04% ± 7.99
MaxEnt + Col.	76.30% ± 3.19	77.56% ± 2.33	52.18% ± 6.94	53.89% ± 6.33
NEC	–	–	49.04% ± 5.33	48.96% ± 5.24

of classification or term weighting strategy. All methods have yielded better results than the baseline NEC, although no statistically significant difference is observed among them. These results show the difficulty of classifying text pertaining to a more general environment, where the lexicon may not be related with a particular domain.

Regarding the training of the classifiers with an additional small amount of text, the results show a positive tendency in all cases (no particular behaviour depending on the features nor the classification principle), but the improvements are not statistically significant. The maximum increment observed is close to 2%. This increment is expected to grow as the size of the extension is enlarged, but then the cost of recording a large dataset should be faced (the trade-off needs to be studied in detail). In summary, the overall results indicate that while it is positive to slightly extend the training dataset, the final effectiveness rates are insignificantly increased.

5. Conclusions

This work intends to contribute to the generalisation of sentence-level TC for expressive TTS synthesis. In this sense, the impact of moving from domain-dependent to sentiment-dependent expressiveness in text is analysed, because the latter has a more sensible direct association with expressiveness. To that end, the ARN-R and MaxEnt TC systems are evaluated. They yield an equivalent effectiveness performance in each of the studied domain-dependent and sentiment-dependent environments. However, for domain-dependent data results are significantly better. This may be due to the strong relation between the data domain and its lexicon. The results delivered by both systems for sentiment data outperform the baseline NEC system, and scoring better in this environment already represents an improvement with respect to the expressiveness generalisation purpose.

The ARN-R presents more flexibility than MaxEnt to use strategies related to the text processing field, such as the term weighting schemes. In this sense, this work has evaluated the impact of unsupervised (ITF) and supervised (RF) term weighting functions, obtaining slightly better results (non-significant) with the supervised method. In the future work, in addition to considering more term weighting aspects, it is expected to extend the exploitation of the similarity of texts in the ARN-R with graph-based measures, like graph distances such as the Pattern Length [9]. These measures are expected to strengthen the need of a graph-based structure to better model the behaviour of sentiment-styled text.

6. References

[1] M. Tatham and K. Morton, *Expression in Speech: Analysis and Synthesis*. New York, NY, USA: Oxford University Press, Inc., 2004.

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, 2009.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[4] A. R. F. Rebordao, M. A. M. Shaikh, K. Hirose, and N. Minematsu, "How to Improve TTS Systems for Emotional Expressivity," in *Proc. of Interspeech'09*, Brighton, UK, Sep. 2009, pp. 524–527.

[5] Z. Wu, H. M. Meng, H. Yang, and L. Cai, "Modeling the Expressivity of Input Text Semantics for Chinese Text-to-Speech Synthesis in a Spoken Dialog System," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1567–1576, Nov. 2009.

[6] G. O. Hofer, K. Richmond, and R. A. J. Clark, "Informed Blending of Databases for Emotional Speech Synthesis," in *Proc. Interspeech*, Sep. 2005.

[7] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proc. of HLT'05*. Morristown, NJ, USA: ACL, 2005, pp. 579–586.

[8] V. Francisco and R. Hervás, "EmoTag: Automated Mark Up of Affective Information in Texts," in *Proc. of the Doctoral Consortium in EUROLAN 2007 Summer School*, C. Forascu, O. Postolache, G. Puscasu, and C. Vertan, Eds., Iasi, Romania, Jul.–Aug. 2007, pp. 5–12.

[9] F. Alías, X. Sevillano, J. C. Socoró, and X. Gonzalvo, "Towards High-Quality Next-Generation Text-to-Speech Synthesis: A Multidomain Approach by Automatic Domain Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1340–1354, Sep. 2008.

[10] C. Strapparava and R. Mihalcea, "SemEval-2007 Task 14: Affective Text," in *Proc. of SemEval'07*, Prague, Czech Republic, Jun. 2007.

[11] A. Trilla and F. Alías, "Sentiment classification in English from sentence-level annotations of emotions regarding models of affect," in *Proc. of Interspeech'09*, Brighton, UK, Sep. 2009, pp. 516–519.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proc. of EMNLP'02*, Philadelphia, PA, USA, Jul. 2002, pp. 79–86.

[13] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," *IEEE T. Pattern. Anal.*, vol. 31, no. 4, pp. 721–735, Apr. 2009.

[14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: The MIT Press, 1999.

[15] F. Sebastiani and C. N. D. Ricerche, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, pp. 1–47, 2002.

Synthesis using Speaker Adaptation from Speech Recognition DB

Sergio Oller, Asunción Moreno, Antonio Bonafonte

TALP Research Center, Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
soller@gps.tsc.upc.edu, asuncion.moreno@upc.edu, antonio.bonafonte@upc.edu

Abstract

This paper deals with the creation of multiple voices from a Hidden Markov Model based speech synthesis system (HTS). More than 150 Catalan synthetic voices were built using Hidden Markov Models (HMM) and speaker adaptation techniques. Training data for building a Speaker-Independent (SI) model were selected from both a general purpose speech synthesis database (FestCat;) and a database designed for training Automatic Speech Recognition (ASR) systems (Catalan SpeeCon database). The SpeeCon database was also used to adapt the SI model to different speakers.

Using an ASR designed database for TTS purposes provided many different amateur voices, with few minutes of recordings not performed in studio conditions. This paper shows how speaker adaptation techniques provide the right tools to generate multiple voices with very few adaptation data. A subjective evaluation was carried out to assess the intelligibility and naturalness of the generated voices as well as the similarity of the adapted voices to both the original speaker and the average voice from the SI model.

Index Terms: speech synthesis, HMM, Adaptation

1. Introduction

Concatenative-based speech synthesis systems have proven to achieve very high quality synthetic voices [1]. These systems need huge and expensive databases preferably recorded from professional speakers, using a phonetically balanced corpus, in very controlled environments and carefully segmented and labelled. Generation of multiple voices implies either to record several voices from professional speakers or to use techniques of speech transformation or speech conversion from clean recordings, usually from a given text.

HTS based systems are versatile. Phonetic units are modelled by a set of Hidden Markov Models (HMM) trained from data from one or more speakers. Speech is synthesised from the parameters (i.e. F0 and cepstral parameters) generated by the HMM in synthesis mode [2]. The quality in terms of intelligibility and naturalness is good and it is known that it is a competitive technology compared with the well established concatenative systems. Multiple voices can be generated by using speaker adaptation techniques to the HMM [3].

In this paper we apply the ideas of [3] to perform a multiple-voice speech synthesis system. We want to test the possibility of adaptation of an average voice to non-professional speakers, with a broad dialectal variety, recorded in non-controlled environment, using both read and spontaneous speech, and few minutes of adaptation data.

This kind of data is typically found for training Automatic Speech Recognition (ASR) systems. ASR databases usually consist of hundreds of speakers with few recordings in noisy environments. In ASR databases, each speaker does not need to utter a phonetically balanced corpus (balance is usually considered among many different speakers) and sentences may be read or spontaneously uttered. Being able to use ASR databases to TTS purposes would provide many more voices at a little extra cost. In order to use an ASR database in TTS, we must deal with the lack of full diphone coverage per each speaker and the noisy and not controlled recording environments. Speaker adaptation seemed a good tool to deal with the lack of balanced phonetic coverage, that's why we chose to use a TTS designed database to train the average voice. As the adaptation data is noisy and very weakly labelled, we combine ASR training data with the TTS designed database to generate the average voice.

The rest of the paper is structured as follows: Section 2 describes the training databases and section 3 describes the adaptation system. Sections 4, 5 and 6 present a subjective evaluation, the results obtained and their discussion. Conclusions and further work are discussed in section 7.

2. Training Databases

An HTS average model voice was built with both, data from a clean database designed for Speech Synthesis purposes, FestCat database, and a noisy database designed for training Speech Recognition systems, named SpeeCon database. For adaptation, only SpeeCon data were used. A short description of the databases follows:

2.1. Catalan FestCat database

The FestCat [4] database was designed for training concatenative speech synthesis systems. The database consists of recordings from 10 native professional Catalan speakers (5 female and 5 male). Eight speakers recorded 1 hour of speech from a phonetically balanced corpus and the other two speakers recorded 10 hours of speech from a broader scope corpus. Recordings were performed in a sound-proof room supervised by an operator. All the data was manually orthographically annotated. The orthographic transcription was phonetically transcribed into the central Catalan dialect with the FestCat transcriber [4]. The phonetic segmentation was performed using HMM-based forced alignments using our in-house automatic speech recognition tool. In order to build the average voice model, only the 1 hour voices were used. 10 hours voices were avoided because such longer corpora could unbalance the average voice model. It is important to bear in mind that as all the FestCat speakers shared the same Catalan central dialect, the speaker independent model is then dialectically biased. Better dialect coverage in the speaker independent model would increase variability in adaptation, thus allowing better adaptation to more speakers. However, that would require more recordings.

2.2. Catalan SpeeCon database

The Catalan SpeeCon database (Speech-Driven Interfaces for Consumer Devices) [5] was designed for training speech recognition systems. The database consists of recordings made by 550 adult speakers; half of them are male and half of them female. Speakers were distributed in four groups of age, four dialects and four environments: Office, Entertainment, Car, and Public hall. The corpus specification is a mixture of spontaneous and read speech, but also continuous utterances and isolated words. Spontaneous sentences were obtained asking the speaker to talk about a selected topic. All the recordings in the SpeeCon database are ortographically transcribed. In addition, the transcription includes a few details that represent clearly distinguishable audible acoustic events (speech and non-speech) present in the corresponding waveform files and not inherent in the environment as such. Events were assigned to one of these four categories [5]:

- [fil]: Filled pause. Are the typical noises used to fill pauses such as: uh, um, er, ah, mm.
- [spk]: Speaker noise. Loud noises uttered by the speakers that are not part of the prompted text are marked.
- [sta]: Stationary noise. This mark is used when a loud background noise is heard in the recordings. Only non expected noises are marked.
- [int]: Intermittent noise. This mark is used to mark intermittent noises like: music, background speech, horn sounds, phone ringing, paper rustle, cross talk, door slam, or ticks by the direction indicator in a car.

Among all the possible environments available in the SpeeCon database, Office and Entertainment environments were used in this project, because the recordings in these environments were less noisy than the recordings in Car or Public hall environments. Among all the utterances recorded in the SpeeCon database, only spontaneous sentences and phonetically rich sentences were used in this project. The recordings with [int] noises or stationary noises [sta] were discarded. After this pruning, a total of 157 speakers were kept. Table 1 shows the gender and accent distribution of the selected speakers. Notice that two thirds of the selected speakers from the SpeeCon database belonged to the central dialect. Non central dialect speakers were also selected to test how adaptation performed from one dialect to another. Table 2 summarizes the minutes of speech selected from each database.

Dialect	Male	Female	*Total*
Central	44	65	109
Gironí	11	13	24
Tortosí	4	8	12
Nord Occidental	6	6	12
Total	65	92	157

Table 1: *Dialect/gender distribution of the selected speakers.*

Model	FestCat speakers	SpeeCon speakers
Female	$4 \times 1h$	$92 \times 3.8 \pm 1.2min$
Male	$4 \times 1h$	$65 \times 3.7 \pm 1.2min$

Table 2: *Training data distribution used for the average model voice.*

3. System description

A complete synthesis system is composed of three parts: text analysis, the Phonetic-Acoustic modelling system, and a waveform generator system.

The text analysis uses Festival [6] with the FestCat frontend [7]. This front-end takes care of processing the text to convert it into phonetic units following the central Catalan dialect rules.

The Acoustic-Phonetic modelling system is based on the standard software HTS [8]. Four different streams are needed, one for the mel-cepstral coefficients and three for the LF0 coefficients that need to be modelled using a Multi-Space Distribution [9] to deal with voiced-unvoiced regions.

For the Acoustic modelling, 24+1 order mel-cepstral coefficients (the +1 accounts for the zeroth order) were extracted using a 25 ms Hamming window and a 5ms frameshift using SPTK [10]. Log F0 was extracted using the Snack library [11]. Dynamic parameters (delta and delta-delta) for mel-cepstral coefficients and LF0 were also computed. In order to prevent over-smoothing caused by the dynamic parameters, global variance is considered in the parameter trajectory optimization.

33 monophone context-independent phonetic units are initially trained. In order to deal with speaker noises and try to improve voice spontaneity and expressiveness, two extra units were added to that set. These units accounted for impulsive speaker sounds [spk] and filling sounds [fil]. Being able to model spontaneous speaker sounds provide a way of synthesising sentences with added noise marks, and this could improve voice expressiveness.

Further, the context-independent units are contextualised and clustered with a decision tree [12]. Given the available amount of data to train, 160k context-dependent phonetic units were trained after the last clustering operation.

Acoustic parameters and waveforms were generated and synthesised with HTS Engine, and the resulting models are ready to be used with the Festival Speech synthesis system.

3.1. Adaptation

The HMM adaptation system used is strongly based on the HTS Adapt demo provided at [8]. Two speaker independent models were built using data from both Festcat and SpeeCon databases: one for male speakers and the other for female speakers. Adaptation to the selected SpeeCon speakers was performed applying constrained maximum likelihood linear regression (CMLLR) to the mean vectors of each stream adapting simultaneously mel-cepstral coefficients and LF0 parameters [13]. A Maximum a Posteriori (MAP) reestimation of the models was also performed because it improves parameter estimation with sparse training data [14].

4. Evaluation Method

Assessment of speech synthesis is needed to determine the system's performance through newer versions and using different synthesis techniques. Due to the dialect bias present in the phonetic transcriber and the FestCat database, the authors perceived on an informal test that the results of adapting to non-central dialect speakers were not good enough. As across-dialect adaptation was not achieved, the evaluation was only performed to speakers from the Central dialect. A subjective evaluation was performed to our system and three voice aspects were asked in the test: Similarity, naturalness and intelligibility.

As the evaluation to the 157 speakers is expensive, the test was performed to a selection of them. For the similarity test, 4 female and 4 male speakers were chosen randomly among all the central-dialect speakers. In order to limit the length of the test, a subset of these speakers was used for the other exercises. Only 2 male, 2 female and the average voices were evaluated in the naturalness exercise and only 3 male, 2 female and the average voices were tested in the intelligibility exercise.

The test was presented to a total of 18 evaluators, mainly non familiar with speech processing. In order to be able to evaluate more speakers, two different question sets were asked. One question set was answered by 10 people and the other by 8 people. Both tests consisted of three tasks, each one related to one of the different aspects to be evaluated:

Similarity: Our main purpose was to build many different voices, so we focused on testing the similarity of the adapted voice, comparing it to both the average speaker-independent voice model and a recording from the same speaker selected at random. Eight sets of three utterances per set were presented to every evaluator. Each set consisted of an original utterance, an average voice utterance and an adapted utterance. The evaluators were asked to move a 5-value slider to either the original utterance or the average voice, depending on which was closer to the adapted utterance.

Naturalness: Six utterances were presented to the listeners from either original recordings, adapted voices or the average voices. The evaluators were asked to rate from 1 (poor quality) to 5 (excellent quality) each utterance. Although the word 'quality' appeared in the ratings of the task, the evaluators were asked to evaluate the naturalness of the recordings.

Intelligibility: Intelligibility was tested by asking to the listeners to transcribe six sentences. The test included sentences from the adapted voices, the average voices and original recordings from the SpeeCon database, as many of them were recorded in noisy environments. Examples of these sentences are *'He vist una placa'* (*'I have seen a plaque'*) or *'Una coŀlecció mundialment famosa de talles de fusta'* (*'A world famous collection of wood carvings'*). The evaluation was revised manually to avoid spelling and typing issues.

5. Results

Similarity: As it can be seen at Figure 1, similarity results vary widely. Most of the listeners agree on three of the adapted speakers (with codes 009, 067 and 161). On the other cases, boxes are bigger and data is more disperse.

Naturalness: Table 3 and Figure 2 show the results. The graph was plotted following the conventions from [15] where results are presented as standard boxplots, the median is represented by a solid bar across a box showing the quartiles, and the mean is represented by a $+$. Whiskers extend to 1.5 the inter-quartile range and outliers beyond this range are represented as circles. Synthetic voices scored around 3 on the 1-5 scale whereas natural voices scored almost perfectly. Results also show that there is not a big difference in naturalness between the average voice and the adapted voices.

Intelligibility: Original recordings from the SpeeCon database scored perfectly. Figure 3 shows the Word Error Rate for the evaluated synthetic voices. The average score of the adapted voices was 5%. The graph was plotted following the conventions from [15] where Word Error Rates (WER) are plotted as bar charts.

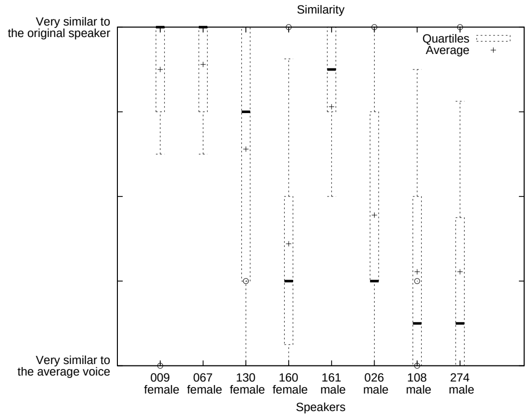


Figure 1: *Standard boxplot showing the similarity of the adapted speakers to either the original recordings or the average voice.*

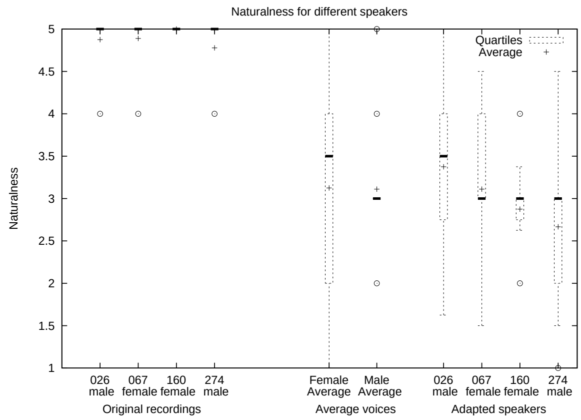


Figure 2: *Standard boxplot showing naturalness evaluated from 1 "unacceptable quality" to 5 "excellent quality".*

Voices	Mean and standard deviation
Original	4.9 ± 0.2
Adapted	3.0 ± 0.5
Average (SI)	3.1 ± 0.7

Table 3: *Global results for the naturalness test.*

6. Discussion

Adaptation results vary widely depending on the speaker. The similarity results give two groups of speakers based on the dispersion of the results. For some speakers (named 009, 067 and 161), it seems clear that the adaptation worked, as the results show that the adapted voice is very similar to the original

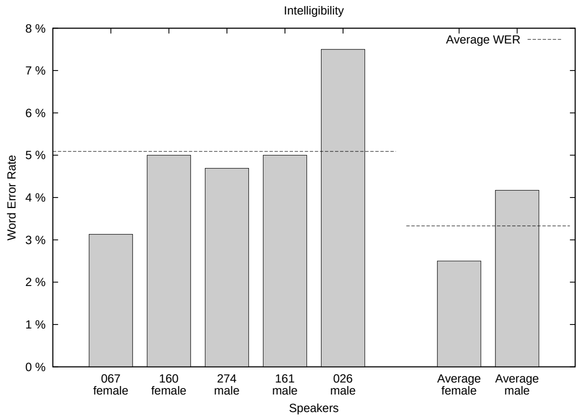


Figure 3: *Word Error Rate for the adapted speakers and the average voices. Lower is better*

speaker. For the other evaluated speakers, the results show a clearly bigger dispersion. A plausible explanation for this group of speakers is that the original speaker was close to the average voice, so the similarity among all three voices was very high and thus the distinction was not clear, making a high dispersion of the results. If the resulting boxes and whiskers had been smaller and centred between the original and the average voice, we would have assumed that the three voices (original, average and adapted) were very different and the adaptation was unsuccessful, but this has not been the case. Comparing directly the original voice and the average voice could confirm this explanation.

The adaptation process degrades slightly the intelligibility of the adapted voices relative to the average voices. Comparing the naturalness results from the adapted voices and the average voices, it can be seen that the adaptation process does not degrade significantly the naturalness of the synthetic voice. However naturalness results still show that there is room for improvement.

Results show that there is still work to do in adaptation at least in the Catalan language. We do not have a proper dialect-aware phonetic transcriber and we were not able to emulate other Catalan accents using adaptation only, mainly because some phonemes in central dialect map to two different phonemes in other dialects.

With the intention of improving the spontaneity and the expressiveness of the synthetic voice, some sentences were synthesised with the speaker sounds [spk] and [fil]. These trained units could not reproduce the typical [spk] or [fil] sounds, only unidentifiable noises were generated in their place. The most likely explanation is that each of these noise marks actually represented a wide variety of different sounds, and one model can not cope with the whole range of sounds (i.e. lip smack, cough, grunt...) However, the inclusion of spontaneous sentences in the training and adaptation corpora may give as a result more expressive and spontaneous synthetic voices even without the speaker noises. Future work is still required to test this and give confirmation.

7. Conclusions and Further work

HMM adaptation techniques can be used to generate multiple voices with reliable results. In this paper we used a combination of a database designed for TTS applications and a database designed for ASR applications to generate HMM able to be adapted to a new speaker, with as few 4 minutes of speech. Results show that intelligibility of the adapted system is acceptable, the average naturalness ranks 3.1 in a MOS scale and there is a high similarity in at least half of the voices evaluated.

Further work is addressed to improve adaptation across dialects and improve voice spontaneity and expressiveness. Different amounts of speakers with different recording contributions may be tested to improve overall quality.

8. Acknowledgements

This work was supported by Government grant TEC2009-14094-C04-01 and FEDER.

9. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 373–376.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, vol. 3. Citeseer, 2000.

[3] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda *et al.*, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech*, vol. 18, 2009, pp. 984–1004.

[4] A. Bonafonte, L. Aguilar, I. Esquerra, S. Oller, and A. Moreno, "Recent work on the FESTCAT database for speech synthesis," *I Iberian SLTech 2009*, p. 131, 2009.

[5] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon-speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, vol. 2002. Citeseer, 2002.

[6] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," 1999.

[7] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Pérez, "Corpus and voices for Catalan speech synthesis," in *Proc. of LREC Conf.*, May 2008, p. 3325–3329.

[8] HTS working group., "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/, July 2010.

[9] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Transactions on Information and Systems E series D*, vol. 85, no. 3, pp. 455–464, 2002.

[10] SPTK working group., "Speech signal processing toolkit (SPTK)," http://sp-tk.sourceforge.net/, December 2009.

[11] S. Sjölander, K.; KTH Stockholm, "The Snack sound toolkit," 2004.

[12] K. Tokuda and Z. H, "Fundamentals and recent advances in HMM-based speech synthesis," in *InterSpeech*, 2009.

[13] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *IEEE ICASSP*, vol. 2. Citeseer, 2001.

[14] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing*, vol. 2, no. 2, 1994.

[15] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Challenge Workshop*, 2007.

Oral Session 2: Speech Recognition

Speaker Tree Generation for Model Selection in Automatic Speech Recognition

David Becerril, Oscar Saz, Carlos Vaquero, Alfonso Ortega, Eduardo Lleida

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Zaragoza, Spain
{davebv,oskarsaz,cvaquero,ortega,lleida}@unizar.es

Abstract

This paper presents the procedure and results for an automated selection of the best acoustic model for an input speaker in Automatic Speech Recognition (ASR). The procedure consists in obtaining a tree which gathers a set of representative speakers of the target population; these speakers are agglomerated by means of the Bayesian Information Criterion (BIC) until all of them are merged in the top. This tree is used when a new user accesses the system by selecting the model that best fits the speech from the speaker in order to improve the performance of the ASR system without relying on speaker dependent models trained with data from the same speaker. The results will show that the BIC metric performs correctly for building the tree, and that the selected model within the tree can outperform the whole speaker independent model in an ASR task.

Index Terms: Speech recognition, adaptation, speaker clustering.

1. Introduction

The performance of Automatic Speech Recognition (ASR) systems depends heavily on how well the acoustic models within the system match the speech characteristics from the incoming speaker. In the best situation, when some data from the speaker is available, speaker adaptation techniques allow to obtain a perfect match and the best recognition rates. Otherwise, if speaker adaptation can not be performed, a general model which covers all possible types of speech considering gender (males and females), age (children, adult or elderly), dialectal type and any other possible characteristic is used.

However, it is possible to improve the performance of a speaker independent system by using a model trained only with similar speakers; i.e. a female model to recognize a female speaker or a child model to recognize children speech. The main difficulty in this approach is to estimate in advance the speech characteristics of the user to select the best suitable model. In some cases, this information might be known a priori, but in general, an automatic procedure has to be designed to achieve this best model selection in an automated way.

The approach presented here aims to organize a set of representative training speakers into a tree which can be used when a new speaker accesses the ASR system. The tree gathers the different speakers in an agglomerative process until a top model containing all the speakers is reached. When data from a new speaker is collected, the tree is evaluated to decide which is the model which best fits the speaker so that the selected model can be used for this speaker as acoustic model in an ASR system or as seed model in a speaker adaptation stage.

This work was supported by national project TIN2008-06856-C05-04 from the Spanish government and Santander Bank scholarships

The paper is organized as follows: Section 2 will explain the proposed method of tree-based agglomerative clustering of speaker models. Section 3 will describe the proposal for selecting the most suitable model within the tree for a given input speaker. Posteriorly, Section 4 will present the experimental setup for the evaluation of the proposed methods with a set of speakers. Then, the possibilities that these techniques arise for enhanced ASR and speaker adaptation will be presented in Section 5; and, finally, Section 6 will serve as discussion for this work and will present the conclusions.

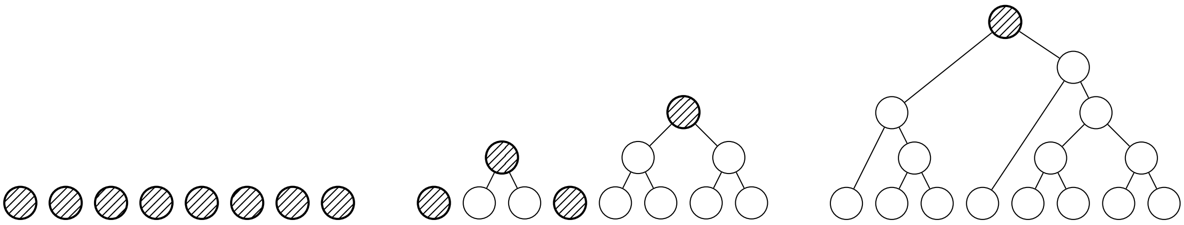
2. Automatic Speaker Tree Generation

The tree has to contain representative data from all the groups to be considered in the system. A selection of speakers has to be done in order to characterize the different groups; this may include a full set of male, female, children, and adult speakers, or a subset of them if the tree aims to model a particular type of speech.

Algorithm 1 Tree Generation

for $i = 1$ to n_{spks} **do**
  $gmm(i) =$ trainGMM()
end for
for $i = 1$ to n_{spks} **do**
  for $j = i + 1$ to n_{spks} **do**
    $distance(i, j) =$ BIC(i, j)
  end for
end for
$nodes_{rem} = n_{spks}$
repeat
  m and n so $(m, n) = \min_{i,j} distance(i, j)$
  $newnode =$ trainGMM($m \cup n$)
  $nodes_{rem} = nodes_{rem} - 1$
  for $node \in NODES \neq newnode$ **do**
    $distance(newnode, node) =$ BIC$(newnode, node)$
  end for
until $nodes_{rem} == 1$

Each speaker (i) in the whole list of speakers (n_{spks}) is characterized by a set of N_i input speech frames $(x_i(1)...x_i(N_i))$ obtained from sufficient speech data from the speaker. A previous Voice Activity Detection (VAD) stage assures that only speech frames are fed to the system, as non-speech frames (silence, noise, etc) would lead to a poorly conditioned tree. The VAD used in the proposed system is based on the Long Term Spectral Divergence (LTSD) [1]. In order to compute the LTSD, the Long Term Spectral Envelope (LTSE) and a noise estimation have to be calculated framewise. The LTSD is calculated proportionally to the ratio between the LTSE



(a) *Initial stage* (b) *Agglomerative stages* (c) *Final tree*

Figure 1: *Example of tree generation (agglomerated nodes in white, active nodes in stripped pattern).*

and the noise estimation. Then, those frames whose LTSD value is above a certain threshold are selected, otherwise they are considered as silence and therefore discarded.

The generation of the tree is performed via an agglomerative bottom-up approach following the procedure described in Algorithm 1, with a graphical example provided in Figure 1. Initially, all the leafs in the tree (bottom nodes) are assigned to a single speaker in the corpus and are marked for agglomeration, as seen in Figure 1(a). All the speech data available from each speaker is used to train a Gaussian Mixture Model (GMM) which models the speaker and the node.

In every recursion of the tree generation algorithm, a Bayesian Information Criterion (BIC) metric is used to decide which are the pair of nodes which have to be agglomerated next. At the initial stage, the metrics between all possible pairs of nodes are calculated to perform the agglomeration of the most similar nodes. When two nodes are agglomerated, the speech data from them is used to train a GMM to model the new node and the BIC metrics between this new node and the remaining nodes are calculated. In every recursion, the number of marked nodes is decreased by one and the process of agglomerating pairs of nodes according to the BIC metric is continued (Figure 1(b)) until the top node that contains all the data from all the speakers is reached (Figure 1(c)). In the end a binary tree is obtained, as nodes are agglomerated pairwise; but the tree can be unbalanced, that is, nodes that contain more than one speaker may merge with a single-speaker node at a different tree depth, as seen in Figure 1(c).

The modeling of each node as a GMM varies depending on the depth within the tree. A fixed number of Gaussian distributions are assigned to the GMM of each bottom node. Posteriorly, when a node agglomerates two lower nodes, the complexity in the data increases. In order to compensate this effect, the number of Gaussian distributions used in the new node model equals the sum of its children's Gaussian distributions. For example, bottom nodes can be modeled as a mixture of 2 Gaussian distributions. Then, when two of these nodes are agglomerated, the resulting node will have a mixture of 4 Gaussian distributions.

When a new node is created, the GMMs of the child nodes are used as seed for the new node GMM training. For the bottom nodes, when no initial data is availabe, the GMM training is initiated using k-means labels.

2.1. BIC-based Metric

BIC criteria are generally used to determine whether two sets of data ($\underline{x} = \{x_1, x_2, \ldots, x_N\}$ and $\underline{y} = \{y_1, y_2, \ldots, y_M\}$) are more likely to be modeled into one single model (alternative hypothesis) or two separate models (null hypothesis) [2].

The decision of accepting or rejecting the alternative hypothesis of merging sequences $\underline{x}$ and $\underline{y}$ in $\underline{z} = \underline{x} \cup \underline{y}$ of length $P = N + M$ with models X and Y in a single model Z is formulated in Equation 1. Alternative hypothesis is accepted when the log-probability of the merging model is greater than the sum of the log-probabilities of the two initial models; where a parameter d represents the penalty on the merging model to compensate its higher complexity.

$$\log P(\underline{z}|Z) - \frac{1}{2} d \log(P) \geq \log P(\underline{x}|X) + \log P(\underline{y}|Y) \quad (1)$$

In our case, where the number of parameters of the merging model Z is equal to the sum of the parameters of the initial models X and Y, a valid variant of the Equation 1 is provided in Equation 2. The complexity penalty is compensated with different parameter size in both BIC hypothesis.

$$\log P(\underline{z}|Z) \geq \log P(\underline{x}|X) + \log P(\underline{y}|Y) \quad (2)$$

In the described system, the BIC-based metric is obtained reformulating the BIC decision threshold in Equation 2 to a metric formula in Equation 3. This metric is computed as the difference between the log-probabilities of the initial single models versus the merging model. The lower this $BIC(\underline{x}, \underline{y})$ is, the more probable that the two models X and Y have to be merged. High values of this metric indicates these models are likely to be different.

$$BIC(\underline{x}, \underline{y}) = (\log P(\underline{x}|X) + \log P(\underline{y}|Y)) - \log P(\underline{z}|Z) \quad (3)$$

3. Model Selection

It is now possible, when data from a new speaker is available, to select the model in the tree which best fits the speech from this incoming speaker. This selection is performed in two steps: First, the tree is pruned to a single branch, a path from the top node to one of the bottom nodes; and then, the best model from that path is chosen. Before these stages, a VAD is applied to the speech data from the speaker to discard silence in the signal, as it was done in the tree generation stage with the training speakers.

3.1. Tree Pruning

The number of nodes that the speaker tree contains is $n_{nodes} = 2 * n_{spks} - 1$. Even with a moderate number of speakers, the evaluation of all the existing models in the tree becomes a computational challenge. However, it is possible to improve this situation by pruning the tree to just one branch.

An in-depth evaluation of the tree is made with the incoming speaker and the best path is obtained evaluating its data on the tree GMM models. Starting from the top, the iterative process consist in choosing as next node-in-path the child node that maximizes the likelihood of the model X to the speaker's data $\underline{y}$: $P(\underline{y}, X)$, until a leaf is reached.

Figure 2, shows an an hypothetical pruned path, highlighted in dots.

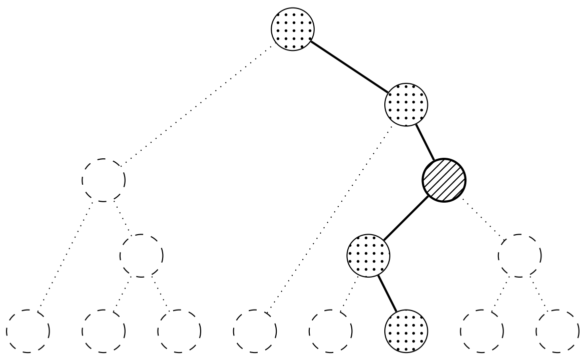


Figure 2: *Example of the selection of the best model (dotted nodes belong to the pruned tree and the stripped node is the selected node).*

3.2. Model Selection

Once the whole tree has been pruned into a single set of models, ranging from the full speaker independent model in the top to a speaker dependent model in the bottom, the best model has to be selected. This decision can be made according to different proposals: Likelihood scoring of the GMMs calculated during the training stage, likelihood scoring of the Hidden Markov Model (HMM) associated to the speech data in each node or a priori, selecting a certain node depending on its depth in the tree. In Figure 2, an hypothetically selected node is highlighted in strips.

4. Experimental Setup and Evaluation

This Section will introduce the experimental framework used in this work and the evaluation results in the speaker tree generation and lookup. The tree generation system was based in BIC criterion and GMM modeling of the nodes as explained in previous Sections. The bottom nodes (single speaker) were modeled with two GMMs, and upper nodes were the sum of the number of Gaussian distributions in their children respectively.

The feature extraction method for the tree generation was based on a standard ETSI front-end using the first 12 Mel Frequency Cepstral Coefficients (MFCC) discarding $c0$. The speech signals were windowed with a Hamming window of 25 ms length, with an overlap of 15 ms.

4.1. Experimental Corpus

The corpus used in this work was the “Alborada-I3A” corpus of disordered speech [3], that contains speech from a group of 232 unimpaired young speakers, used to build the speaker tree, and 14 young disabled speakers as potential users of the recognition system. ASR for disabled speakers is a complex task due to the many effects or their physical and cognitive impairments in their speech. For this reason, they require a proper matching to their characteristics when selecting acoustic models to be used in recognition. In our proposal, the 232 unimpaired peers will serve as reference speakers to select the best fitting model to every impaired speaker.

The group of unimpaired speakers represents the speech of individuals ranging in age 11 to 18 years old. This corpus contains one session peer speaker of the 57 words in the Registro Fonológico Inducido (RFI) [4], for a total of 13,224 isolated-word utterances.

The 14 young disabled speakers are distributed as 7 boys and 7 girls in a similar range of age (from 11 to 21 years old). Each speaker recorded 4 sessions of the RFI vocabulary, for a total of 3,192 isolated-word utterances. These speakers suffer from different developmental disorders that affect their language acquisition, resulting in a great number of mispronunciations (substitution and deletions) at the phonetic level. Physiological disorders in their vocal tract components, due to multiple physical impairments, also affect their production of speech.

4.2. Evaluating the Speaker Tree: Speaker Identification

The evaluation of the abilities of the proposed methods to build a useful speakers tree and detect correctly the best model for a given speaker was made with the following experiment. Three different trees were built with the data from the 232 reference speakers in the corpus. Each tree contained two thirds of the data (38 words) from each speaker, while the remaining third (19 words) was saved for evaluation purposes. The first tree (Set 1) comprised words 1,2,4,5,7,8... while words 3,6,9... were meant for evaluation; the second tree (Set 2) was built with words 1,3,4,6,7,9... from each speaker, keeping words 2,5,8... for evaluation; and, finally, the third tree (Set 3) had words 2,3,5,6,8,9... for generating the tree and words 1,4,7... for the evaluation.

The tree was pruned according to Section 3.1 to a single path from the top node to one of the bottom nodes. The remaining data from each speaker was evaluated through the tree as a new speaker, measuring the accuracy in which this pruned tree lead to the bottom node which was built from speech data from the same speaker.

Table 1: *Search tree accuracy.*

	Set 1	Set 2	Set 3	Average
Accuracy	79.74	79.31	81.04	80.03

The results in this task for the 3 proposed sets in Table 1 assured the method used for the tree generation and pruning as in 80.03% of the cases, the tree allowed to reach the speaker who was actually evaluating the system. In this speaker identification task, there were 232 possible competing speakers, with each speaker having just an average of 32.07 seconds of speech for training data and 16.03 seconds for the evaluation data. Only 2 Gaussian distributions formed the final nodes to model the training speaker data. This competitive performance of the tree in detecting similar speakers was encouraging in its possibilities of providing an improvement in the ASR task.

5. Use of the Speaker Tree in ASR

The purpose of the use of the tree in an ASR task is to improve the recognition rates by using an acoustic model that best matches the speech from the speaker. The experiments presented here were based on the recognition of the 14 impaired speakers on the corpus presented in the previous Section. The initial model used for the recognition was trained purely on adult speech with the 44108 noise-free signals of the adult Spanish speech databases SpeechDat-Car [5], Albayzin [6] and Domolab [7]. These data was used to train an speaker independent HMM acoustic model with a set of 744 context-dependent phonetic units and two units to model begin-end silence and interword silence; all units were modeled as 1-state units with 16 Gaussian distributions per state. 39 MFCC parameters were used for recognition, with 12 static parameters and log-energy plus their first and second derivatives. The Word Error Rate (WER) of the impaired speakers with this model was 36.69%, showing up the big influence of the impairments of the speakers in their ability to use speech recognition; while the WER of the 232 unimpaired peers was 3.99%.

The proposal divided the data from each speaker into two subsets; one for a initial development stage and the other for evaluating the new models obtained in the development stage. A set of experiments were designed to create all the 14 different possibilities using 1, 2 and 3 sessions for the development stage and the complementary 3, 2 and 1 sessions for evaluation. The purpose was to learn how the amount of data used to tune up and adapt the system influenced the performance of the recognition stage. All the adaptation processes carried out in these experiments followed a Maximum A Posteriori (MAP) implementation [8].

Table 2: *ASR results without adapting to the speaker data.*

Development data	Model	WER
N/A	From 232 speakers	28.20%
1 session	From best node	27.41%
2 sessions	From best node	27.25%
3 sessions	From best node	27.50%

From this starting point, two cases of study were evaluated. In the first case, it was considered that the data separated for development was not transcribed, hence adaptation of the models to the speaker was not possible (although unsupervised algorithms could have been applied). If no tree were available, adaptation to the whole 232 speakers was the only possible option. The result through all the speakers with the model adapted to the unimpaired children was 28.20%, as seen in first row of Table 2. When the tree was used, the best node in the tree was estimated through the evaluation of the likelihood of the speaker data to each one of the models in the pruned tree as seen in Section 3.1. Afterwards, a new model was adapted through MAP with the data of the speakers in the selected node of the tree and this model was used in the ASR stage. The average results in this case for the three possible amounts of development data are shown in Table 2, with improvements in all cases over the system which did not use the tree information.

Table 3: *ASR results adapting to the speaker data*

Development data	Model	WER
1 sessions	From 232 speakers	19.36%
	From best node	18.96%
2 sessions	From 232 speakers	16.50%
	From best node	16.50%
3 sessions	From 232 speakers	15.29%
	From best node	14.85%

The second case of study considered that the data for the initial stage was transcribed, hence it could be used to perform supervised adaptation via MAP algorithm. Again, when no information about the tree was used, the model trained from the 232 unimpaired children was used as seed for the adaptation algorithm with the specific speakers' sessions. The results in these cases are presented in Table 3, showing the more amount of data from the speaker in adaptation, the lower the WER was. Whenever the tree was used, the best node was estimated considering the likelihood of the transcribed utterances from the speaker in a HMM-based Viterbi forced alignment, comparing the HMM models assigned to each node in the pruned tree. This model served as seed in the MAP algorithm and the results are also shown in Table 3. Again, the use of the tree implied an improvement, depending of the available training datata.

6. Discussion and Conclusions

The results presented in this work showed how the proposed method produced certain improvements in terms of WER for the ASR task. The small impact of these improvements is due to the compact group of speakers available for the tree building and recognition process (all children and young adults). Further work in tasks with sets of more differentiated speakers such as adults (especially males) and children should perform better in terms of relative improvement of the WER.

The process of generating the tree was validated by means of a speaker identification task with 232 different speakers. Different utterances from the same speakers whom built the tree, were used to obtain the most likely speaker according to the tree. The 80% of accuracy showed the ability of the algorithm to build an effective speaker tree and develop a lookup system to reach the most similar nodes for an incoming speaker.

Several aspects have appeared as result of these experiments as further work. First, other techniques can be proposed for the selection of the best node to fit the speaker's data; this may include likelihood of the GMM model or the HMM model as proposed here, as well as other techniques which may consider different issues in the algorithm, such as the amount of data that would be available for creating the HMM of each node or the depth within the tree. New tasks can also be proposed where the use of a speaker tree, like the one in this work, can produce an improved performance. Its use in speaker identification or verification task could be studied in the future, as well as the possibilities in ASR situations where very sparse data from the speaker is available to train adapted models; the selection of the best speaker independent model could help to provide better recognition in rapid adaptation systems.

7. References

[1] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio, "Efficient voice activity detector algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[2] S. S. Chen and P. Gopalakrishnan, "Clustering via the bayesian information criterion with applications in speech recognition," in *Proceedings of the 1998 ICASSP*, 1998, pp. 645–648.

[3] O. Saz, E. Lleida, C. Vaquero, and W. R. Rodríguez, "The Alborada-I3A corpus of disordered speech," in *Proceedings of the 7th LREC*, 2010, pp. 2814–2819.

[4] M. Monfort and A. Juárez-Sánchez, *Registro Fonológico Inducido (Tarjetas Gráficas)*. Madrid, Spain: Ed. Cepe, 1989.

[5] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speech Dat Car. A large speech database for automotive environments," in *Proceedings of the II LREC*, Athens, Greece, June 2000.

[6] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. Mariño, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proceedings of the 1993 Interspeech)*, Berlin, Germany, September 1993, pp. 175–178.

[7] R. Justo, O. Saz, V. Guijarrubia, A. Miguel, M.-I. Torres, and E. Lleida, "Improving dialogue systems in a home automation environment," in *Proceedings of the Ambi-Sys 2008*, Québec City, Canada, February 2008.

[8] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

Dealing with Acoustic Noise and Packet Loss in VoIP Recognition Systems

José L. Carmona, Antonio M. Peinado, José L. Pérez-Córdoba, José A. González, Angel M. Gómez

Dpto. Teoría de la Señal, Telemática y Comunicaciones, University of Granada
{maqueda, amp, jlpc, joseangl, amgg} @ ugr.es

Abstract

In this paper the robustness of Network Speech Recognition (NSR) systems is analyzed. In NSR the speech signal is transmitted using a conventional speech codec from the client to the server, where the recognition task is carried out. The use of speech codecs degrades the performance of such systems, mainly in presence of acoustic noise and packet losses. First, we study the effects of possible degradation sources. Then, we propose a new NSR solution based on a robust feature extractor and an efficient packet loss concealment (PLC) algorithm, which compensates the possible degradations by means of a cepstral compensation and linear interpolation. The experimental results are obtained for a well-known speech codec, AMR 12.2 kbps, using a noisy database (Aurora-2) and several packet loss conditions. The results show that our proposal achieves noticeable improvements over the baseline results.

Index Terms: Network speech recognition, robust speech recognition, packet loss concealment.

1. Introduction

IP Packet switching networks have originated a global network of networks (Internet). Voice transmission over this type of network, called Voice over IP (VoIP), has shown strong growth during the past years, and it has turned into one of the key aspects of the current state of telecommunications. In parallel with voice and data convergence provided by VoIP platforms, new standards of wireless Internet access have led to a convergence of IP and mobile telephony networks. This paradigm will give rise to a new concept of nomadic access, hybrid of fixed and mobile access, linked to the incorporation of IP technologies and provided by suppliers of these new technologies.

Under this parardigm, automatic speech recognition offers a natural oral interaction and fast access to information. Unfortunately, there are several problems to implement a powerful automatic speech recognition subsystem into mobile terminals due to their size restrictions and limited computation capacity. Distributed speech recognition (DSR) avoids these hardware constraints by placing the most complex computational requirements of speech recognition into a remote server [1]. Moreover, the structure of a remote recognition system is well suited for the IP model, since it is the provider who implements the recognizer depending on its needs.

Although during the last years several DSR standards have been issued [2, 3], the lack of DSR codecs in the existing devices supposes a barrier for its deployment. Thus, most of the current DSR systems employ a conventional speech codec in order to transmit the speech signal to the server, where the recognition task is performed. This architecture is also known as network-based speech recognition (NSR), since the whole speech recognizer resides in the network from the client's point of view. NSR does not require any modification in the client terminal, since it uses deployed VoIP platforms. However, speech coding involves an information loss that may reduce speech recognition performance. Moreover, we have to take into account this performance reduction in presence of other implicit problems of remote speech recognition, such as acoustic noise (the acoustic context of the terminal may vary) and degradations introduced by the communication channel (packet loss for IP networks) [1].

This paper focuses on analyzing the impact of acoustic noise and packet losses on NSR systems. The NSR architecture based on decoded speech allows us to employ robust feature extractors in adverse acoustic conditions, such as the advanced front-end (AFE) proposed by the Aurora-2 working group [3]. However, as we will see, packet loss involves a drastic performance reduction in this kind of NSR systems. Thus, an analysis of the possible degradation sources is carried out prior to propose new solutions to the packet loss problem for NSR architecture working in noisy acoustic conditions.

The structure of this paper is the following. In Section II we present the experimental framework. Section III is devoted to analyze the possible degradation sources in robust NSR systems. In Section IV we propose a new framework in order to increase the robustness of NSR systems in adverse acoustic and channel conditions. Finally, in Section V we summarize our conclusions.

2. Experimental framework

The experimental setup is based on the framework proposed by the ETSI STQ-Aurora working group for the Aurora-2 database [4]. The Advanced front-end [3] provides a 14-dimension feature vector containing 13 MFCC (Mel Frequency Cepstral Coefficients) plus log-Energy. Furthermore, these vectors are extended by appending the first and second derivatives of the features. The recognizer is the one provided by Aurora and uses eleven 16-state continuous HMM word models (plus silence and pause, that have 3 and 1 states, respectively) with 3 Gaussians per state (except silence, with 6 Gaussians per state). The training and testing data are extracted from the Aurora-2 database (connected digits). Training is performed with 8400 clean sentences and test is carried out over set A. This test set contains 4 subsets (1001 sentences each) contaminated with four different types of additive noise (subway, babble, car and exhibition) at different SNRs (clean, 20, 15, 10, 5, 0 and -5 dB). For every SNR, the word accuracy (WAcc) is obtained by averaging the word accuracies of the four subsets. A mean word accuracy is computed by averaging the results obtained for all the SNRs excluding those of clean and -5 dB.

In the analysis of possible degradation sources carried out in this paper, we have used two widely used CELP-based codecs: G.729A [5] and AMR (Adaptive Multi-Rate) [6]. In addition, iLBC (internet Low Bit-rate Codec) [7] is also included, since its design is oriented to increase the robustness against packet losses.

The channel burstiness exhibited by lossy packet networks

Condition	p	q	L_{avr}	PL_r
C0	0	–	–	0%
C1	0.0526	1.0000	1	5%
C2	0.0555	0.5000	2	10%
C3	0.0588	0.3333	3	15%
C4	0.0625	0.2500	4	20%

Table 1: Packet loss conditions.

is modeled by a 2-state Markov model. The transition probabilities between states, p and q, can be set according to an average burst length (L_{avr}) and a packet loss ratio (PL_r). The performance of the NSR systems presented in this paper is tested under the channel conditions listed in Table 1.

3. Effect of Channel and Acoustic Degradations on NSR systems

The performance of NSR systems will be determined by the intrinsic codec robustness. Table 2 shows word accuracy (WAcc) results obtained by different remote speech recognition systems in packet loss conditions. The results are obtained using two training stages. The first one, called T1, refers to train the speech recognizer using original speech, i.e. non-coded speech. The second one, labeled T2, corresponds to carry out the training stage with decoded speech. As shown, the results are consistently higher for training T2, since using decoded speech in training reduces the mismatch in testing. The results obtained by the DSR system defined in [3] are included as reference. This system obtains noticeable improvements respect to NSR systems since it is specifically oriented to remote speech recognition. The WAcc result obtained directly from original speech, i.e. without using any coding scheme, is 87.74 and, thereby, the quantization stage used in [3] does not involve any performance reduction. On the contrary, there exists some performance reduction for NSR systems, although it is somewhat alleviated when the training is carried with decoded speech (T2). The speech codecs based on CELP do not achieve an optimum performance in lossy channel conditions because they use predictive techniques. For example, G.729 achieves good results in clean channel conditions transmitting linear spectrum pair (LSP) coefficients by means of a differential predictive quantifier. However, this strategy makes the codec more vulnerable to consecutive packet losses, since once a packet loss is finished, the LSP prediction is still significantly degraded. This justifies that even AMR (4.75 kbps) achieves better results than G.729 (8 kbps) for non-ideal conditions. iLBC tackles these problems by removing all types of inter-frame dependencies in the encoding process [8]. However, the price to pay is a considerable increase of bit-rate (15.2 kbps). In general, the performance of NSR is particularly lower than that of DSR when packet losses are grouped in bursts.

Speech decoders try to reduce the perceptual impact of packet losses by means of packet loss concealment (PLC) algorithms. These algorithms are usually based on repetition and progressive muting of the last received speech segment. The purpose of repetition is to conceal the effect of lost frames, whilst the progressive muting avoids the generation of annoying sounds in case of several consecutive lost frames. Nevertheless, this progressive muting leads to an increase on the insertion errors in the recognizer (artificial silences). In addition, we can observe a degradation of the decoded speech signal corresponding to correct frames after a packet loss. This degradation is inherent to the predictive nature of the encoding process of CELP-based codecs, such as G.729 and AMR 12.2 kbps. Moreover, when the decoded speech signal is used as input of a robust feature extractor, such as AFE, the error propagation is strengthened. In particular, AFE includes a noise reduction block which estimates noise characteristics in order to reduce its negative effect. In this sense, packet losses prevent this estimation process, hence a new source of degradation appears.

Our objective now is to distinguish the effects of the 'repetition and muting' effect during the burst from the 'codec error propagation' and 'AFE error propagation' after the burst. In order to do so, we can study the impact of each type of degradation on the reduction of the recognition accuracy. The following experiments are carried out by substituting speech samples and feature vectors by those corresponding to a clean transmission using AMR 12.2 kbps.

1. 'Repetition and Muting' experiment: Speech samples belonging to correctly received frames are replaced by the corresponding correct samples. Thus, the only degraded samples that remain are those where the repetition and muting algorithm is applied.
2. 'Codec Error Propagation' experiment: Speech samples belonging to lost frames are replaced by their corresponding correct samples, so that the error propagated by the speech decoder is the only remaining degradation.
3. 'AFE Error Propagation' experiment: In this case, a double substitution is carried out. First, we follow the same procedure as in the 'Repetition and Muting' experiment (that is, to replace those samples affected by codec error propagation). Second, the extracted feature vectors corresponding to the lost packets are replaced by the original ones. Thus, the remaining degradation is mainly due to the corruption of the internal states of AFE after a packet loss.

The results of these experiments are presented in Table 3. As shown, the main source of degradation is given by the 'repetition and muting' effect. PLC algorithms included in speech decoders are based on perceptual considerations that are unsuitable for recognition tasks. Nonetheless, both propagation effects that appear after a packet loss are also a considerable source of degradation. As can be observed, codec error propagation reduces the speech recognition performance at all SNR conditions, whilst AFE error propagation is more appreciable for low SNR conditions. These results are consistent, since good noise estimations are more significant for those test conditions with low SNRs. Note that AFE error propagation does not reduce the recognition performance in clean acoustic conditions, while its effect is more detrimental than codec error propagation for those conditions with SNR below 5 dB.

4. Improving NSR from decoded speech

In this section we propose a new scheme that allows us to reduce the impact of the degradation sources studied in the previous section. First, we will introduce some modifications in the Advanced Front-End (AFE) in order to make their spectral estimates more robust against packet losses. Later, we will describe a packet loss concealment oriented to speech recognition, and a compensation technique that reduces the impact of the remaining error propagation. We will test the performance of our proposal using AMR 12.2 kbps.

		iLBC		*G.729*		*AMR 12.2*		*AMR 7.95*		*AMR 4.75*		DSR	
Training		*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*
Conditions	*C0*	84.78	86.80	84.22	85.23	85.47	86.07	83.75	83.96	82.06	83.48	87.39	87.81
	C1	83.89	86.12	76.91	78.31	81.62	82.92	79.63	80.56	77.78	79.82	87.25	87.61
	C2	80.70	82.89	68.30	68.75	76.49	77.56	74.50	74.88	72.68	74.30	86.04	86.47
	C3	76.14	78.36	62.98	63.07	70.89	71.81	69.08	68.90	67.78	68.36	83.69	84.13
	C4	71.41	73.51	58.48	58.40	65.54	65.75	63.85	62.83	62.74	62.51	79.79	80.35

Table 2: Recognition accuracy (WAcc (%)) for NSR systems based on different speech codecs.

	Chan. Cond.	SNR *Clean*	*20 dB*	*15 dB*	*10 dB*	*5 dB*	*0 dB*	*-5 dB*	*Avg.*
AMR 12.2	C0	99.13	97.90	96.61	92.38	82.98	60.48	28.90	86.07
	C4	84.10	81.11	78.00	71.50	59.06	39.07	17.73	65.75
Repetition and Muting	C4	85.95	83.69	80.79	74.95	62.10	40.29	16.41	68.36
Codec Error Propagation	C4	97.69	95.49	93.54	89.32	78.96	56.79	26.42	82.82
AFE Error Propagation	C4	99.13	97.47	95.57	90.22	77.99	51.82	20.71	82.62

Table 3: Recognition accuracy (WAcc (%)) for substitution experiments using AMR 12.2 kbps and test A of Aurora-2 (clean training).

4.1. Modified AFE

In comparison with non-advanced front-ends, the AFE standard introduces two noise reduction techniques based on a Wiener filtering (WF) and an SNR-dependent waveform processing. Since the WF block is the main noise reduction technique applied in this ETSI standard we will focus on it. In particular, AFE is based on a two-stage mel-warped WF technique [9]. Its basic principle is a double WF filtering (the output of the first stage is the input to the second one). The WF filter is computed for every block of $M = 80$ samples. In order to obtain the WF coefficients is necessary to buffer 4 blocks of samples and to compute an estimate of the power spectral density (PSD) using frames of length $N_{in} = 200$ samples (between the samples 60 and 259 of the input buffer). Obviously, the buffer necessary for the WF design produces an expansion of the impact of packet loss as shown in Figure 1. The speech codec frame division is represented in the upper part. A 20-ms frame duration has been assumed, such as AMR 12.2 kbps. The two WF stages included in the noise reduction block work on subframes of 80 samples, which are represented in the second line. Finally, the bottom of the diagram represents the way that feature vectors (FV frames) are extracted from the denoised samples. Feature vectors are computed from overlapping speech segments of 25-ms length (200 samples) and 10-ms frame shift. In addition the diagram shows how subframes and feature vectors are affected by one packet loss. As can be seen, a packet loss corresponding to a speech frame of 20 ms can affect upto 9 feature vectors.

In practice the PSD estimation is carried out after applying a Hanning window of 200 samples over those samples stored in each buffer. We can assume that spectral estimates are mainly obtained using only the two central subframes of each buffer. For this reason, we can consider that a loss actually affects 7 feature vectors and its effects are more detrimental when the second stage of the noise reduction block is involved. In order to reduce the corruption of the spectral estimates, we define two non-updating indicators (NUI) by means of the following mapping functions,

$$\mathrm{NUI}_1(n) = \begin{cases} 1 & \text{if } \mathrm{PLI}(m) = 1, m \in \left(\lceil \frac{n+1}{2} \rceil, \lceil \frac{n+4}{2} \rceil\right) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\mathrm{NUI}_2(n) = \begin{cases} 1 & \text{if } \mathrm{PLI}(m) = 1, m \in \left(\lceil \frac{n-1}{2} \rceil, \lceil \frac{n+2}{2} \rceil\right) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

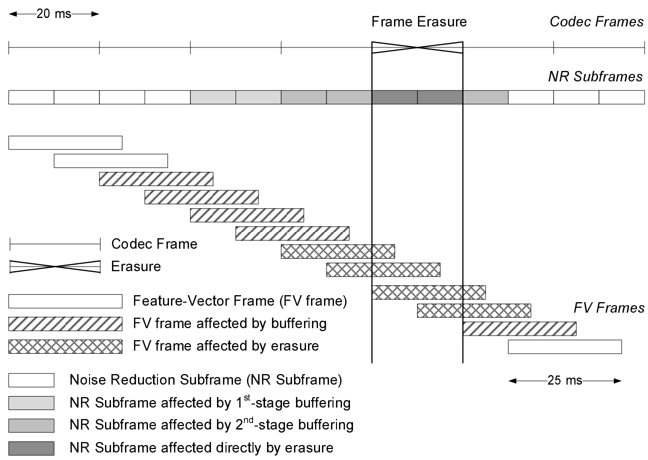


Figure 1: Feature vectors affected by a frame erasure.

where $\mathrm{PLI}(m)$ is the packet loss indicator for the codec frame m (1 for packet loss, 0 otherwise), and n is the time index of a given noise-reduction subframe. $\mathrm{NUI}_i(n) = 1$ indicates that the PSD estimate of the ith stage must not be updated for the subframe with index n, while $\mathrm{NUI}_i(n) = 0$ corresponds to normal updating.

4.2. PLC Algorithm

Although AFE propagation error is limited thanks to the modifications proposed in the previous subsection, there will still be a remaining degradation because spectral estimates have not been updated. In addition, we also have to consider that the decoded speech signal will be corrupted after a burst due to the codec error propagation [10]. We can assume that this degradation behaves as an additive noise that affects those frames after a burst. Under this approach, we can compensate this degradation by means of a cepstral normalization. In particular, we have shown [11, 10] that this remaining degradation can be effectively compensated by means of FCDCN (Fixed Codeword-Dependent Cepstral Normalization). The principle of this technique is to apply an additive correction vector **r** to the noisy feature vector **y** that depends on the length of the burst, the po-

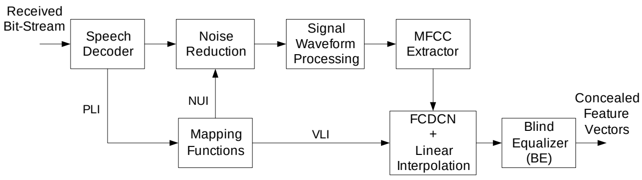


Figure 2: Block diagram of the proposed robust feature extractor including PLC techniques.

	Channel Conditions				
PLC Algorithms	*C0*	*C1*	*C2*	*C3*	*C4*
AFE	86.07	82.92	77.56	71.81	65.75
AFE+LI	–	82.68	78.31	72.87	66.99
MAFE	–	82.98	78.12	73.29	68.13
MAFE+LI	–	83.1	79.5	75.36	70.69
MAFE+FCDCN+LI	–	83.89	80.48	76.18	71.36

Table 4: WAcc results for different PLC techniques using AMR 12.2 kbps over test A of Aurora-2 (clean training).

sition after the burst, and vector **y** itself. Since the correction depends on the observed vector, this is quantized and a compensation is computed for every quantizer cell during a stereo training. This estimation can be carried out for different burst lengths by simulating as many frame erasures as needed in order to obtain a accurate compensation. Further details about this PLC compensation can be found in [11]. In order to obtain a fine representation of the cepstral space, we have used split vector quantizers with the same number of centroids than those ones employed in the ETSI DSR standard [3]. As in [11], we have considered 20 positions after every loss and a maximum burst length of 5 frames.

In addition to FCDCN compensation, it is necessary to define a PLC algorithm in order to substitute those feature vectors affected directly by a frame erasure (see Figure 1).Thus, feature vectors corresponding to lost frames can be reconstructed by means of a simple linear interpolation between the last and first correct vectors before and after a packet loss,

$$\hat{\mathbf{x}}(t) = \mathbf{x}(t_s) + \frac{\mathbf{x}(t_e) - \mathbf{x}(t_s)}{t_e - t_s}(t - t_s) \qquad t_s < t < t_e \qquad (3)$$

where $\hat{\mathbf{x}}(t)$ is the estimated feature vector at time t, and $\mathbf{x}(t_s)$ and $\mathbf{x}(t_e)$ are the last and first correct feature vectors before and after a burst, respectively. This technique has proven to be more powerful than the repetition of the nearest neighbor vector in NSR systems [10].

Figure 2 presents a block diagram of the proposed feature extractor, where the feature vectors corresponding directly to a loss are marked by a vector loss indicator (VLI). As shown, PLC techniques are inserted between the AFE blocks (noise reduction, signal waveform processing, MFCC extractor and blind equalizer). Table 4 shows the results obtained with AMR 12.2 kbps using the proposed PLC techniques. The baseline (AFE) corresponds to carrying out the recognition task from decoded speech including the PLC algorithm defined by the legacy codec. The second row shows the results obtained by applying only linear interpolation (AFE+LI). The algorithms named MAFE refer to those solutions based on the modified AFE explained in the previous section, which uses non-updating indicators in order to avoid the corruption of the AFE internal states. Finally, the results labeled as MAFE+FCDCN+LI correspond to carrying out the FCDCN compensation and linear interpolation described in this section. As shown, this last approach achieves the best results.

5. Conclusions

In this work we have analyzed the robustness of NSR systems in adverse acoustic and transmission channel conditions. In particular the analyzed NSR architecture is based on the use of decoded speech as input of the advanced front-end (AFE) defined by ETSI. First, we have identified three possible sources of degradation when a packet loss appears. These can be summarized as follows. The first one is generated by the PLC algorithms included in the speech decoder. These PLC algorithms are usually based on perceptual considerations that are not appropriate for speech recognition. The second one is the error propagation associated to those speech codecs based on the CELP paradigm. The third source of degradation is caused by the corruption of the internal states (spectral estimates) of AFE during a loss burst. Secondly, we have proposed a new framework in order to reduce the impact of these degradation sources. Our proposal is based on a modified version of AFE, which partially avoids the corruption of its internal states, and a PLC algorithm oriented to speech recognition, which is based on a cepstral compensation technique and linear interpolation.

6. Acknowledgments

This work was supported by the Spanish MEC in the project FEDER TEC2007-66600.

7. References

[1] A.M. Peinado and J.C. Segura. “Speech recognition over digital channels. Robustness and standards”, Wiley, 2006.

[2] ETSI ES 201 108. “Front-end feature extraction algorithm; Compression algorithms”, 2000.

[3] ETSI ES 202 050. “Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms”, 2005.

[4] H.G. Hirsh and D. Pearce, “The Aurora experimental framework for the performance evaluations of speech recognitions systems under noise conditions”, ISCA ITRW ASR, 2000.

[5] Recommendation ITU-T G.729. “Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited-linear-prediction”, 1996.

[6] 3GPP TS 26090. “AMR speech codec; Transcoding functions”, 1999.

[7] S. Andersen, W. Kleijn, R. Hagen, J. Linden, M. Murthi, and J. Skoglund, “iLBC - a linear predictive coder with robustness to packet losses”. in IEEE Workshop on Speech Coding, 2002.

[8] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, V. Sánchez and A.M. Gómez. “iLBC-based transparameterization: A real alternative to DSR for speech recognition over packet networks”, in Proc. of ICASSP’07, Honolulu, USA, 2007.

[9] A. Argawal and Y. Cheng. “Two-stage mel-warped Wiener filter for robust speech recognition”, in Proc. of ASRU’99, Keystone, USA, 2009.

[10] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba and A.M. Gómez. “MMSE-based packet loss concealment for CELP-coded speech recognition”, IEEE Trans. Audio Speech Lang. Process., in Press. Online: http://dx.doi.org/10.1109/TASL.2009.2033891, accessed on 18 Jun 2010.

[11] A.M. Gómez, A.M. Peinado, V. Sánchez and A.J. Rubio, “Recognition of coded speech transmitted over wireless channels”, IEEE Trans. Wireless Commun., Vol. 5, No. 9, 2006.

The L2F Broadcast News Speech Recognition System

Hugo Meinedo[1], *Alberto Abad*[1], *Thomas Pellegrini*[1], *João Neto*[1,2], *Isabel Trancoso*[1,2]

[1]L2F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
[2]Instituto Superior Técnico, Lisboa, Portugal

{hugo.meinedo, alberto.abad, thomas.pellegrini, joao.neto, isabel.trancoso}@l2f.inesc-id.pt

Abstract

Broadcast news play an important role in our lives providing access to news, information and entertainment. The existence of an automatic transcription is an important medium that not only can provide subtitles for inclusion of people with special needs or be an advantage on noisy and populated environments, but also because it enables data search and retrieve capabilities over the multimedia streams. In this work we will describe and evaluate the automatic speech recognition systems developed for two Iberian languages, European Portuguese and Spanish and also for Brazilian Portuguese, African Portuguese and English. The developed systems are fully automatic and capable to subtitling in real-time Broadcast News stream with a very small delay.

Index Terms: Speech Recognition, Broadcast News, Iberian languages, Accent, Online processing

1. Introduction

The Broadcast News (BN) processing system developed at the Spoken Language Systems Lab of INESC-ID integrates several core technologies, in a pipeline architecture: jingle detection, audio segmentation, automatic speech recognition, punctuation, capitalization, topic segmentation/indexation, summarization, and translation. The first modules of this system were optimized for on-line performance, given their deployment in the fully automatic speech recognition subtitling system that is running on the main news shows of the public TV channel in Portugal (RTP), since March 2008.

To our knowledge, the majority of subtitling systems described in the literature rely on speech-to-text alignment rather than full automatic speech recognition [1]. Re-speakers also are commonly used to simplify the original speech, and speech recognition engines are adapted to the captioner voice [2].

This paper concerns the third module in the pipeline - speech recognition, emphasizing the most recent improvements, and our efforts to port it to other languages (English and Spanish), and to other varieties of Portuguese, namely those spoken in the South American and African continents.

The development of a system for a new language is a challenging task due to the need of new acoustic training data, vocabulary definition, lexicon generation and language model estimation [3].

The paper starts with a description of the main modules of our recognition engine, emphasizing the two language independent components - feature extraction and decoder. The next three sections are devoted to the three varieties of Portuguese covered by our system: the original one (European Portuguese, henceforth designated as EP), Brazilian Portuguese (BP), and African Portuguese (AP). The porting efforts for the other two languages (European Spanish and American English) are described in Sections 6 and 7, respectively. For each of these sections, we shall detail the corpora, vocabulary, and lexical and language model generation, ending with performance results. The final section discusses the main advantages and shortcomings of these systems, namely in what concerns real time close captioning applications.

2. Automatic Speech Recognition

Our Broadcast News automatic speech recognition engine named Audimus [4, 5] is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs). A block diagram is shown in Figure 1.

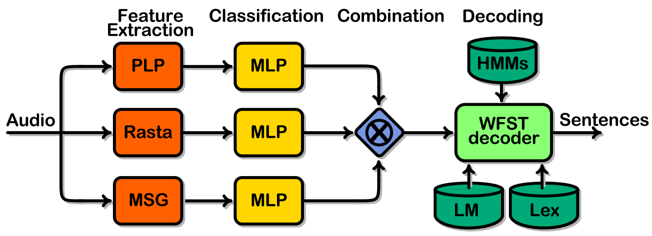


Figure 1: *Audimus block diagram.*

2.1. Feature extraction

The MLP/HMM acoustic model combines posterior phone probabilities generated by three phonetic classification branches. Different feature extraction and classification branches effectively perform a better modeling of the acoustic diversity, in terms of speakers and environments, present in BN data. The first branch extracts 26 PLP (Perceptual Linear Prediction) features, the second 26 Log-RASTA (log-RelAtive SpecTrAl) features and the 3rd uses 28 MSG (Modulation Spectrogram) coefficients for each audio frame. Each MLP classifier incorporates local acoustic temporal context via an input window of 13 frames (the MSG branch uses 15 frames) and two fully connected non-linear hidden layers. The number of units of each hidden layer as well as the number of softmax outputs of the MLP networks differs for every language. Usually, the hidden layer size depends on the amount of training data available, while the number of MLP outputs depends on the characterisitc phonetic set of each language.

2.2. Decoding process

The Audimus decoder is based on the Weighted Finite-State Transducer (WFST) approach, where the search space is a large

WFST that results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model one [6]. This decoder uses a specialized WFST composition algorithm of the lexicon and language model components in a single step. Furthermore it supports lazy implementations, where only the fragment of the search space required in runtime is computed. Besides the recognized words, the decoder outputs a series of values describing the recognition process. In order to generate a word confidence measure these features are combined through a maximum entropy classifier, whose output represents the probability of each word being correct [6]. Confidence measures for the recognized text are fundamental not only to select new acoustic training data but also to filter the output text in the subtitling composition stage.

3. European Portuguese

The initial EP acoustic model (EP baseline) was trained with 46 hours of manually annotated BN data collected from the public Portuguese TV. Currently automatically collected and transcribed data is being reused to perform unsupervised training. Recognized words that have a confidence measure above 91.5% are chosen for new training data. This is an iterative and never ending process while we get better performance with more data. The first iteration (EP iteration 1) used 378 hours of useful training speech data, 332 of which were automatically annotated using word confidence measures. The current iteration (EP iteration 2) used a total of 1000 hours of data mostly news shows from several EP TV channels. The EP MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three state monophones of the EP language plus a single-state non-speech model (silence) and 385 phone transition units which were chosen to cover a very significant part of all the transition units present in the training data.

The Language Model (LM) is a statistically 4-gram model and results from the interpolation of three specific LMs. The first is a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts, collected from the Web from 1991 to 2005 (out-of-domain corpus). The second LM is a backoff 3-gram LM estimated on a 531k word corpus of broadcast news transcripts (in-domain corpus). The third model is a backoff 4-gram LM estimated on the EP web newspapers texts collected from the previous seven days. These three LMs were linearly interpolated. For weight optimization we have used the automatically transcribed texts from the last twenty one days of news shows from RTP channels 1 and 2. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing, 100k words (or 1-gram), 7.5M 2-gram, 14M 3-gram and 7.9M 4-gram.

The EP engine uses a 100k word vocabulary adapted on a daily basis to reflect the new words that appear in web newspaper texts [7]. This daily modification of the vocabulary implies a re-estimation of the language model and retraining of the word confidence measures classifier. In order to validate the new vocabulary and language model generated, a benchmark test with one hour long news show was created, running after the daily adaptation process. This validation data is then used to retrain the confidence measure classifier in order to linearize the confidence threshold.

After the 100k word vocabulary adaptation, the pronunciation lexicon is built automatically by dividing the words into two categories. The "known" ones, for which we are able to produce a correct pronunciation, and the "unknown" ones. The correct pronunciation is either retrieved from an in house lexicon, or generated by our rule-based grapheme-to-phone (GtoP) conversion module [8]. This module can only process words which follow the Portuguese pronunciation rules, so spelled acronyms and foreign words have to be filtered out. These unknown words are then automatically split into spelled acronyms and foreign words. For spelled acronyms, rule-based pronunciations are generated. For the foreign words, a further subdivision is made, in order to identify the ones that exist in the public domain lexicon provided by CMU, for which a nativized version was produced. For the words not included in the CMU lexicon, grapheme nativization rules are applied prior to using the EP GtoP module to generate the pronunciation. The final multiple-pronunciation EP lexicon generally includes 114k entries.

Table 1 summarizes the Word Error Rate (WER) obtained in one of our BN evaluation test sets, RTP07, which is composed by six one hour long news shows from 2007. EP iteration 1 system reduced the WER by using more training data and switching to multi-state-monophones and transition units. Our current ASR, denoted EP iteration 2 significantly reduces the WER by using an extended training set with 1000 hours, larger MLPs and also the daily adaptations of the vocabulary, lexicon and language models.

Training data	Train	WER (%)
EP baseline	46 h	23.5
EP iteration 1	378 h	21.5
EP iteration 2	1000 h	18.4

Table 1: *Word Error Rates (WERs) achieved on RTP07 evaluation test set for our European Portuguese BN Recognition systems.*

4. Brazilian Portuguese

The need for porting all the key modules of the speech recognizer was first stated by using the EP BN transcription system for testing Brazilian Portuguese (BP) BN data. The result of this preliminary experience was an expected low performance of 56.6 % word error rate (WER), using the 100k vocabulary version. In order to overcome the confirmed mismatch between EP and BP, several adaptation/development steps were mandatory in the baseline speech recognizer. More concretely: the use of a GtoP module for BP in order to build new lexicon models, the development of new acoustic models based on BP data, and building new language models that could model the syntactic differences.

Details on the G2P developed for BP can be found in [9]. The performance achieved by the recognition system with the integration of the BP GtoP module was 46.2 %, which is still far from the performance achieved for EP, but already represents a significant improvement.

Due to the reduced amount of data available for training compared to EP, the size of the two nonlinear hidden layers of the MLPs is of 600 units. The process for the estimation of the monophone classification MLP networks, which corresponds to 40 outputs, consisted of several iterations of re-alignment and re-training until a stable phone classification rate is achieved in the development data set. The usefulness of the new acoustic models together with the Brazilian Portuguese GtoP was validated in the same evaluation data set of the previous experiments, achieving a WER performance of 31.6 %.

The last stage for porting the EP recognition system to a first complete BP version consists of the adaptation of the language model. We built an initial vocabulary with all the words of the transcriptions of the training corpus, and completed it with the most frequent words of the newspaper corpus, in order to achieve 100k different word forms. The next step was the automatic addition of multiple pronunciations in order to take into account some of the different variations that can be obtained as a result of word co-articulation rules.

The language model is a 4-gram backoff model created by interpolating three individual LMs built from three different sources: the CETENFolha corpus which has around 24M words, the recent newspapers corpora automatically obtained from the Internet which amounts to 18M words and the manual transcriptions of the training set. The language models were smoothed using Knesser-Ney discounting and entropy pruning. The perplexity obtained in the development set is 197.

The use of this new language model and new vocabulary together with the BP GtoP conversion module and by incorporating multiple-state phones and transitions, which caused the MLPs to have 320 outputs obtained an improvement in WER of 25.5 %. Afterwards we increased the training data with more 33 hours of automatically transcribed material and we were able to achieve 22.4 % of WER. Finally we built a new 100k vocab and LM using more 44M words of web newspaper texts which resulted in a 21.6 % WER performance.

5. African Portuguese

The EP trained speech recognizer was tested on the AP data, using the 100k vocabulary version, yielding a WER of 29.7 %, which is as expected worse than the one obtained for the EP test data, but is significantly better than the one obtained for the BP test data. Given that the GtoP module was not yet ported to AP, and no pronunciation lexicon was available for AP, our first efforts concerning porting the ASR to AP were thus restricted to porting the acoustic and the language models, starting with the latter.

The language model for AP is a 3-gram model created by interpolating three individual language models. The first is the same 4-gram LM from the EP, the second LM was built from recent AP newspapers automatically obtained from the Internet with 1.6M words and the third LM was built from the manual AP transcriptions of the training set which amount to 86k words. The language models were smoothed using Knesser-Ney discounting and entropy pruning. The perplexity obtained in the test set with only the EP model were 165.4. The use of the final AP interpolated LM resulted in perplexity of 150.6.

In terms of acoustic model training two distinct approaches were followed. The first one consisted of training new acoustic models using only the small amount of AP training data available which is around 7.5 hours and using around 17 hours of automatically detected and transcribed AP data [10]. The automatic transcription was done using the EP recognizer and selecting only words with a high confidence score. Nevertheless this very short amount of training material was the cause that we attributed for the worse performance, 27.9 % WER.

The second approach consisted of adapting the EP acoustic models using only the 7.5 hours of manually transcribed AP data. This resulted in 23.7 % WER which is a very significant improvement relative to the value obtained using only the EP acoustic models that had obtained 29.7 % WER.

6. European Spanish

In order to create an acoustic model for European Spanish (ES), an initial audio/phone alignment was necessary. We decided to transform the EP BN acoustic model output to bootstrap the ES phone alignment. A mapping was created between the ES phone set (29 phones plus silence) and EP phone set (39 plus silence) by choosing the EP phone with the most similar sound to the ES phone. With this transformed phone set, an initial ES alignment was created for the training data which consisted of 10 hours of manually annotated news shows from the national Spanish TV station (TVE). We made 4 realignment / MLP training iterations, until there was no significant improvement in the recognition results. With this ES acoustic model it was possible to automatically transcribe more data, and reuse it for training. Currently we have available an additional 148 hours of BN automatically recognized data. To improve phone classification, the 29 ES monophones where extended to multi-state-monophones (87 units), plus 112 phone transition units plus silence, totalizing 200 output units. To cope with the increased number of output units and the larger training set, the hidden layers of the MLP classifiers were enlarged to 1500 units each.

Similar to the EP the ES LM is also a statistically 4-gram model but results from the interpolation of four specific LM. The first is a 4-gram LM trained on a 1.2G word corpus of newspapers texts named Spanish Gigaword from LDC catalog. The second is also a 4-gram LM composed by data from several online ES newspapers text raging from 2001 to 2008 totalizing 72M words. This data is more recent than the Gigaword corpus and its content is similar to the BN news shows. The third is a 3-gram LM estimated on the BN training transcriptions which has 466k words. The fourth model is a 4-gram LM estimated on the ES web newspapers texts collected from the previous seven days to better cover and reflect the vocabulary adaptation. These four LMs were linearly interpolated with optimization of the weights on the automatic transcription texts from the last twenty one days of news shows from TVE channel. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing, with 100k words (1-gram), 6.6M 2-gram, 12.5M 3-gram and 10M 4-gram.

Audimus ES ASR system also uses the vocabulary adaptation process developed for the EP system. A new vocabulary containing 100k words is built in a daily base to reflect the new words that appear in web newspaper texts [7]. After the 100k word vocabulary adaptation the pronunciations lexicon is built automatically by dividing the words according to three categories. The acronyms, foreign words and “normal” words. For the “normal” words a correct pronunciation is built using in house lexica and our rule-based GtoP conversion system. For the acronyms, rule base pronunciations are generated. For the foreign words grapheme transformation rules are applyed after our ES GtoP system is used to generate the pronunciation. The final multiple-pronunciation EP lexicon includes 100k entries.

Training data	Train	WER (%)
ES baseline	10 h	22.6
ES current	158 h	15.7

Table 2: *Word Error Rates (WERs) achieved on TVE evaluation test set for our European Spanish BN recognition systems.*

Table 2 presents the evaluation results for the ES recognition system. These results were obtained in a test set with nine BN news shows from TVE station totalizing aproximately 9 hours of audio. The first line refers to the ES baseline which was trained using just the 10 hours of manually anotated audio and used only monophones. The second line shows a significant improvement by using much more acoustic training data (158 h), larger MLPs and multi-state-monophones plus phone transition units. Both systems used the same 100k vocabulary, lexicon and LM.

7. American English

The HUB-4 1996 (LDC97S44) and 1997 (LDC98S71) data sets were used to train MLP networks. Both data sets are distributed by LDC, and contain respectively 73 and 67 hours of manually transcribed speech, coming from ABC, CNN and CSPAN television networks and NPR and PRI radio networks. A set of 39 multiple-state monophones plus two single-state non-speech models (one for silence and one for breath) and 336 phone transition units (chosen to cover more than 90% of all the transition units present in the training data), was trained. The number of final recognition units (output layer size) totalizes 455 units. The use of context dependent units led to gains larger than 20% relative, when compared to context independent ("monophones") units.

For language modeling, we built one LM per source, including speech transcripts and written text sources. Nine LMs were linearly interpolated with optimization of the weights on a subset of the HUB-4 1997 training corpus used as development corpus. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing, comprised of 64k words (or 1-gram), 12 M 2-gram, 5.8M 3-gram and 4.5M 4-gram.

The 64k word vocabulary consists of all the words contained in the HUB-4 training set plus the most frequent words in the broadcast news texts and Newspapers texts. The pronunciations were extracted from the public domain lexicon provided by CMU. For words not included in this lexicon, a rule-based GtoP conversion system was used. The multiple-pronunciation lexicon included 70k entries.

Table 3 shows the best performances achieved on four official NIST test sets. The use of multi-state monophones and transition units greatly improved the performance with 15 % relative gain on average.

Corpus	Eval'97	Eval'98	Eval'99	Eval'03
WER (%)	22.0	20.4	23.3	20.6

Table 3: *Word Error Rates (WERs) achieved on four NIST evaluation test sets for our American English BN system.*

8. Conclusions

These ASR systems are the result of several years of research and development in the BN area for the Portuguese language, which were recently expanded to encompass other languages and varieties.

The main characteristic of these systems is their low latency requirements, which makes them suitable for BN subtitling. The evaluation results presented on this paper are strong and the systems have several and innovative differences. We will continue in the near future to explore these differences, mainly introducing language and variety identification, new speaker adaptation techniques, gender and bandwidth dependent acoustic models and improving language modeling and reducing vocabulary size to better accommodate the word types generated each day, in order to further improve the system performance.

9. Acknowledgements

This work was funded by FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitrio de Apoio III".

10. References

[1] A. Ortega, J. Garcia, A. Miguel, and E. Lleida, "Real-time live broadcast news subtitling system for spanish," in *Proc. Interspeech 2009*, Brighton, September 2009.

[2] "Speech Recognition in Assisted and Live Subtitling for Television, BBC R&D White Paper WHP 065," September 2003.

[3] T. Schultz, A. Black, S. Badaskar, M. Hornyak, and J. Kominek, "SPICE: Web-based tools for rapid language adaptation in speech processing systems," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[4] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in portuguese," in *Proc. ICASSP 2008*, Las Vegas, USA, 2008.

[5] H. Meinedo, "Audio pre-processing and speech recognition for broadcast news," Ph.D. dissertation, IST, Lisbon, Portugal, 2008.

[6] D. Caseiro and I. Trancoso, "A specialized on-the-fly algorithm for lexicon and language model composition," *IEEE Transactions on Audio, Speech and Lang. Proc.*, vol. 14, no. 4, Jul. 2005.

[7] C. Martins, A. Teixeira, and J. Neto, "Dynamic language modeling for a daily broadcast news transcription system," in *Proc. ASRU 2007*, Kyoto, Japan, 2007.

[8] D. Caseiro, I. Trancoso, L. Oliveira, and C. Viana, "Grapheme-to-phone using finite state transducers," in *Proc. 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, Sep. 2002.

[9] A. Abad, I. Trancoso, N. Neto, and C. Viana, "Porting an european portuguese broadcast news recognition system to brazilian portuguese," in *Proc. Interspeech 2009*, Brighton, UK, 2009.

[10] O. Koller, A. Abad, I. Trancoso, and C. Viana, "Exploiting variety-dependent phones in portuguese variety identification applied to broadcast news transcription," in *Proc. Interspeech 2010*, Makuhari, Japan, 2010.

Predictive vector quantization using the M-algorithm for distributed speech recognition

Jose Enrique Garcia, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

Communications Technology Group (GTC)
Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
jegarlai,ortega,amiguel,lleida@unizar.es

Abstract

In this paper we present a predictive vector quantizer for distributed speech recognition that makes use of a delayed decision coding scheme, performing the optimal codeword searching by means of the M-algorithm. In single-path predictive vector quantization coders, each frame is coded with the closest codeword to the prediction error. However, prediction errors and quantization errors of future frames will be influenced by previous quantizations, in such a way that choosing an instantaneous coding with the best codeword for each frame do not offer the optimal codeword sequence. The M-algorithm presents the advantage of obtaining a global minimization of the quantization error by maintaining the M-best quantization hypotheses for each frame, in a multipath coding approach outperforming the single-path predictive vector quantizer. In this work, the chosen cost function is the Euclidean distance between the sequence of prediction errors and the sequence of quantized values. The method has been tested for coding MFCC coefficients in Distributed Speech Recognition systems, making use of a non-linear predictive vector quantization on a large vocabulary task. Experimental results show that using this global optimization, lower bit rates can be achieved than using the single-path coding non-linear predictive vector quantizer without degradation in terms of WER.

Index Terms: distributed speech recognition, predictive vector quantizer, delayed decision coding, M-algorithm

1. Introduction

Distributed Speech Recognition (DSR) is the paradigm in which high performance automatic speech recognition applications (ASR) can be developed, releasing more expensive computing resources in the client side. A DSR system is composed of two main modules, the client or user module, where speech acquisition, feature extraction and feature compression are performed, and the server or recognition module, where both feature decompression and ASR decoding are carried out. DSR is usually the solution adopted when the client computing capability is limited, as it occurs in mobile devices, or just for releasing memory and processing resources in the client side, as it occurs in speech-enabled web browsing applications [1]. The main feature of DSR is that low bit-rate compression algorithms can be used without degrading the recognition accuracy. Another option for performing ASR in a client-server architecture consists of sending coded speech instead of coded acoustic features, which is known in the literature as Network Speech Recognition (NSR). However, several studies have shown that the performance is drastically reduced using state of the art speech codecs at low bit-rate conditions [1] [2]. The main reason is that most speech coding algorithms are designed for maximizing speech perceptual quality, not for maximizing speech recognition performance. Because the available network bandwidth is a scarce resource, it is convenient to use compression algorithms that provide transmission rates as low as possible, provided that recognition rates are not reduced.

Differential Vector Quantization (DVQ) is a compression method that exploits both inter-frame and intra-frame mutual information, existing in feature vectors (e.g. MFCC). On the one hand, temporal correlation between adjacent frames, due to both, the overlapping of the windowing step and the relatively slow variation of speech production, is exploited by means of linear prediction. On the other hand, intra-frame redundancy is exploited by means of Vector Quantization. A previous work [3] concluded that using DVQ in a connected digit task, a bit-rate as low as 2.1 kbps could be reached obtaining the same recognition performance than without quantization, while traditional VQ methods obtained a poor recognition performance at bit-rates lower than 3.5 kbps, even worse when noisy channels were evaluated.

The differential vector quantizer was improved by means of a non-linear predictive Vector Quantization scheme based on a Multi-Layer Perceptron (PVQ-MLP) [4]. It makes use of Artificial Neural Networks for predicting each coefficient individually using additional energy information, while prediction errors are quantized jointly by using Vector Quantization. With this non-linear predictive schema, both prediction gain and recognition accuracy improvements were reported, compared to the DVQ that makes use of an order one linear predictor.

In this work, another step for improving the compression method is presented. The proposed optimization algorithm solves the limitation of conventional single-path predictive vector quantizers, where the closest codeword for representing a single frame is chosen and sent out to the decoding side, without taking into account that future predictions can offer less quantization errors if a different codeword would be chosen. In order to tackle with this limitation, a global optimization can be done in a delayed decision coding approach, using the M-algorithm [5]. It preserves the M-best quantization hypotheses in each frame, where a minimum cost criterion is followed for choosing them. For the experiments presented in this paper, the Euclidean distance between the sequence of prediction errors and the sequence of codewords has been chosen as the cost function, however other functions for maximizing the recognition accuracy could be chosen.

The M-algorithm optimization, evaluated in the non-linear

This work has been partially funded by the national project TIN2008-06856-C05-04.

predictive vector quantizer schema, has been compared to PVQ-MLP, DVQ, VQ and the codebooks of the ETSI standard. All of them evaluated using the Advanced ETSI Front End (AFE) and Aurora 4 corpus which is a 5kword task with different acoustic environmental conditions including severe noise scenarios.

The remainder of this paper is organized as follows. First, the basics of DSR are briefly introduced in Section 2. In Section 3 an introduction to conventional predictive vector quantizers are presented, in section 4 the optimization algorithm applied to predictive vector quantizers is presented, while the experimental setup and performance evaluation are given in Section 5. Finally, the conclusions are provided in Section 6.

2. Distributed Speech Recognition

A feature compression algorithm is usually the last stage of the Front-End in DSR, in order to reduce the transmission bit-rate as much as possible. One of the most extended compression methods for DSR is Vector Quantization (VQ), which uses intra-frame redundancy of feature vectors for reducing the bit-rate providing good recognition performance [1]. The European Telecommunication Standards Institute (ETSI) has incorporated VQ as compression technique for all of its Front-End standards: ETSI 201 108, 202 050 and 202 212.

The ETSI standards Front-Ends offer 13 cepstral coefficients, and the log-energy coefficient, with noise reduction algorithms for the Advanced version and along with fundamental frequency and voicing class information in the Extended Advanced version. The compression stage is based on Vector Quantization of feature vectors pairs, resulting in 7 quantized pairs, in which C_0 is jointly quantized with log-energy, and the rest, quantized in adjacent pairs. The bit-rate obtained using this VQ is 4.4 kbps without channel error protection and without pitch and voicing class information.

However, using only VQ in the compression stage presents the main drawback that fail to exploit the strong inter-frame redundancy existing in MFCC vectors. Exploiting such inter-frame redundancy, along with intra-frame redundancy, would potentially lead to an increase in the compression rate. This can be done with a predictive vector quantization scheme, as the systems proposed in [3] [4], and the system presented in this paper. The idea of such schemes is the design of a more efficient source coding algorithm that removes all the non structured redundancy existing in the MFCCs, assuming that this new representation will be more sensitive to channel errors. However, the effect of channel errors can be neutralized by adding structured redundancy, in a lesser amount, by using channel coding techniques. In fact, the most important mobile networks (p.e. WiMaX or WiFi) provide error protection modules, and additionally for IP networks, the TCP protocol can be employed. Of course, these and other interesting known issues regarding the transmission of compressed acoustic features in ASR worth to be studied but they are beyond the scope of this work.

3. Predictive Vector Quantization of MFCC

Several compression schemes that make use of signal prediction jointly with Vector Quantization of the residual prediction error, have been successfully used in video and audio compression and, more recently, in DSR [3] [4]. Other predictive approaches for compressing acoustic features for DSR have been studied using order one linear prediction with scalar quantization in [2] and with a two-stage Vector Quantization in [6].

Predictive Vector Quantization with Multi-Layer Perceptron (PVQ-MLP) [4] performs prediction of each coefficient making use of a non-linear function that has an input layer with the latest quantized coefficients, and the latest energy quantized coefficients, outperforming DVQ [3], that employs an order one linear predictor, both in quantization error and recognition accuracy.

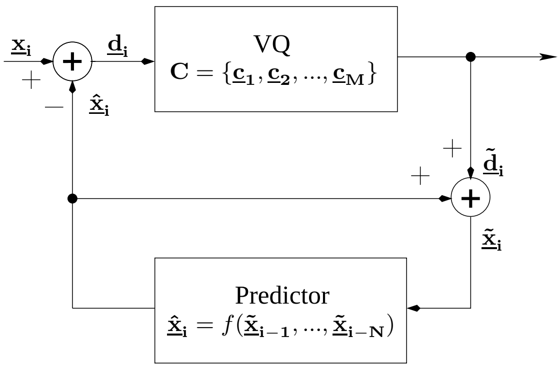


Figure 1: *Block Diagram of a predictive vector quantization schema.*

The scheme for compressing a group of coefficients with a differential vector quantizer is shown in figure 1. For the $\mathbf{i^{th}}$ frame, each group of cepstral coefficients is denoted as $\underline{\mathbf{x}}_\mathbf{i}$. Over this tuple, a prediction is done by using the previous quantized values,

$$\hat{\underline{\mathbf{x}}}_\mathbf{i} = f(\tilde{\underline{\mathbf{x}}}_\mathbf{i-1}, ..., \tilde{\underline{\mathbf{x}}}_\mathbf{i-N}), \tag{1}$$

where $\mathbf{N}$ is the predictor order.

In a conventional differential vector quantizer, the prediction error, $\underline{\mathbf{d}}_i = \underline{\mathbf{x}}_\mathbf{i} - \hat{\underline{\mathbf{x}}}_\mathbf{i}$, is quantized by means of a codebook composed of $\mathbf{L}$ codewords, $\mathbf{C} = \{\underline{\mathbf{c}}_\mathbf{1}, \underline{\mathbf{c}}_\mathbf{2}, ..., \underline{\mathbf{c}}_\mathbf{L}\}$. The quantization procedure consists of choosing, for each frame, the closest codeword $\underline{c}_j$, using the Euclidean distance:

$$\tilde{\underline{\mathbf{d}}}_\mathbf{i} = \arg\min_{\underline{c}_j}\{|\underline{c}_j - \underline{\mathbf{d}}_\mathbf{i}|^2\} = \underline{\mathbf{d}}_\mathbf{i} + \underline{\mathbf{e}}_\mathbf{i} \tag{2}$$

where $\underline{c}_j$ is the $\mathbf{j^{th}}$ codeword, that will be sent out to the decoder side, and $\underline{\mathbf{e}}_\mathbf{i}$ is the quantization error. The quantized prediction error $\tilde{\underline{\mathbf{d}}}_\mathbf{i}$ is used to obtain the reconstructed coefficients $\tilde{\underline{\mathbf{x}}}_\mathbf{i} = \hat{\underline{\mathbf{x}}}_\mathbf{i} + \tilde{\underline{\mathbf{d}}}_\mathbf{i}$, that are also obtained in the decoder, and employed for predicting the forthcoming frames using (1).

Note that the reconstructed coefficients can be also expressed as

$$\tilde{\underline{\mathbf{x}}}_\mathbf{i} = \underline{\mathbf{x}}_\mathbf{i} + \underline{\mathbf{e}}_\mathbf{i}, \tag{3}$$

where it can be observed that the quantization error of the coefficients is the same than the quantization error of the prediction error.

4. M-algorithm optimization for Predictive Vector Quantization

Let $\mathbf{X} = \{\underline{\mathbf{x}}_\mathbf{1}, ..., \underline{\mathbf{x}}_\mathbf{t}, ..., \underline{\mathbf{x}}_\mathbf{T}\}$ be the original $\mathbf{T}$ frame sequence of coefficients to be quantized, $\tilde{\mathbf{X}} = \{\tilde{\underline{\mathbf{x}}}_\mathbf{1}, ..., \tilde{\underline{\mathbf{x}}}_\mathbf{t}, ..., \tilde{\underline{\mathbf{x}}}_\mathbf{T}\}$ the reconstructed coefficient sequence, $\tilde{\mathbf{D}} = \{\underline{\mathbf{d}}_\mathbf{1}, ..., \underline{\mathbf{d}}_\mathbf{t}, ..., \underline{\mathbf{d}}_\mathbf{T}\}$ the prediction error sequence, and $\mathbf{U} = \{\underline{\mathbf{u}}_\mathbf{1}, ..., \underline{\mathbf{u}}_\mathbf{t}, ..., \underline{\mathbf{u}}_\mathbf{T}\}$ the sequence of chosen codewords sent out to the decoder side, where $\underline{\mathbf{u}}_\mathbf{t} \in \mathbf{C}$.

The minimum squared quantization error for the whole sequence can be computed as:

$$\xi = \min_{\{\tilde{\mathbf{x}}_1,...,\tilde{\mathbf{x}}_\mathbf{T}\}} \sum_{t=1}^{\mathbf{T}} |\underline{\mathbf{x}}_t - \underline{\tilde{\mathbf{x}}}_t|^2 = \min_{\{\mathbf{u_1},...,\mathbf{u_T}\}} \sum_{t=1}^{\mathbf{T}} |\underline{\mathbf{d}}_t - \underline{\mathbf{u}}_t|^2 \quad (4)$$

However, note that for the instantaneous decision method (2) used in single-path predictive vector quantizers, there is no guarantee that ξ could be obtained due to the fact that,

$$\min_{\{\mathbf{u_1},...,\mathbf{u_T}\}} \sum_{t=1}^{\mathbf{T}} |\underline{\mathbf{d}}_t - \underline{\mathbf{u}}_t|^2 \leq \sum_{t=1}^{\mathbf{T}} \min_{\mathbf{u_t}} |\underline{\mathbf{d}}_t - \underline{\mathbf{u}}_t|^2, \quad (5)$$

where the second term in (5) is the squared error obtained by a typical predictive vector quantizer that performs decisions in a frame by frame basis, as in (2). The inequality (5) is valid for predictive vector quantizers, since the term $\underline{\mathbf{d}}_t$ (containing the prediction error for frame $\mathbf{t}$) depends on previous codeword decisions, and previous values of the signal,

$$\underline{\mathbf{d}}_t = h(\underline{\mathbf{u}}_{\mathbf{t-1}}, ..., \underline{\mathbf{u}}_{\mathbf{1}}, \underline{\mathbf{x}}_{\mathbf{t}}, \underline{\mathbf{x}}_{\mathbf{t-1}}, ..., \underline{\mathbf{x}}_{\mathbf{1}}). \quad (6)$$

Note that in (5), the equality holds if $\underline{\mathbf{d}}_t$ is memoryless.

The problem that we want to solve is to choose the codeword sequence $\mathbf{U}$ that minimizes the Euclidean distance between the sequence of original coefficients $\mathbf{X}$ an the sequence of reconstructed coefficients $\tilde{\mathbf{X}}$. The exact solution to this coding problem could be obtained by using a brute force approach, computing the Euclidean distance for all possible codeword sequences $\mathbf{U}$. However, that is computationally intractable even for small values of $\mathbf{T}$ and $\mathbf{L}$.

In this paper we make use of the M-algorithm [5] in order to get an approximate solution to this problem that performs better than the single frame decision. The method consists of a synchronous evaluation algorithm, where in a frame by frame basis the M-best hypotheses (with minimum accumulated cost) are mantained. Before frame $\mathbf{t}$ is evaluated, a hypothesis is composed of an accumulated cost $\mathbf{a_{t-1}}$, an index history $\underline{\mathbf{i}}_{\mathbf{t-1}} = \{\mathbf{i_1}, ..., \mathbf{i_{t-1}}\}$, and a history of reconstructed coefficients $\underline{\tilde{\mathbf{r}}}_{\mathbf{t-1}} = \{\underline{\tilde{\mathbf{x}}}_{\mathbf{1}}, ..., \underline{\tilde{\mathbf{x}}}_{\mathbf{t-1}}\}$, For each one of the $\mathbf{M}$ active hypothesis at frame $\mathbf{t}$, an instantaneous prediction error is extracted:

$$\underline{\mathbf{d}}_{\mathbf{t}} = \underline{\mathbf{x}}_{\mathbf{t}} - f(\underline{\tilde{\mathbf{x}}}_{\mathbf{t-1}}, ..., \underline{\tilde{\mathbf{x}}}_{\mathbf{t-N}}) \quad (7)$$

With this prediction error, the instantaneous Euclidean distance $\mathbf{o_{t,j}}$ is obtained for each codeword index $\mathbf{j} = \mathbf{1}..\mathbf{L}$,

$$\underline{\mathbf{o}}_{\mathbf{t,j}} = |\underline{c}_j - \underline{\mathbf{d}}_{\mathbf{t}}|^2 \quad (8)$$

Finally, the accumulated cost for that hypotheses propagated through the codeword index $\mathbf{j}$ is,

$$\mathbf{a_{t,j}} = \mathbf{a_{t-1}} + \underline{\mathbf{o}}_{\mathbf{t,j}} \quad (9)$$

If the evaluated hypothesis is selected as valid, the accumulated cost, the index history and the reconstructed coefficients are updated. The total number of prediction hypotheses that are evaluated at frame $\mathbf{t}$ become $\mathbf{ML}$. However, only the M-best hypothesis (with less accumulated cost $\mathbf{a_{t,j}}$) are conserved for processing the next frame. When the last frame $\mathbf{T}$ is reached, the index history $\underline{\mathbf{i}}_{\mathbf{T}}$ with the lowest accumulated cost $\mathbf{a_T}$ is sent out to the receiver side, that performs the reconstruction of the coefficients like in a conventional predictive vector quantizer.

In a typical single-path predictive vector quantizer the codebook index of a quantized frame is chosen in a frame by frame basis independently of future quantizations (2), and only one prediction hypothesis is conserved in each frame. The optimization algorithm for $\mathbf{M} = \mathbf{1}$ is equivalent to the typical single-path predictive vector quantizer, however, with higher values of $\mathbf{M}$ lower quantization errors are obtained.

Table 1: Different bit-allocations for different bit-rates explored from 1.4 to 2.0 kbps

Bit-rate	C_1C_2	C_3C_4	C_5C_6	$C_7C_8C_9$	$C_{10}C_{11}C_{12}$	C_0 E
1.4	3	3	2	2	2	2
1.6	3	3	3	3	2	2
1.8	3	3	3	3	3	3
2.0	4	4	3	3	3	3

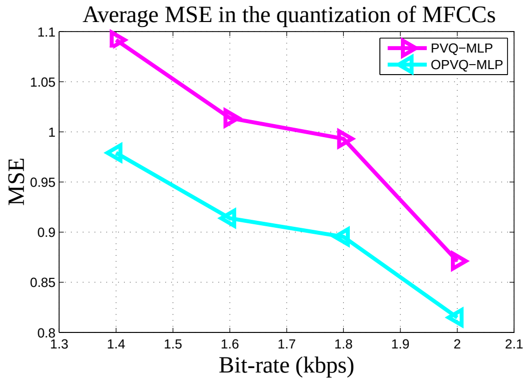


Figure 2: Average MSE in the quantization of MFCCs, in the test01 of Aurora 4

5. Performance Evaluation

In order to evaluate the performance of the proposed optimized predictive vector quantization method, OPVQ-MLP an extensive set of recognition experiments was carried out on a large vocabulary task, under different channel and noise conditions. The presented quantization scheme has been compared to the rest of quantization techniques exposed in [3][4] (PVQ-MLP, DVQ, variable length VQ, and the fixed length ETSI VQ).

The number of hypotheses per frame in the optimization algorithm OPVQ-MLP, $\mathbf{M}$, was fixed to 10, in a trade-off between computational complexity and quantization error performance, since it was observed that higher values of $\mathbf{M}$ did not reduce significantly the quantization error. The optimization algorithm was applied to each sub-vector group individually, in such a way that there was 10 hypothesis by frame and group. In the OPVQ-MLP and PVQ-MLP methods, MFCC sub-vectors were grouped as shown in Table 1, for testing bit-rates between 1.4 and 2.0 Kbps. For testing bit-rates between 700 and 4200 bps in the methods PVQ-MLP, DVQ and VQ, MFCC sub-vectors were grouped by pairs, as defined in the ETSI standard encoder, and the bit-rate was obtained assigning the same number of bits for each one of the 7 pairs, in such a way that using 1,2,3,4,5,6 bits by pair, a bit-rate of 700, 1400, 2100, 2800, 3500, 4200 bps is obtained, since a 100 frames per second rate is considered,

For all the experiments, the codebooks of OPVQ-MLP, PVQ-MLP, DVQ, and VQ were trained using different numbers of codewords under different train/test conditions. For training all the quantizers, we used the same training set that the one

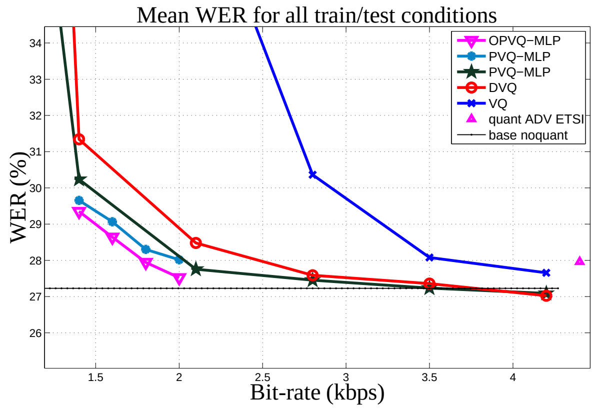


Figure 3: Mean results for all train-test combinations

used for training the acoustic models, so acoustic models were always adapted to the compression algorithms under all conditions.

The experiments for the ASR performance evaluation were carried out with the 8 kHz part of Aurora 4 database [7], designed by the Aurora Working Group of the ETSI. This database was conceived for developing robust Front-Ends and speech processing modules to be used in DSR systems. It is composed of a 5kword vocabulary based on DARPA Wall Street Journal (WSJ0) and contains 3 training sets (with 7138 utterances each one) and 14 test sets (with 330 utterances each one). Several acoustic environments are defined for composing 3 different train sets.

The recognizer and training tool employed for all the experiments was HTK, using a similar setup to that used in HIWIRE project for evaluating Aurora 4 database [8], that is, ETSI Advanced Front-End (AFE), cross-word tree-based tied-state triphones for acoustic models, with 3 states in each unit, and a GMM of 6 components for modeling the observation probability in each state. The language model employed was a back-off bigram.

Fig. 2 shows the quantization error for both methods with bit-rates ranging from 1.4 kbps to 2.0 kbps. It can be seen that OPVQ-MLP curve is always under PVQ-MLP curve, for all the evaluated bit-rates showing that a better quantization has been obtained thanks to the proposed optimization algorithm. These curves were obtained with test number one of Aurora 4, and the codebooks trained with clean signal. However, for all train-test combinations explored, the quantization performance is better for the OPVQ-MLP method than for the PVQ-MLP method.

Fig. 3 shows the mean Word Error Rate for all the experimental conditions described before (each one of 42 combinations train set - test set). As it can be seen, the DVQ performance is superior to the one obtained with conventional VQ methods for all code-book lengths, and OPVQ-MLP schema outperforms PVQ-MLP, DVQ and VQ methods. The degradation of DVQ method compared to a system without quantization is small for bit-rates above 2.1 kbps. However, the PVQ-MLP method can reach a bit-rate as low as 1.8 kbps with similar WER to ETSI quantizer, at 4.4 kbps, and slightly better than WER achieved by DVQ at 2.1 kbps, and the proposed OPVQ-MLP method can reach similar bit-rates than PVQ-MLP, but with less WER.

In comparative terms, it is worth pointing out that OPVQ-MLP at 1.6 kbps, PVQ-MLP at 1.8 kbps and DVQ at 2.1 kbps perform as well as VQ at 3.5 kbps and the AFE compression method at 4.4 kbps, with a small WER degradation over the baseline. Respect to the recognition results in different conditions It was observed that the behavior of the compression methods for different matching conditions is very homogeneous in comparative terms.

6. Conclusion

In this paper, a delayed decision coding algorithm for predictive vector quantization, the M-algorithm, has been evaluated in recognition experiments on a large vocabulary task, using Aurora 4 database. This algorithm extracts an optimal codeword sequence in the quantization process, in an efficient way, without evaluating all possible codeword combinations by maintaining the M-best hypotheses for each frame. Experimental results with M=10 showed that the proposed OPVQ-MLP quantizer outperforms the method PVQ-MLP, that decides the current codeword that must be sent to the back end in a single frame decision approach. The bit-rate that can be reached using OPVQ-MLP is 1.6 Kbps with a 5.1% of relative WER degradation with respect to an ASR system without quantization. Similar results in terms of WER can be obtained with the ETSI standards compression method, which makes use of 4.4 kbps, implying 175% of bandwidth increase relative to the proposed 1.6 kbps OPVQ-MLP method.

This study shows that the M-algorithm, in a delayed decision coding approach for predictive vector quantizers, can be used in order to get a minimum cost global optimization. In addition, a reduction in the necessary bit-rate without WER degradation has been reported on a large vocabulary task under different noise conditions, which highlights the benefits of using this method as compression stage in Distributed Speech Recognition.

7. References

[1] V. Digalakis, L. Neumeyer, and M. Perakakis, “Quantization of cepstral parameters for speech recognition over the world wide web,” in *proceedings of Icassp*, 1998.

[2] N. Srinivasamurthy, A. Ortega, and S. Narayanan, “Efficient scalable encoding for distributed speech recognition,” *Speech Communication*, 2006.

[3] J. E. Garcia, A. Ortega, A. Miguel, and E. Lleida, “Differential vector quantization of feature vectors for distributed speech recognition,” in *Interspeech*, 2009.

[4] J. E. Garcia, A. Ortega, A. Miguel, and E. Lleida, “Non-Linear Predictive Vector Quantization of Feature Vectors for Distributed Speech Recognition,” in *Interspeech*, 2010.

[5] F. Jelinek and J.B.Anderson, ”Instrumentable Tree Encoding of Information Sources”, IEEE Trans. on Information Theory, pp. 118-119, January 1971

[6] G. N. Ramaswamy and P. S. Gopalakrishnan, “Compression of acoustic features for speech recognition in network environments,” in *Icaasp*, 1998.

[7] G. Hirsch, “Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task,” ETSI STQ Aurora DSR Working, Tech. Rep., June 2001.

[8] J. C. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P. A. Breton, R. G. V. Clot, M. Matassoni, and P. Maragos, “The HIWIRE database, a noisy and non-native english,” Hiwire consortium, Tech. Rep., april 2007. [Online]. Available: http://www.hiwire.org

Speech signal- and term-based feature contribution to hit/false alarm classification in a spoken term detection system

Javier Tejedor[1]*, Doroteo T. Toledano*[2]*, Miguel Bautista*[2]*, José Colás*[1]

[1]Human Computer Technology Laboratory,
[2] ATVS-Biometric Recognition Group,
Universidad Autónoma de Madrid, Spain
javier.tejedor@uam.es

Abstract

There are many factors that lead to decrease the final performance on spoken term detection (STD) systems. They are mainly related to the properties of the terms to be searched, the speech signal conditions and so on. This paper proposes and analyses a set of factors that can enhance or disminish the hit/false alarm (FA) ratio based on certain features. Our study reflects that detections corresponding to short-length terms, detections corresponding to a term similar to some other, short duration detections and lower confidence values assigned to each putative detection can lead to a FA whereas the opposite is shown to correspond to a hit in an open-vocabulary STD system.

Index Terms: spoken term detection, feature analysis, speech recognition.

1. Introduction

Speech information retrieval has received much interest for years, focusing on finding relevant information from audio archives. It encouraged many groups to develop practical systems [1–5] and NIST to conduct the first Spoken Term Detection (STD) evaluation [6], which aims at finding a list of terms fast and accurately in huge audio repositories. The standard STD architecture consists of a Speech Recogniser to produce word/sub-word lattices, a Term Detector to hypothesise putative detections and a Confidence Measure component to decide if each putative detection is reliable, as it is depicted in Figure 1.

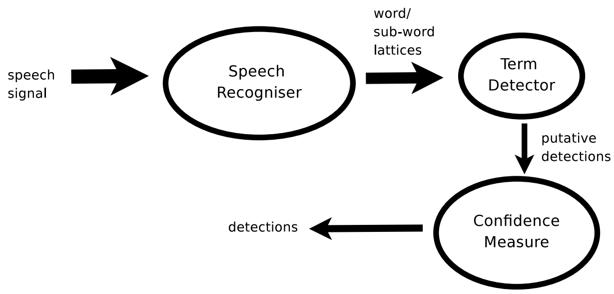


Figure 1: *The standard STD architecture.*

The *Confidence Measure* component plays a very important role in STD systems. It examines each putative detection and decides if it is considered to be a hit or a false alarm (FA). A *hit* occurs when a hypothesised detection appears in the speech signal. A *FA* occurs when the detection does not appear in the speech signal. An occurrence which is not hypothesised by the system is called a *miss*. Most of the works related to STD have proposed different confidence measures from which the final STD performance, in terms of ATWV (Actual Term Weighted Value, defined by NIST [6] for the STD task) and DET curves [7], is enhanced. Some are based on the scores produced by the speech recogniser [8,9]. Other such as n-best lists [10,11], minimum edit distance [12,13] and discriminative confidence [14–16] have been also explored. However, these works hardly make any analysis about which term properties or feature values derived from the speech signal are more likely to produce more hits or FAs. Actually, this hit/FA tradeoff measures the system performance. Therefore, this work aims at proposing a putative set of features, mainly term-based features, detection-based features and speech signal-based features and analyses their influence in the final hit/FA ratio. It must be noted that there are related works [17,18] which analyse the Word Error Rate (WER) contribution of individual words in an Automatic Speech Recognition (ASR) Large Vocabulary Continuous Speech Recognition (LVCSR) system. Our work is slightly different since we analyse the performance, in terms of hits and FAs in an open-vocabulary STD task. In addition, new features are also proposed and explored for this STD task.

The rest of the paper is organised as follows: Section 2 describes the sets of features explored in this work. Section 3 presents the experimental setup. An histogram-based analysis and linear regression-based analysis are presented in Section 4 and Section 5 respectively. Finally, the work is concluded in Section 6.

2. Feature class description

Inspired by the previous works [17,18], the following sets of features have been studied:

- Lattice features: This set of features comprises: the lattice-based confidence (score) for each detection (i.e., $c_f(d_i^K)$, computed as in [19] from standard forward-backward recursions), R0 (i.e., the effective occurrence rate for each term defined by Equation 1) and R1 (i.e., the effective false alarm rate for each term defined by Equation 2).

$$R_0(K) = \frac{\sum_i c_f(d_i^K)}{T} \tag{1}$$

$$R_1(K) = \frac{\sum_i (1 - c_f(d_i^K))}{T} \tag{2}$$

where $c_f(d_i^K)$ represents the lattice-based confidence of the i-detection of the term K and T is the total length of the audio.

- Lexical features: This set of features contains the total number of graphemes, phones, vowel graphemes, consonant graphemes, vowel phones and consonant phones for each term.
- Levenshtein distance features: The maximum, minimum and mean Levenshtein distance for each term against the others.
- Duration features: This set of features contains the duration of each detection, the duration divided by the number of phones (phone speech rate) and divided by the number of vowels (vowel speech rate) of each detection.
- Position: It represents if the detection appears the first in the lattice, the last in the lattice or in any other position.
- Prosodic features: They contain the pitch (maximum, minimum and mean pitch for each detection), the intensity (maximum, minimum and mean intensity for each detection) and the voicing percentage (i.e., the percentage of voiced speech for each detection in the speech signal). These features were collected using Praat [20].

The new features introduced in this work compared with the previous works [17,18] are the lattice-based features, all the lexical-based features except the number of phones, the Levenshtein distance features, the vowel speech rate within the duration features and the voicing percentage within the prosodic features.

3. Experimental setup

The geographical domain of the Albayzin database [21] was used for the experiments. 500 OOV terms, selected from the geographic corpus, which amount 12651 occurrences in the geographic training set, were used as list of terms. They were chosen based on their number of occurrences in this set.

A phone-based system was built from the HTK tool [22] in N-best mode to produce the phone lattices. It used state-clustered triphone models and 39-dimensional MFCC features. A bigram was used as LM trained from the phonetic training set of the Albayzin database. A grapheme-to-phone conversor was used to predict pronunciations for the *OOV terms*. As term detector, we used the *Lattice2Multigram* tool developed by Brno University of Technology (BUT), which hipothesises dectections based on an exact match of the phone transcription of each term and the paths in the phone lattice.

The STD system was run on the 500 OOV terms and the geographic training set and detections were labeled as hit or FA to carry out the analysis of which features are more likely to produce hits and FAs.

4. Histogram-based analysis

Each individual set of features explained in Section 2 is analysed from a histogram by plotting each feature contained in each group as it is presented in Figures 2-6. Inspecting the Figure 2, we see that, as expected, hits posses a higher score than FAs since it actually corresponds to the confidence assigned to each detection. Therefore, detections with higher scores are more likely to be hits and detections with lower scores should be considered as FAs. Consistent results are observed from the R0 and R1 features since terms with higher R0 and lower R1 are more likely to produce hits than FAs due to the former represents the effective occurrence rate and the latter represents the effective false alarm rate. Inspecting the Figure 3, where the lexical features per term are plotted, it can be seen that short-length terms (both in terms of phones are graphemes) are more likely to produce more FAs than long-length terms since the former can be a part of a long term or even a concatenation of the end and beginning of two different terms. This analysis is also consistent with the number of vowels and number of consonants (both for phones and graphemes). From the Figure 4, we observe that terms with a lower mean Levenshtein distance are more likely to be confused with some other and therefore they will produce more FAs than terms with higher mean Levenshtein distance, whose confusability with the rest is lower. However, extreme values (i.e., those derived from the maximum and minimum Levenshtein distances), does not separate hits from FAs in such a way that any clear conclusion is reached. Inspecting the duration-based features in Figure 5, we can see that a detection with a shorter duration is more likely to be a FA than a detection with longer duration, both in terms of absolute duration, phone speech rate and vowel speech rate. This may be due to many times FAs are produced by speech recognition errors that tend to cause awkward durations. The position of each detection found in the lattice does not discriminate between hits and FAs at all and therefore, the two plots are mostly overlapped in Figure 5. Finally, inspecting the Figure 6, we can see that detections corresponding to speech signal intervals with low intensity are more likely to be hits than FAs since higher values of minimum intensity may be caused by poor speech signal conditions and that the rest of the prosodic features do not discriminate between the hit and FA classes at all.

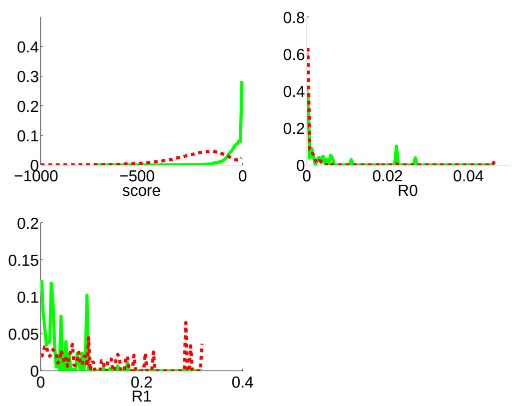


Figure 2: *Histogram analysis for the lattice-based features. Green bars represent hits and red bars represent false alarms.*

5. Linear regression-based analysis of variance

As an alternative analysis to the one presented in the former section, in this section we perform an analysis based on linear regression in which we analyse the amount of variance in the binary variable hit/FA, represented as a 1 or a 0, that can be explained by a linear regression using each of the individual features defined in Section 2. This analysis is performed using the stepwise function of MATLAB and computing the R^2 statistic.

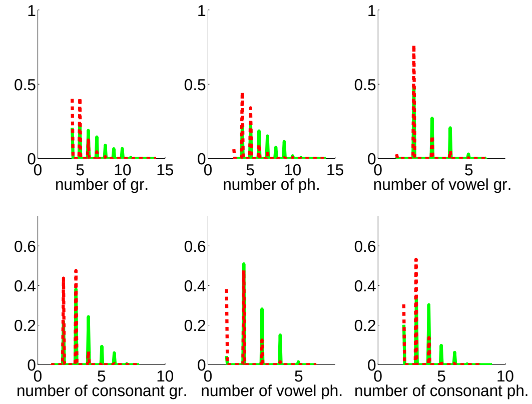


Figure 3: *Histogram analysis for the lexical features.* ph. *denotes phones and* gr. *denotes graphemes. The layout is the same as in Figure 2.*

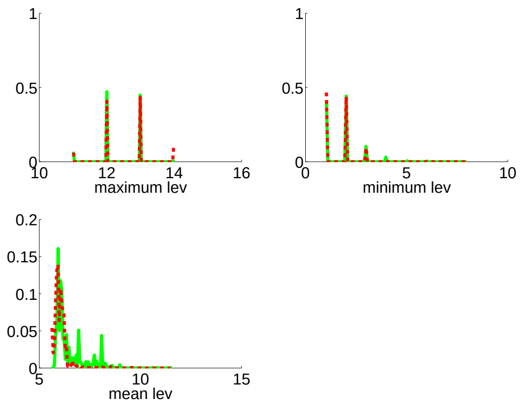


Figure 4: *Histogram analysis for the Levenshtein (lev) distance-based features. The layout is the same as in Figure 2.*

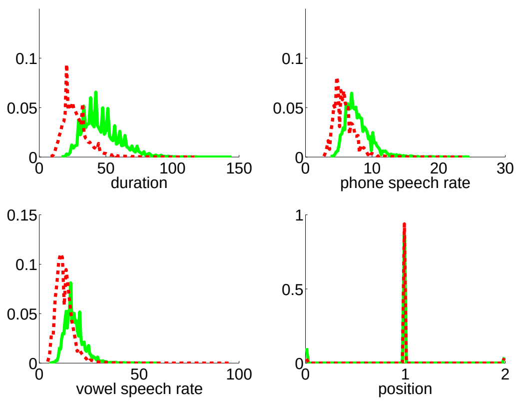


Figure 5: *Histogram analysis for the duration- and position-based features. The layout is the same as in Figure 2.*

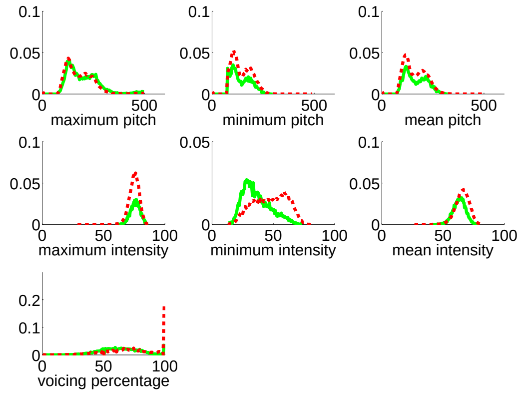


Figure 6: *Histogram analysis for the prosodic features. The layout is the same as in Figure 2.*

A similar approach was successfully used in [23] to choose the set of features that provides more information to discriminate between hits and FAs. There we showed that the conclusions obtained from the multiple linear regression analysis were in accordance with results obtained with a more complex (neural network) confidence estimator. Here our interest is different, because we are not interested in training a confidence estimator, but in determining the most interesting features in isolation. For this reason we do not group the features as we did there and only the percentage of reduction of variance achieved by using a single feature is analysed. Results of these analyses on the same set used in Section 4 are presented in Table 1.

This analysis yields basically the same conclusions obtained in the previous section, but with a numerical result that can be used to compare the amount of information provided by each individual feature in a more principled manner than by looking at the amount of overlapping of the histograms. Therefore, those feature histograms with a less overlapping between hit and FA classes lead to a higher R^2 contribution, which derives in a better hit/FA discrimination. Not surprisingly, the score is the feature that provides with the highest R^2, since it represents the confidence that the detection is considered to be a hit. It is consistent with the histogram-based analysis, where the score possesses the best hit/FA discrimination among all the features explored in this work. On the other hand, when the R^2 contribution of a certain feature is small, the histogram reveals a high degree of overlapping, meaning that such feature does not disriminate between both classes at all.

6. Conclusions

This work has investigated the individual contribution to the hit/FA classification in an STD system of both term- and detection-dependent properties and speech signal-based features. It has been shown that short terms are more likely to produce more errors and therefore more FAs in STD systems. Terms which posses a similar phone sequence are more likely to be confused with each other, leading to an increase in the FA rate, and short duration detections also contribute with a high FA rate.

Future work will investigate new features based on the most informative ones explored in this work, since it has been shown that lattice-, duration- and Levenshtein distance-based features

Feature Class	Feature	R^2 (%)
Lattice	score	42.48
Lattice	R0	4.06
Lattice	R1	20.29
Lexical	Number of graphemes	15.31
Lexical	Number of phones	18.81
Lexical	Number of vowel graphemes	10.97
Lexical	Number of consonant graphemes	13.15
Lexical	Number of vowel phones	20.58
Lexical	Number of consonant phones	8.59
Levenshtein distance	Maximum	1.23
Levenshtein distance	Minimum	1.02
Levenshtein distance	Mean	11.33
Duration	Duration of each detection	38.73
Duration	Phone speech rate	23.57
Duration	Vowel speech rate	21.84
Position	Position	0.98
Prosodic	Maximum Pitch	1.44
Prosodic	Minimum Pitch	0.25
Prosodic	Mean Pitch	0.10
Prosodic	Maximum Intensity	1.04
Prosodic	Minimum Intensity	17.12
Prosodic	Mean Intensity	5.64
Prosodic	Voicing percentage	4.49

Table 1: *Linear Regression analysis of variance results. Results show the R^2 statistic in percentage attributed to each feature, which can be interpreted as the percentage of variance explained by each particular feature.*

and lexical and prosodic features make an important contribution to hit/FA classification.

7. Acknowledgements

This work has been funded by the CAM S2009/TIC-1542 MA2VICMR project.

8. References

[1] I. Szöke, M. Fapšo, M. Karafiát, L. Burget, F. Grézl, P. Schwarz, O. Glembek, P. Matějka, S. Kontár, and J. Černocký, "BUT system for NIST STD 2006 - English," in *Proc. NIST Spoken Term Detection Evaluation workshop (STD'06)*. Gaithersburg, Maryland, USA: National Institute of Standards and Technology, December 2006.

[2] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 2393–2396.

[3] S. Parlak and M. Saraçlar, "Spoken term detection for Turkish broadcast news," in *Proc. ICASSP'08*, Las Vegas, Nevada, USA, March 2008, pp. 5244–5247.

[4] C. Parada, A. Sethy, and B. Ramabhadran, "Balancing false alarms and hits in spoken term detection," in *Proc. ICASSP'10*, vol. 1, March 2010, pp. 5286–5289.

[5] J. Mamou and B. Ramabhadran, "Phonetic query expansion for spoken document retrieval," in *Proc. Interspeech'08*, Brisbane, Australia, September 2008, pp. 2106–2109.

[6] NIST, *The spoken term detection (STD) 2006 evaluation plan*, 10th ed., National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, September 2006. [Online]. Available: http://www.nist.gov/speech/tests/std

[7] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *Proc. Eurospeech*, September 1997, pp. 1895–1898.

[8] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting," in *Proc. ICASSP'89*, Glasgow, UK, May 1989, pp. 627–630.

[9] S. Cox and R. Rose, "Confidence measures for the SWITCHBOARD database," in *Proc. ICASSP'96*, vol. 1, Atlanta, Georgia, USA, May 1996, pp. 511–514.

[10] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," in *Proc. ICASSP'95*, vol. 1, Detroit, Michigan, USA, May 1995, pp. 297–300.

[11] A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting recognition errors via discriminative utterance verification," in *Proc.ICSLP'96*, Philadelphia, USA, October 1996, pp. 602–605.

[12] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *Proc. Interspeech'07*, Antwerp, Belgium, August 2007, pp. 2385–2388.

[13] K. Thambiratmann and S. Sridharan, "Rapid yet accurate speech indexing using dynamic match lattice spotting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 346–357, January 2007.

[14] A. G. Hauptmann, R. E. Jones, K. Seymore, S. T. Slattery, M. J. Witbrock, and M. A. Siegler, "Experiments in information retrieval from spoken documents," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*, Lansdowne VA, February 1998, pp. 175–181.

[15] K. Sudoh, H. Tsukada, and H. Isozaki, "Discriminative named entity recognition of speech data using speech recognition confidence," in *Proc. ICSLP'06*, Pittsburgh, USA, September 2006, pp. 1153–1156.

[16] Z. Shafran, B. Roark, and S. Fisher, "OGI spoken term detection system," in *Proc. NIST spoken term detection workshop (STD 2006)*, Gaithersburg, Maryland, USA, December 2006.

[17] S. Goldwater, D. Jurafsky, and C. D. Maning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2009.

[18] S. Goldwater, D. Jurafsky, and C. D. Maning, "Which words are hard to recognize? lexical, prosodic, and disfluency factors that increase asr error rates," in *Proc. ACL/HLT*, June 2008, pp. 280–388.

[19] D. Wang, S. King, J. Frankel, and P. Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Proc. Interspeech'09*, Brighton, UK, September 2009, pp. 2139–2142.

[20] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, University of Amsterdam, Spuistraat 210, Amsterdam, Holland, 2007. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[21] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. M. no, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proc. Eurospeech*, September 1993, pp. 653–656.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Engineering Department, Cambridge University, March 2006.

[23] J. Tejedor, D. T. Toledano, M. Bautista, S. King, D. Wang, and J. Colás, "Augmented set of features for confidence estimation in spoken term detection," in *To appear in proc. Interspeech'10*, September 2010.

Poster Session 2

Speech production models for ASR in Spanish language

Javier Mikel Olaso, María Inés Torres

Universidad del País Vasco
javiermikel.olaso@ehu.es, manes.torres@ehu.es

Abstract

In this paper we undertake the extraction of phonological features applied to Spanish language. Also propose a method to integrate these features into an HMM based speech recognition system using an architecture that uses independent feature streams. In the experimental results we find that higher recognition accuracies and less computational cost can be obtained.
Index Terms: speech recognition, acoustic modeling, phonological features

1. Introduction

The majority of speech recognition systems are currently based on the use of the acoustic properties of speech to establish its characteristics. This method has to tackle various difficulties, such as, [2, 3, 13]:

- Phonation differences due to the diversity of speakers.
- Coarticulation effects.
- Spontaneous speech.
- Problems with pronunciation dictionaries, mainly in the English language.
- Ambient noise and interferences.

Other approaches have alternatively been proposed. One such approach seeks to incorporate information relating to the way speech is produced in terms of articulatory gestures. This approach is considered to be highly beneficial for automatic speech recognition systems, mainly due to the invariance of critical articulators, those mostly involved in sound production, and the lower susceptibility of the articulatory space to the effects of coarticulation, [1, 2]. This approach has to deal with two main problems. On the one hand, the speaker's utterances needs to be represented in terms of these articulatory gestures, and on the other hand a system is need that is able to interpret the articulatory gestures based representation. Some studies have attempted to solve these problems. The seemingly most successful method has been the use of Time Delay Neural Networks (TDNN) [5] for articulatory gestures detection, and the re-scoring of lattices obtained using a system based on HMMs defined over mel frequency cepstrum [1].

This paper is twofold. On one hand, we want to undertake the extraction of phonological features applied to the Castilian variety of Spanish. On the other hand, we propose a method to integrate these features into a speech recognition system. TDNN was used for the extraction of the features and an alternative method based on treating the vectors representing the phonological features as observation vectors of HMM models for the integration. Two types of experiments were carried out. The first only used articulatory information and the second combined both articulatory and acoustic information. We should point out that these experiments focused on the Castilian variety of Spanish and it was a challenge given that it was the first ever attempt to carry out this task.

The structure of the article is as follows. Section 2 provides a short description of the different methods studied to obtain articulatory information and describes how we decided to implement this phase. Section 3 describes the architecture of the speech recognizer used in our experiments. Section 4 contains the results of our experiments. And the paper ends with the concluding remarks and acknowledgements in Sections 5 and 6, respectively.

2. Phonological feature extraction

Several methods have been proposed for the extraction of the phonological features. These methods fall into one of two approaches. On the one hand, there are the methods based on extraction of information directly from the measurement of the positions or the articulatory organs responsible for speech generation, such as those presented in [6] where measures of the articulator's positions taken with X ray are used. On the other hand, there are the methods based on indirect measurements. Examples of the indirect methods can be found in [7], where visual information of the mouth is used, or in [8, 10, 11, 12], where the phonological information is taken from the surface waveform.

The most common of these two approaches seems to be the indirect one, and more specifically when information is taken from the surface waveform. This is mainly due to the fact that direct measurements require expensive and invasive devices, such as an electropalatograph. On the other hand, different methods are used to extract phonological information from the surface waveform, such as, the use of artificial neural networks [8, 10], dynamic Bayesian networks [4, 9] or Hidden Markov Models [14], among others.

We used neural networks in this study, and more specifically, Recurrent Time Delay Neural Networks [5], a type of neural networks that combines time-delay windows and recurrent connections to capture the dynamic information of the speech signal.

We therefore needed to define the set of sounds (phonemes) used in our experiments and how they were described in terms of articulatory features. Using the theoretical classification shown in Table 1, and after a set of tests to maximize the classification accuracy, we defined the articulatory feature sets shown in Table 2. It can be seen that it corresponds to the theoretical classification, plus a class *silence* in all features except sonority, a class *vowel* in manner and place of articulation, and a *non-vowel* class for vowel/non-vowel features.

Place of articulation	Manner of articulation							
	Plosive		Fricative	Affricate	Lateral	Trill	M. Trill	Nasal
	unvoiced	voiced	unvoiced		voiced			
Bilabial	p	b						m
Labiodental			f					
Linguodental			z					
Alveolar	t	d	s	ch	l	r	rr	n
Palatal					ll			ñ
Velar	k	g	j					

	Front	Central	Back
Close	i		u
Close-Mid	e		o
Open		a	

Table 1: Theoretical classification for phonemes in spanish language.

Sonority	
Voiced	a,e,i,o,u,b,d,g,l,ll,r,rr,m,n,ñ
Unvoiced	p,t,k,f,z,s,j,ch

Manner	
Plosive	p,t,k,b,d,g
Fricative	f,z,s,j
Affricate	ch
Lateral	l,ll
Trill	r
M. Trill	rr
Nasal	m,n,ñ
Vowel	a,e,i,o,u
Silence	*SIL*

Place	
Bilabial	p,b,m
Labiodental	f
Linguodental	z
Albeolar	t,d,s,ch,l,r,rr,n
Palatal	ll,ñ
Velar	k,g,j
Vowel	a,e,i,o,u
Silence	*SIL*

Vowel - Non Vowel			
Front	i,e	Open	a
Central	a	Mid-Close	e,o
Back	o,u	Close	i,u
Non Vowel	*rest*	Non Vowel	*rest*
Silence	*SIL*	Silence	*SIL*

Table 2: Classification used for the phonological features.

3. Speech recognizer based on phonological features

Different systems have been developed that make use of the phonological features. For example, a system is presented in [1], [10], that uses phonological features to re-score the lattices generated by a MFCC based HMM phone recognizer.

In this paper, we propose a system based on a classical acoustic speech recognition system, based on HMMs, with two main differences. On one hand, we replaced, or combined, the acoustic feature vectors with vectors representing the phonological information, that were obtained via the feature extractors mentioned in section 2. On the other hand, we followed an approach of integrating the feature vectors using independent feature streams.

Let,

$$O = o_1, o_2, ..., o_T \tag{1}$$

be a sequence of speech vectors or observations where o_t is the speech vector observed at time t. When o_t are elements of a continuous observation alphabet, and in case of using Gaussian mixtures as probability distribution function, the observation symbol probability matrix, $b_j(o_t)$, for an HMM can be written as:

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(o_t; \mu_{jm}, \Sigma_{jm}) \tag{2}$$

where $\mathcal{N}(o_t, \mu_{jm}, \Sigma_{jm})$ denotes m'th Gaussian, with μ_{jm} mean vector and Σ_{jm} variance matrix, for state j. M is the number of Gaussians in the mixture and c_{jm} is the weight of the m'th component in the mixture, satisfying:

$$\sum_{m=1}^{M} c_{jm} = 1 \tag{3}$$

Well, to integrate the features defined in Table 2 we propose to use an architecture with independent feature streams. Let S be the number of independent feature streams, e.g. those defined in Table 2, and O_{st} a vector defined as:

$$O_{st} = o^1_{st}, o^2_{st}, \ldots, o^n_{st} \tag{4}$$

that represents an observation in stream s and time t, and with n its dimension, which may vary for each feature stream. With this approach the observation symbol probability matrix, $b_j(o_t)$, can be rewritten as:

$$b_j(o_t) = \prod_{s=1}^{S}(\sum_{m=1}^{M_s} c_{jms}\mathcal{N}(O_{st}; \mu_{jms}, \Sigma_{jms})) \tag{5}$$

where M_s is the number of Gaussians in the mixture of stream s.

Likewise, in the case of using discrete symbol streams, the matrix, $b_j(o_t)$, can be written as:

$$b_j(o_t) = \prod_{s=1}^{S} b_{js}(O_{st}) \tag{6}$$

where $b_{js}(O_{st})$ is the observation symbol probability matrix of stream s.

4. Experimental results

This section is dedicated to a more detailed description of the implementation of the system presented. First, we provide a short description of the corpus used. The process for the phonological feature extraction is then described, and finally the different configurations, and the recognition results of the speech recognition system used are given.

4.1. Database description

The speech corpus used in this paper was Albayzin [15]. This is a corpus in the Castilian variety of Spanish recorded at 16KHz divided in three sub-corpus: a phonetic corpus without syntactic-semantic restrictions, which was used in this study, a second corpus including those restrictions and a third corpus designed for noisy environments. The phonetic corpus is divided in a training set of 200 sentences pronounced by 4 speakers and 25 sentences more pronounced by 160 speakers, making a total of 4800 sentences, 42144 words (712 different) and 187848 phonemes, along with a test set with 50 sentences pronounced by 40 speakers, making a total of 2000 sentences, 21052 words (1856 different) and 93696 phonemes. Table 3 contains a short description of the phonetic corpus.

	Training	Test
Speakers	160	40
Sentences	4800	2000
Words	42144	21052
Different Words	712	1856
Phonemes	187848	93696

Table 3: Summary of the phonetic subcorpus of Albayzin speech corpus.

On the other hand, the representation of the corpus in terms of the phonological features needed to be obtained prior to training the HMM models. This representation was obtained by making previously trained networks, see section 4.2, act on the acoustic representation of the corpus.

Finally, the corpus was transcribed using a set of 24 phonetic units, 23 phonemes and 1 silence, and therefore 24 HMM models were trained.

4.2. Phonological feature extraction results

Based on the study in [5], we used Recurrent Time Delay Neural Networks for phonological feature detection. Five neural networks were used to detect each of the following features:

- Sonority
- Vowel-NonVowel (2)
- Articulation manner
- Place of articulation

These neural networks had multiple outputs and the classes to be detected for each feature were those described in Section 2. The inputs of all the neural networks were 12 first Mel Frequency Cepstral Coefficients plus energy, which were extracted in 25 ms Hamming windowed frames with an overlapping of 10 ms. The outputs of the neural networks were real values ranging from 0 to 1. Although these values could be treated as the posterior probabilities of the features, we applied a more basic implementation and used them as simple vectors, as if they were MFCCs.

Table 4 contains the detection accuracies of the different phonological feature sets. It can be seen very high detection accuracy was obtained in the case of both sonority and vowels detectors, given the very good vowel-nonvowel detector obtained. In the case of manner and place of articulation, good overall detection accuracies were obtained, but this poor detection in some classes. For example in the case of the articulation manner, poorer detection accuracies were obtained for the classes *trill*(r) and *multiple trill*(rr), which were mostly detected as vowels. This is due to the fact that these phonemes do not have their own spectrum in Spanish and they inherit the spectrum of the preceeding vowel and give it an intermittent pattern.

4.3. Recognition results

The HMM topology used was the classical left-to-right of three states with transitions from one state to itself and to the adjacent one. Two types of experiments were likewise carried out. On

Sonority	
Class	% correct
Voiced	93.3
Unvoiced	96.0

Manner	
Class	% correct
Total	83.8

Place	
Class	% correct
Total	83.4

Vowel - Non Vowel			
Class	% correct	Class	% correct
Front	81.1	Open	78.3
Central	84.1	Mid-Close	79.0
Back	73.8	Close	75.1
Non Vowel	90.8	Non Vowel	90.3
Silence	96.0	Silence	93.6

Table 4: Classification accuracies for the different phonological features.

the one hand, the system was tested using phonological information only, and on the other hand, phonological and acoustic information was combined.

When using phonological information only, two ways of integrating the information were used. The first used independent feature streams. The second used a unique feature stream resulting from the concatenation of the vectors of each of the independent streams. When acoustic information was also integrated, it took place as four independent feature streams. These streams corresponded to the first twelve cepstral coefficients, their first and second derivatives, and an additional stream with the energy and first derivative of the energy, per frame.

We also used discrete models. In this case, codebooks needed to be generated both for the case of a unique feature stream and of various feature streams. In the case of a unique stream, it was generated using the LBG algorithm to the concatenation of the independent feature vectors. For the various streams case, the codebooks were generated as follows: for each independent feature, the representative vector for each class was obtained as the mean vector of all the vectors belonging to that class. And were these representative vectors what we used as the codebook's vectors.

We then proceeded to train and test the models. It should be noted that tests varying the number of Gaussians in the mixtures for continuous models and the number of vectors of the codebooks for discrete models were carried out. Table 5 contains the results obtained, together with the recognition results for the acoustic based system used as baseline. The topology of this baseline system was identical to the topology of the system presented, with the same four independent acoustic feature streams used in the combinations with the phonological features. On the other hand, a codebook of 1024 classes in the case of discrete models and 32 mixture Gaussians in the case of continuous models were used in the results given for this baseline system .

It can be seen that a recognition improvement was obtained in the case of discrete HMM models, however in the case of continuous HMM models only when combining phonological and acoustic information we obtain recognition accuracies similar to the baseline system. We believe that this could be due to the fact that the phonological space is highly discretized which favours the use of discrete models. On the other hand, and comparing the systems with just phonological information and with both phonological and acoustic information, it can be seen that the systems combining both types of information have better recognition accuracies.

	DHMM		CHMM	
BASELINE	69.40		75.15	
	$S = 1$	$S = 5$	$S = 1$	$S = 5$
PH.	72.93	72.46	70.35	70.23
	$S_{ph} = 1$	$S_{ph} = 5$	$S_{ph} = 1$	$S_{ph} = 5$
PH.+AC.	75.83	75.72	75.06	74.24

Table 5: Recognition results for DHMM and CHMM. When combining phonological (PH) and acoustic (AC) spaces, we have $S = S_{ph} + S_{ac}$ and $S_{ac} = 4$, being S_{ph} and S_{ac} the number of independent feature streams for phonological and acoustic spaces respectively.

We find that the results obtained for the discrete models are pretty good because they have proved to be computationally faster than continuous ones. In Table 6 we show computation times for the recognition process of the continuous HMM models based baseline system and the different implementations used with discrete HMM models, normalized with the value of the baseline system. It can be seen that using both phonological and acoustic features has higher computational cost than using only phonological features, although this cost is less than the cost of the baseline system and therefore is a reasonable cost because the gain in recognition accuracy is higher. Alternately, also can be seen that when speaking of computational cost is better to use phonological features in independent streams rather than concatenate them in one stream.

	PH.		PH. + AC.	
	$S = 1$	$S = 5$	$S_{ph} = 1$	$S_{ph} = 5$
DHMM	0.13	0.03	0.27	0.17
BASELINE	1			

Table 6: Normalized computation times for baseline and discrete HMM models. When combining phonological (PH) and acoustic (AC) spaces, we have $S = S_{ph} + S_{ac}$ and $S_{ac} = 4$, being S_{ph} and S_{ac} the number of independent feature streams for phonological and acoustic spaces respectively.

5. Concluding remarks

In this work we have undertaken the problem of using phonological features for speech recognition in Castilian variety of Spanish. Also we have proposed a method for integrate these features in a speech recognition system based on HMM models.

We have found, that the use of phonological features could be highly beneficial above all in the case of using discrete HMM models where we have obtained better results than the baseline system used, both in accuracie rate and in computational cost.

6. Acknowledgements

This work has been partially supported by the University of the Basque Country under grant GIU07/57, Spanish CICYT under grant TIN2008-06856-C05-01 and by the Spanish program Consolider-Ingenio 2019 under grant CSD2007-00018.

7. References

[1] Rose, R. and Momayyez, P., “Integration of multiple feature sets for reducing ambiguity in ASR.“, ICASSP 2007, Volume 4, 325-328, 2007.

[2] Rose, R. et ál, “An investigation of the potential role of speech production models in automatic speech recognition”, Proceedings ICSLP-94, pp. 575-578, 1994.

[3] Koreman, J. and Andreeva, B., “Can we use the linguistic information in the signal¿‘, Phonus (Institute of Phonetics, University of the Saarland) 5: 47-58, 2000.

[4] Livescu, K. et ál “Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report“, Technical Report, Center for Language and Speech Processing, Johns Hopkins University, 2007.

[5] Strom, N., “Phoneme probability estimation with dynamic sparsely connected artificial neural networks.“, The free speech journal, Vol 1, Issue #5, 1997.

[6] Blackburn, C.S. and Young, S.J., “Pseudo-Articulatory speech synthesis for recognition using automatic feature extraction from X-Ray data.“, In proceedings ICSLP 96, 969-972, 1996.

[7] Saenko, K. et ál, “Articulatory features for robust visual speech recognition.“, ICMI’04, 2004.

[8] King, S. and Taylor, P., “Detection of phonological features in continuous speech using neural networks.“, Computer Speech & Language, 333-353, 2000.

[9] Frankel, J. et ál, “Articulatory feature recognition using dynamic Bayesian networks“, Computer Speech and Language archive, Volume 21 , Issue 4, 620-640, 2007.

[10] Parya, M. at ál, “Exploiting complementary aspects of phonological features in automatic speech recognition.“, IEEE Workshop on Automatic Speech Recognition & Understanding”, 47-52, 2007.

[11] Stouten, F. and Martens, J.P., “On the use of phonological features for pronunciation scoring.“, In proceedings ICASSP, 229-232, 2006.

[12] Stouten, F. and Martens, J.P., “Speech Recognition with Phonological Features: Some issues to attend.“, Interspeech-06, paper 1081-Mon2BuP-4, 2006.

[13] BenZeghiba, M. et ál, “Automatic speech recognition and intrinsic speech variation“, ICASSP 2006, 31st International Conference on Acoustics, Speech, and Signal Processing, May 14-19, 2006.

[14] Abu-Amer, T. and Carson-Berndsen, J., “HARTFEX: A multidimentional system of HMM based recognisers for articulatory features extraction.“, In proceedings NOLISP-2003, paper009, 2003.

[15] Casacuberta, F. et ál, “Desarrollo de corpus para investigación en tecnologías del habla (Albayzin).”, Procesamiento del lenguage natural, 12:35-42, 1992.

FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

A feature compensation approach using VQ-based MMSE estimation for robust speech recognition

José A. González, Antonio M. Peinado, Angel M. Gómez, José L. Carmona, and Juan A. Morales-Cordovilla

Dpto. de Teoría de la Señal, Telemática y Comunicaciones, University of Granada, Spain
{joseangl,amp,amgg,maqueda,jamc}@ugr.es

Abstract

We describe a novel feature compensation algorithm based on the minimum mean square error (MMSE) estimation and stereo training data for robust speech recognition. The proposed algorithm can be viewed as a piece-wise linear transformation between the noisy and clean feature spaces, where both spaces are modeled by means of vector quantization (VQ) codebooks. By means of this VQ modeling, we show that a very efficient estimator can be obtained in terms of computational cost and recognition accuracy. Also, two approaches are proposed in order to compensate the acoustic noise distortion. First, we propose a novel formulation for the normalization of noisy feature vectors. Second, a novel subregion-based modeling is applied to obtain a better representation of the differences between noisy and clean domains. The experimental results on noisy digit recognition show a relative improvement of 61.49% over the baseline when clean acoustic models are used. Furthermore, important improvements are achieved in comparison with other similar approaches.

Index Terms: robust speech recognition, feature compensation, MMSE estimation, stereo data

1. Introduction

It is well known that the performance of automatic speech recognition (ASR) systems degrades as the mismatch between testing and training conditions increases. Thus, there are several sources of mismatch that directly affect to the ASR performance, such as variety of speakers, accents, channels and noise conditions [1]. Many algorithms have been developed to compensate this mismatch. These algorithms are usually grouped into two categories [3]: feature-based and model-based approaches. Feature-based techniques focus on modifying or enhancing the feature vectors to be closer to the clean training condition or to be less sensitive to the variability introduced by the aforementioned sources of mismatch. On the other hand, model-based approaches adapt the acoustic model parameters to the testing conditions. These approaches often yield better performance than feature-based ones, especially in low SNR conditions. Nevertheless, feature-based techniques have the advantage that can be seamlessly implemented into existing systems, since only a module that pre-process the feature vectors before they are fed into the speech recognizer is needed. In addition, feature compensation is usually less computationally expensive, especially if the acoustic environment is rapidly changing.

Stereo data are widely used in order to achieve noise robustness in ASR systems. In this way, a stereo database including both clean and noisy features can be used to learn the statistical relationship between both domains. The earliest approach based on stereo data was proposed in [3] with the SNR-Dependent Cepstral Normalization (SDCN) and Codeword-Dependent Cepstral Normalization (CDCN). Since then, more sophisticated techniques have appeared, as multivariate Gaussian based cepstral normalization algorithm (RATZ) [4], Stereo based Piecewise LInear Compensation for Environments (SPLICE) [5], Multi-Environment Models based LInear Normalization (MEMLIN) [6] and Stereo-based Stochastic Mapping (SSM) [7]. The later techniques are based on a Minimum Mean Squared Error (MMSE) estimation, where the clean and/or noisy domains are represented by means of Gaussian Mixture Models (GMMs).

In this paper we are also interested in MMSE estimation for feature compensation, although a different approach is followed to represent the clean and noisy domains. Thus, instead of modeling the clean and noisy feature spaces with GMMs, we characterize each of these spaces with a set of cells obtained by means of vector quantization (VQ). As it will be shown, VQ quantization provides much more efficient compensation techniques, but their results are known to be inferior, due to the hard decision involved (a cell is represented by a centroid instead of a probability function). For this reason, in this paper we present a novel MMSE formulation which can cope with this disadvantage. In addition, we show that the recognition accuracy can be significantly improved by considering that every VQ cell contains a set of overlapping subregions with provide a more accurate mapping between the clean and noisy spaces.

This paper is organized as follows. In Section 2, the mathematical formulation for the proposed VQ-based MMSE estimation is derived. The experimental framework is described in Section 3 while the results are presented and discussed in Section 4. Finally, Section 5 presents the conclusions and some directions for future work.

2. Derivation of the proposed MMSE estimator

We denote by $\boldsymbol{y}$ the observed feature vector representation of a noisy speech segment distorted by acoustic noise and by $\boldsymbol{x}$ its corresponding unknown clean version obtained when the segment is not affected by noise. In this work, we seek for a compensation function that provides an estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$, i.e., $\hat{\boldsymbol{x}} = f(\boldsymbol{y})$. Among others, a plausible option to derive this function is by means of the MMSE criterion. In this case, the estimate of clean speech is given by

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\boldsymbol{y}] = \int_{\boldsymbol{x}} \boldsymbol{x} \cdot p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \tag{1}$$

where $p(\boldsymbol{x}|\boldsymbol{y})$ is the conditional probability of $\boldsymbol{x}$ given $\boldsymbol{y}$. Different approaches have been proposed to model this distribution. For example, RATZ [4] models the clean feature space by means of a GMM and it assumes an additive effect of the noise on the MFCC domain. On the other hand, SPLICE [5] models the distorted feature space and, as RATZ, also an additive effect of the noise is assumed. A more complex modeling is applied by SSM [7] in which the conditional distribution is derived from the joint distribution of clean and noisy feature vectors $p(\boldsymbol{x}, \boldsymbol{y})$. In this work, however, we follow a different approach. We assume that the clean and noisy feature spaces can be independently represented by means of probability density function (pdf) mixtures in the following way,

$$p(\boldsymbol{x}) = \sum_{k_x} p(\boldsymbol{x}|k_x)P(k_x) \tag{2}$$

$$p(\boldsymbol{y}) = \sum_{k_y} p(\boldsymbol{y}|k_y)P(k_y) \tag{3}$$

where k_x and k_y are components (e.g., Gaussian pdfs) of the mixtures that model the clean and noisy spaces, respectively.

Using the previous models, the conditional probability $p(\boldsymbol{x}|\boldsymbol{y})$ can be expressed as,

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &= \sum_{k_x}\sum_{k_y} p\left(\boldsymbol{x}, k_x, k_y|\boldsymbol{y}\right) \\ &= \sum_{k_x}\sum_{k_y} p\left(\boldsymbol{x}|k_x, k_y, \boldsymbol{y}\right) p\left(k_x|k_y, \boldsymbol{y}\right) p\left(k_y|\boldsymbol{y}\right) \end{aligned} \tag{4}$$

Finally, applying (4) to (1), the MMSE estimation takes the following form,

$$\hat{\boldsymbol{x}} = \sum_{k_x}\sum_{k_y} E\left[\boldsymbol{x}|k_x, k_y, \boldsymbol{y}\right] P\left(k_x|k_y, \boldsymbol{y}\right) P\left(k_y|\boldsymbol{y}\right) \tag{5}$$

where $P\left(k_y|\boldsymbol{y}\right)$ and $P\left(k_x|k_y, \boldsymbol{y}\right)$ are obtained using the marginal distributions of eqns. (2)-(3) and stereo training data. In contrast to other methods, such as MEMLIN [6], where these two distributions are modeled by means of GMMs, we propose the use of vector quantization (VQ) codebooks. In this way, every feature vector space is modeled by means of a VQ codebook that partitions its space into a set of disjoint cells. We will notate $\{C_X^{(i)}\ (i = 1, \ldots, M)\}$ as the set of cells corresponding to the clean feature space X and $\{C_Y^{(j)}\ (j = 1, \ldots, N)\}$ as the cells of the noisy space Y. These cells will hereinafter play the role of pdfs k_x and k_y in eqn. (5).

The VQ codebook of the noisy feature space can be used now to compute the *a posteriori* probability $P(k_y|\boldsymbol{y})$ in eqn. (5) as,

$$P\left(C_Y^{(j)}\middle|\,\boldsymbol{y}\right) = \begin{cases} 1 & C_Y^{(j)} = C_Y^* \\ 0 & otherwise \end{cases} \tag{6}$$

where $C_Y^* \equiv C_Y^*(\boldsymbol{y})$ is the cell that contains the input feature vector $\boldsymbol{y}$ according to the following distance,

$$C_Y^*(\boldsymbol{y}) = \underset{j}{\operatorname{argmin}} \left\{ (\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y})^T \operatorname{diag}(\boldsymbol{\Sigma}_Y^{(j)})^{-1} (\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y}) \right\} \tag{7}$$

where $\operatorname{diag}(\cdot)$ returns a diagonal matrix with the elements of the main diagonal of the input matrix, and $\boldsymbol{\mu}_Y^{(j)}$ and $\boldsymbol{\Sigma}_Y^{(j)}$ are the mean vector (centroid) and covariance matrix of $C_Y^{(j)}$.

Applying (6) to (5), the MMSE estimation can be rewritten as,

$$\begin{aligned} \hat{\boldsymbol{x}} &= \sum_{i=1}^{M}\sum_{j=1}^{N} E[\boldsymbol{x}|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}] P(C_X^{(i)}|C_Y^{(j)}, \boldsymbol{y}) P(C_Y^{(j)}|\boldsymbol{y}) \\ &\approx \sum_{i=1}^{M} E\left[\boldsymbol{x}\,\middle|\, C_X^{(i)}, C_Y^*, \boldsymbol{y}\right] P\left(C_X^{(i)}\,\middle|\, C_Y^*\right) \end{aligned} \tag{8}$$

We will refer to this estimation as VQ-based MMSE estimation (VQ-MMSE). As can be observed, the conditional probability $P(k_x|k_y, \boldsymbol{y})$ of eqn. (5) is simplified to $P(\,C_X^{(i)}\,\big|\,C_Y^*)$ in VQ-MMSE. This probability can be estimated using stereo data. This simplification, along with the application of VQ for the computation of the noisy component posterior in eqn. (6), leads to a very efficient implementation of the MMSE estimation. It is important to note that the original input vector $\boldsymbol{y}$ remains in the expected value $E[\boldsymbol{x}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$ of eqn. (8) in spite of the VQ modeling. That is, the VQ modeling is applied to compute the probabilities required by the MMSE estimation, but it does not necessarily involve a quantization of the input that could lead to a performance reduction.

The term $E[\boldsymbol{x}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$ in eqn. (8) defines the transformation of feature vectors between cells $C_X^{(i)}$ and C_Y^* due to acoustic noise. In order to accurately model this transformation, we introduce in the following the concept of subregion of a VQ cell. We will consider that every clean cell $C_X^{(i)}$ is composed by a set of subregions $\{C_X^{(i,j)}, (j = 1, \ldots, N)\}$, where $C_X^{(i,j)}$ represents all the clean feature vectors whose corresponding distorted ones belong to the noisy cell $C_Y^{(j)}$. Similarly, $C_Y^{(i,j)}$ represents the subregion of $C_Y^{(j)}$ where the feature vectors of $C_X^{(i)}$, once transformed by noise, are mapped. It is interesting to point out that this refined modeling can also be seen as a cross-modeling between the clean and noisy domains. Thus, a subregion defined in a given feature space can be considered as a part of the projection of a cell defined in the other space.

In order to compensate the noise distortion, we propose to apply a linear transformation to every feature vector. To do so, we assume that the subregions in the clean and noisy feature spaces are Gaussian distributed, i.e., $C_X^{(i,j)} \sim \mathcal{N}(\boldsymbol{\mu}_X^{(i,j)}, \boldsymbol{\Sigma}_X^{(i,j)})$ and $C_Y^{(i,j)} \sim \mathcal{N}(\boldsymbol{\mu}_Y^{(i,j)}, \boldsymbol{\Sigma}_Y^{(i,j)})$, where $\boldsymbol{\mu}_X^{(i,j)}$,$\boldsymbol{\mu}_Y^{(i,j)}$ are the mean vectors and $\boldsymbol{\Sigma}_X^{(i,j)}$,$\boldsymbol{\Sigma}_Y^{(i,j)}$ the corresponding covariance matrices. Then, the proposed transformation takes the following form,

$$E\left[\boldsymbol{x}\,\middle|\, C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right] = \boldsymbol{A}^{(i,j)}\boldsymbol{y} + \boldsymbol{b}^{(i,j)} \tag{9}$$

where $\boldsymbol{A}^{(i,j)}$ and $\boldsymbol{b}^{(i,j)}$ are computed in order to eqn. (9) firstly normalizes the noisy feature vectors regarding the mean and covariance of the noisy subregion, and then transforms them to the clean domain. Thus, eqn. (9) can be seen as a whitening and mapping transformation whose parameters are computed as,

$$\boldsymbol{A}^{(i,j)} = \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2} \left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \tag{10}$$

$$\boldsymbol{b}^{(i,j)} = \boldsymbol{\mu}_X^{(i,j)} - \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2} \left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \boldsymbol{\mu}_Y^{(i,j)} \tag{11}$$

where these terms can be precomputed offline for every pair of cells $(C_X^{(i)}, C_Y^{(j)})$. In addition, the mean vectors and covariance matrices can be easily computed from a stereo database using the feature vectors assigned to each subregion.

Finally, the proposed VQ-MMSE estimation in (8) becomes

$$\begin{aligned}
\hat{\boldsymbol{x}} &= \sum_{i=1}^{M} E\left[\boldsymbol{x}\left|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right.\right] P\left(C_X^{(i)}\middle|C_Y^*\right) \\
&= \sum_{i=1}^{M}\left(\boldsymbol{A}^{(i,*)}\boldsymbol{y}+\boldsymbol{b}^{(i,*)}\right) P\left(C_X^{(i)}\middle|C_Y^*\right) \\
&= \underbrace{\left(\sum_{i=1}^{M} P\left(C_X^{(i)}\middle|C_Y^*\right)\boldsymbol{A}^{(i,*)}\right)}_{\boldsymbol{A}^*}\boldsymbol{y} + \underbrace{\sum_{i=1}^{M} P\left(C_X^{(i)}\middle|C_Y^*\right)\boldsymbol{b}^{(i,*)}}_{\boldsymbol{b}^*} \\
&= \boldsymbol{A}^*\boldsymbol{y}+\boldsymbol{b}^* \qquad (12)
\end{aligned}$$

where $\boldsymbol{A}^*$ and $\boldsymbol{b}^*$ can be precomputed offline. Thus, we can see that the proposed compensation depends only on the noisy cell C_Y^* which the input feature vector $\boldsymbol{y}$ belongs to.

3. Experimental framework

Experiments are performed under the framework proposed by ETSI STQ-Aurora working group using the Aurora-2 database [8]. This database consists of utterances of connected digits spoken by American English speakers. For our purposes, we have extracted the speech data from the clean training set and the clean utterances from the test *set A* of this database. The European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) [9] is used in this work. It provides a 13-dimension feature vector containing 12 Mel-Frequency Cepstral Coefficients (MFCCs) (the 0th order one is discarded), plus the log-energy feature. The recognizer is the one provided by Aurora-2 using whole word acoustic models trained on clean speech. Each digit is modeled by means of a 16-state continuous HMM with 3 Gaussians per state. On the other hand, the silence and short pause models are modeled by means of HMMs with 3 and 1 states, respectively, and 6 Gaussians per state.

The speech features extracted by ETSI FE are directly processed by VQ-MMSE. After the compensation, the dynamic speech features are computed. VQ codebooks are trained for every available acoustic condition using a k-means algorithm which applies the weighted Euclidean distance defined in eqn. (7). Through this set of codebooks, the compensation parameters are estimated for the proposed technique using stereo data. These compensation parameters account for the possible transformations due to acoustic noise between the clean feature space and the noisy one, both modeled by means of VQ codebooks with the same number of cells. In order to compare our proposal with other MMSE-based estimators, GMMs are also trained. Thus, one GMM with diagonal covariance matrices is estimated for every available training condition using the Expectation-Maximization (EM) algorithm.

A set of 9 acoustic noises is used for training purposes, namely: airport, highway, babble, bar, beach, pedestrian street, restaurant, street, and train station. Every noise recording is split into two parts: two-thirds are employed to train the proposed MMSE estimator while the remaining third is reserved for testing. The training part is added to the *clean* training set of Aurora-2 at 6 different SNRs (20, 15, 10, 5, 0, and -5 dB), resulting in 54 environmental noisy training conditions plus a clean condition (55 training conditions in total). In order to evaluate the performance of our proposal, two different test sets are defined. The first set, called *Set A*, is intended to show the performance of the different techniques when considering the same environments used for training. Thus, 55 testing conditions are defined by artificially contaminating the clean test set of Aurora-2 with the testing part of the noises. The second set, called *Set B*, is created in the same way, but using five new different noises (pedestrian square, car, bus station, heavy sea, and heavy traffic avenue) at 5 new different SNRs (17.5, 12.5, 7.5, 2.5, and -2.5 dB). Thus, we can evaluate the influence of considering different environments that the ones used for training.

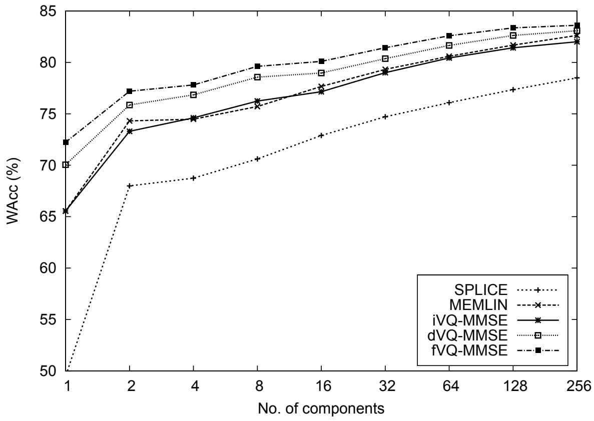


Figure 1: Oracle results for different feature compensation algorithms based on MMSE regarding the number of components (Gaussians or VQ cells) used.

4. Results

In the first part of this section, we give results for the aforementioned digit recognition task in non-mismatch conditions. Later, we provide results when the proposed technique is tested for unknown acoustic noises.

4.1. Oracle experiments

Fig. 1 shows the average word accuracy (WAcc), in percent, achieved by different estimators for Set A. For these experiments, oracle information about the acoustic noise is assumed, i.e., each utterance is compensated using a set of compensation parameters trained under the same noise. It must be pointed out that this information is not available in practice. However, the oracle results provide an estimate of the best performance that could be expected from every technique.

The baseline system applies acoustic models trained with clean speech and no compensation. This configuration achieves a WAcc of 50.83%. Three different versions of the proposed VQ-MMSE estimation are evaluated: iVQ-MMSE, dVQ-MMSE, and fVQ-MMSE. These versions assume identity, diagonal, and full covariance matrices, respectively, for the computation of the expected value in eqn. (9). Also, two well-known MMSE estimators using GMMs are considered: SPLICE [5] and MEMLIN [6]. As can be seen, our 3 proposals greatly improve the results achieved by the baseline system and SPLICE for all GMM and VQ sizes. This improvement shows the benefits of modeling both feature spaces, clean and noisy, instead of only one space such as SPLICE. Thus, the transformation applied by our approach is more accurate than in SPLICE. MEMLIN achieves a performance slightly better than iVQ-MMSE (82.62% vs. 82.02% for 256-component codebooks) . In fact, both techniques are quite similar, although our approach is more computationally efficient. Further improvements can be ob-

	Set A	Set B	Avg.	Imp.
Baseline	50.83	40.28	45.56	–
SPLICE	72.99	59.45	66.22	45.35
MEMLIN	77.21	62.89	70.05	53.75
iVQ-MMSE	77.29	65.44	71.37	56.64
dVQ-MMSE	79.04	67.18	73.11	60.47
fVQ-MMSE	79.54	67.61	**73.57**	61.49

Table 1: Average Word Accuracy (%) achieved by SPLICE, MEMLIN, and VQ-MMSE in the soft-compensation experiments for Set A and Set B (256-components codebooks).

tained when a more complex mapping is applied. This is the case of dVQ-MMSE and fVQ-MMSE, which compensate the shifts and scales in the feature domain due to environmental noise. In this case, our approaches obtain better results than MEMLIN.

4.2. Soft-compensation experiments

The proposed techniques are also evaluated in a more realistic scenario in which the acoustic noise that distorts the speech is unknown. In such a scenario, the clean feature vector estimate is obtained by means of a soft-compensation approach [10]. Thus, an estimate $\hat{\boldsymbol{x}}_e$ is obtained for every possible environmental condition e. The final estimate is computed as a linear combination of the estimates obtained for all environments. To do so, GMMs trained on every environmental condition are employed as environment classifiers to obtain the required probabilities $P(e|\boldsymbol{y})$. It must be pointed out that these GMMs are the same as those employed by SPLICE and MEMLIN, although a more sophisticated environment modeling could be applied.

Table 1 shows the recognition results achieved in the soft-compensation experiments for Set A and Set B when codebooks (GMMs or VQ codebooks) with 256 components are used. The average word accuracy (Avg.) and the relative improvement over the baseline in percent (Imp.) are also shown. As can be seen, all techniques suffer a performance degradation regarding the oracle experiments for Set A. This degradation is produced by mismatches in the environment identification. Furthermore, all methods yield poorer results for Set B. This is one of the lacks of the soft-compensation approach: the performance drops in mismatch situations. Nevertheless, these results demonstrate again the superior performance of our proposal. Thus, fVQ-MMSE achieves relative improvements of 11.10% and 5.03% in comparison with SPLICE and MEMLIN, respectively. Furthermore, now iVQ-MMSE outperforms MEMLIN.

5. Conclusions

In this paper, we have presented a novel feature compensation technique based on MMSE estimation and stereo training data for robust speech recognition. As a result, a piece-wise linear function between the noisy feature space and the clean one is obtained. We show that the application of VQ codebooks for the modeling of the feature spaces allows an efficient implementation of the MMSE estimator. Also, a novel subregion modeling is applied in order to accurately represent the acoustic noise distortion.

Two sets of experiments are carried out. Firstly, oracle experiments are conducted to obtain an upper bound of the performance that could be expected under non-mismatch. Secondly, the proposed techniques are also tested with unknown noises. In these experiments, we follow a soft-compensation approach in which the clean feature vector estimate is obtained as a linear combination of the estimates obtained for several defined environments. A relative improvement of 61.49% regarding the baseline is achieved for these experiments. Furthermore, relative improvements of 11.10% and 5.03% are obtained in comparison with two other well-known MMSE-based compensation algorithms: SPLICE and MEMLIN.

The experimental results show the importance of modeling both feature spaces (clean and noisy) in order to obtain an accurate probability model for the MMSE estimation. Furthermore, the proposed normalization of noisy feature vectors and the more accurate representation of the noise distortion by means of the proposed subregion modeling, lead to further improvements. Finally, we think that the application of the proposed feature compensation algorithm to scenarios where stereo data is unavailable is an interesting issue that deserves more research in the future.

6. Acknowledgments

This work has been supported by an FPU grant from the Spanish Ministry of Science and Innovation and by project MEC-FEDER TEC2007-66600.

7. References

[1] A.M. Peinado and J.C. Segura, “Speech Recognition over digital channels. Robustness and Standards”, *Wiley*, 2006.

[2] X. Huang, A. Acero, and H. Hon, “Spoken language processing: A guide to theory, algorithm, and system development”, *Prentice Hall*, 2001.

[3] A. Acero, “Acoustical and environmental robustness in automatic speech recognition”, *Kluwer Academic Publishers*, Norwell, MA, U.S.A, 1993.

[4] P.J. Moreno, “Speech recognition in noisy environments”, Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1996.

[5] J. Droppo, L. Deng, and A. Acero, “Evaluation of the SPLICE algorithm on the Aurora2 database”, in *Proc. Eurospeech 2001*, pp. 217–220.

[6] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, “Cepstral vector normalization based on stereo data for robust speech recognition”, *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 3, pp. 1098–1113, Mar. 2007.

[7] M. Afify, X. Cui, and Y. Gao, “Stereo-based stochastic mapping for robust speech recognition”, *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 7, pp. 1325–1334, Sep. 2009.

[8] H.G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluations of the speech recognition systems under noisy conditions”, in *ISCA ITRW ASR2000*, 2000.

[9] “ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms”, ETSI.

[10] José A. González, A.M. Peinado, A.M. Gómez, José L. Carmona, and Juan A. Morales-Cordovilla, “Efficient VQ-based MMSE estimation for robust speech recognition”, in *Proc. ICASSP 2010*, pp. 4558–4561.

Missing Feature Techniques Combination for Speaker Recognition in Noisy Environment

Dayana Ribas[1], *Jesús A. Villalba*[2], *Eduardo Lleida*[2] *and José R. Calvo*[1]

[1] Advanced Technologies Application Center (CENATAV), La Habana, Cuba.
[2] Communications Technology Group (GTC), Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain.
{dribas, jcalvo}@cenatav.co.cu,{villalba, lleida}@cenatav.co.cu

Abstract

In order to handle speech signals corrupted by noise in speaker recognition and provide robustness to systems, this paper evaluates the use of missing feature (MF) approach with a novel combination of techniques. A mask estimation based on spectral subtraction is used to determine the reliability of spectral components in a speech signal corrupted by noise. A cluster based reconstruction technique is used to remake the damaged spectrum. The recognition performance was evaluated through a speaker verification experiment with signals corrupted by white noise under different signal to noise ratios. The results were promising since they reflected a relevant increase of speaker verification performance, applying MF approach with this combination of techniques.

Index Terms: speaker recognition, missing feature

1. Introduction

Dealing with noisy signals is a fact in real life, background noise can markedly degrade performance of any speaker recognition system. In order to handle environmental noise to improve the robustness of recognition performance, many techniques have been proposed [1]. Most of them were originally designed and applied in speech recognition application. MF method [2] is an example of that.

MF approach is a group of techniques developed to compensate for noise. Unlike other compensation methods MF does not require to know a priori the characteristics of noise to handle unknown noise. Because of that, it has a lot of potential to ensure robustness in speaker recognition applications which process speech signals acquired in noisy environmental conditions with unknown features. This situation is very frequent in real applications.

The MF approach has two steps. The first determines the level of noise corruption in each time-frequency region of speech spectrum to set up a map of binary labels called spectrographic mask. The mask tags as unreliable *(U)* the time-frequency spectral components that are so corrupted by noise that can cause poor recognition performance, and tags as reliable *(R)* the time-frequency spectral components that are not very corrupted by noise. The second step is compensation of unreliable region, it could be bypassing the spectral unreliable locations in the recognition process, known as marginalization, or reconstructing unreliable spectrum location and keeping the recognition process with the new reconstructed spectrum.

Until now, most of the MF development has occurred on the speech recognition field, while only a few works have been done on speaker recognition [3][4][5][6]. This work presents a novel combination of MF techniques for robust speaker recognition with noisy speech. For estimate the MF mask we proposed to use the SNR criterion. For MF compensation we proposed to use a reconstruction method which estimate *U* components from *R* ones. This kind of reconstruction have not been previously used for speaker recognition. We evaluate the performance impact of this MF setup through speaker verification experiment in noisy environments.

From now on, this paper is organized as follows. Section 2 describes mask estimation technique. Section 3 explains the MF compensation technique used. Section 4 presents speaker verification experiments and results. Finally, section 5 a discussion of results and conclusions.

2. Mask estimation

The success of the MF approach in providing robustness to speaker recognition system will depend on the mask accuracy [2]. To estimate the masks, the SNR criterion is the most widely used in previous works because of SNR-based masks are very easy to compute [7].

In this paper we proposed, as MF detector, the identification of *U* spectral components based on spectral enhancement technique used frequently in speech processing. This approach was applied to MF mask estimation in the previous work [8]. This is an effective technique in the detection of corrupted components that is known as Negative Energy Criterion.

This method uses a frame by frame spectral subtraction algorithm as MF detector and is based on an estimated noise spectrum. The reliability decision of spectral components is done following this rule:

$$\begin{array}{lll} |Y(f,s)|^2 \leq |\hat{N}(f,s)|^2 & then & Y(f,s) \leftarrow U \\ |Y(f,s)|^2 > |\hat{N}(f,s)|^2 & then & Y(f,s) \leftarrow R \end{array} \quad (1)$$

where, f and s are the frame (time) and subband (frequency) spectrographic representation of the signal power spectrum, respectively. If the power spectrum in a component is less than the estimated noise power spectrum in it, this component is assumed as *U*, otherwise the component is tagged as *R*.

3. Cluster-based reconstruction

Until now, most speaker verification systems using the MF approach, to improve performance in noisy environments, have been based on modifying the classifier to work with the reliable components of the spectrographic representation of the speech signal. That is the case of the works of Drygajlo et al. [8]

or Padilla et al. [3]. In these systems, the unreliable log-Mel spectral components are integrated out of the GMM distributions to get the speaker likelihood. This technique is known as marginalization.

Marginalization has several drawbacks. On the one hand, recognizers are constrained to use Mel spectral features that are known to produce worse performance than Mel frequency cepstral coefficients (MFCC). On the other side, by using incomplete spectrographic data we are not able to apply certain feature processing steps that are known to improve considerably the results. These processing steps include mean normalization, feature warping [9] or added time derivatives .

For these reasons, in this paper we are taking an alternative approach by trying to estimate the true values of the unreliable spectrographic components from the reliable ones. Once we get the complete time frequency representation of the signal, we are able to compute MFCC features, and apply whatever post-processing step to the features. Besides, we do not need to modify the recognizer so we can use anyone at our disposal. The algorithm we have chosen to compensate for the *U* components is cluster-based reconstruction which has proven to be very effective in speech recognition tasks as it is reported in the work of Raj et al. [10] [11].

3.1. The algorithm

The Cluster-based Reconstruction (CBR) algorithm estimates the *U* components of the spectral vector from the *R* ones of the same vector using a statistical model that relates both of them. This method is based on the assumption that the sequence of observations is an independent, identically distributed random process. This assumption is used by the most successful text independent speaker verification approaches too. Therefore, it is expected to have good results for MF compensation in speaker verification systems.

This algorithm models the distribution of log-Mel spectral vectors for clean signals as a mixture of Gaussian distributed clusters. The mean, covariance and a priori probability of each cluster can be estimated from a training corpus using maximum likelihood estimation via the expectation maximization (EM) algorithm [12].

Let Y be the noisy spectral vector and X the reconstructed spectral vector and let Y_r, X_r and Y_u, X_u be their *R* and *U* components respectively. The first step to compensate for the *U* components is to determine the noisy vector probability of belonging to each cluster. This is given by

$$P(k|Y) = \frac{w_k P(Y|k)}{\sum_{j=1}^{k} w_j P(Y|j)} \tag{2}$$

where w_k is the a priory cluster probability.

To calculate the term $P(Y|k)$ we have to take into account that Y has *R* and *U* components, and that $X_r = Y_r$ and $X_u \leq Y_u$ for additive noises. Therefore we can evaluate the Gaussian distribution in the *R* components and integrate out the *U* ones. This integration supposes additive noise so, the estimated *U* components need to be less than the measured components

$$P(Y|k) = P(X_r, X_u \leq Y_u|k) = \int_{-\infty}^{Y_u} P(X_r, X_u|k) dX_u \tag{3}$$

If we suppose that the covariance matrices are diagonal this can be written as

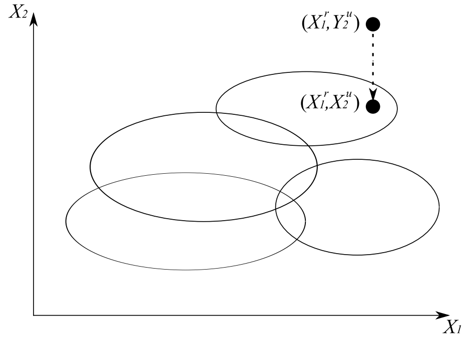


Figure 1: *CBR Process.*

$$\begin{gathered} P(Y|k) = \Pi_{i|X_i \epsilon X_r} \frac{1}{\sqrt{2\pi\sigma_{ki}}} exp(-\frac{1}{2}\frac{(X_i-\mu_{ki})^2}{\sigma_{ki}^2}) \times \\ \Pi_{i|X_i \epsilon X_u} \frac{1}{2}(1 + erf(\frac{Y_i-\mu_{ki}}{\sqrt{2}\sigma_{ki}})) \end{gathered} \tag{4}$$

where erf is the Gauss error function.

We can get an estimation of the clean value of the unreliable components from each cluster based on its distribution maximizing its likelihood given the measured reliable and unreliable components as

$$\hat{X}_u^k = \arg\max_{X_u} \{P(X_u|k, X_u \leq Y_u, X_r = Y_r)\} \tag{5}$$

Assuming diagonal covariance matrices this can be reduced to

$$\hat{X}_u^k = min(Y_u, \mu_{kr}) \tag{6}$$

where μ_{kr} is the Gaussian means of the unreliable components of the associated cluster.

Finally, we can get the overall unreliable components using the posterior membership probabilities to combine, by a weighted sum, the unreliable components estimations given by each cluster.

$$\hat{X}_u = \sum_{k=1}^{K} P(k|Y) \hat{X}_u^k \tag{7}$$

Figure 1 shows an example of the algorithm. Using the reliable component X_1, the procedure determines what cluster the feature vector belongs to, and substitute the unreliable component Y_2 with a clean estimation X_2. Once we have recovered the full Mel spectral vector, we are able to calculate the MFCC with their time derivatives and apply any preprocessing technique we need prior to the recognizer input.

4. Experiments and results

In order to evaluate the behavior of the MFs techniques combination in front of corrupted signals, a speaker verification experiment was carried out using the 1conv4w-1conv4w task of the 2006 NIST SRE [13].

4.1. Detection and compensation of unreliable components

To implement the mask estimator based on spectral subtraction we used the classical algorithm of Berouti et al. [14] and the noise estimator of Martin work [15].

The noisy signals were segmented with 25 msec. Hamming window overlapped 15 msec. and passed through 24 Mel filters bank. Then, noise estimator was applied, taking decision of reliability presented in equation 1, to obtain the unreliable components of the noise corrupted speech.

Once the mask estimation was done, the Cluster-based Reconstruction algorithm makes an estimation of the unreliable components. These reconstructed log-Mel spectra are then used to calculate the MFCC features that will be the input to the speaker verification system.

4.2. Speaker verification protocol

In this task, the enrollment and test utterances contain around 2 minutes of speech after voice activity detection. There are a total of 810 target models with 3176 true trials and 42079 false trials. It has used clean speech to train the target models and contaminated test signals with different levels of white noise selected to get several mean SNR, from 5 to 20 dB.

Our acoustic features are 15 MFCC plus first and second derivatives and C0 derivatives resulting in a total of 47 features. On the one hand, we have got results using no feature normalization at all to prove the capacity of our MF approach to cope with noise on its own. On the other hand, we have repeated the experiments using feature warping over 3 seconds in order to proof the benefits of being able to use feature normalization techniques together with the MF approach.

A gender dependent Universal Background Model (UBM) of 512 Gaussians is used. This model is trained using NIST SRE 2004 database containing 124 male and 184 female speakers with several utterances each one of them. The means of target models are adapted from the UBM using relevance MAP [16]. Classification is performed evaluating the log-likelihood ratio between the target and the UBM model for the test signal. Gender dependent cluster models for CBR are trained from the same dataset as UBM using different number of Gaussians.

4.3. Results

The first experiment we have conducted was intended to determine the optimal number of Gaussians needed for reconstruction. For that purpose, we have got results comparing baseline and MF cluster-based reconstruction with different number of clusters between 64 and 1024 using test signals contaminated with a SNR of 10 dB. The experiment has been repeated using feature warping and no feature normalization. In Table 1, we show the equal error rate (EER) and improvement percentage relative to the baseline of this experiment.

Figure 2 shows DET curves using no feature normalization and feature warping respectively, results with number of cluster over 256 are not plotted in order to preserve clarity.

We have got an amazing improvement when no feature normalization is applied nearly reaching clean signal performance. When using feature warping the challenge is bigger, but MF achieves a considerable improvement. The great capacity of feature warping of increasing robustness against channel mismatch, additive noise or even headset non-linearity it is well known. As a matter of fact, most sites participating in NIST evaluations use it in their systems. As we can see in Table 1, feature warping on its own is able to provide better results than MF compensation alone. That means it does a great deal of the same job as MF does. However, the benefits of using both techniques together are not negligible producing around a 17 percent of improvement compared to using feature warping only. This encourages us to think reconstruction of missing spectral component is the right path to follow in order to take advantage easily of the existing techniques to build robust speaker recognition systems.

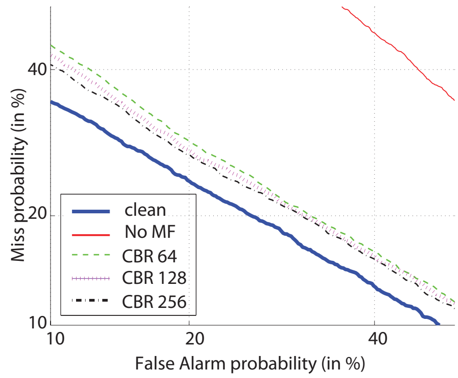


Figure 2: DET curves to SNR=10 dB

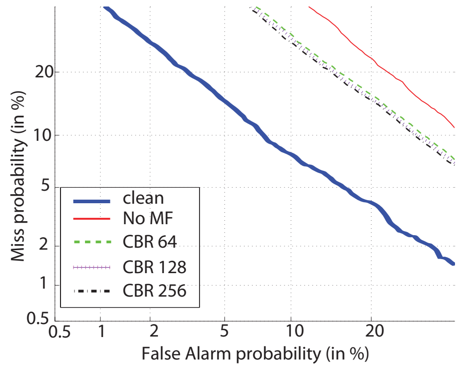


Figure 3: DET curves to SNR=10 dB with features normalization

Results show there is little improvement as we increase the number of clusters getting the best performance with 256 with no feature normalization and 512 with feature warping. We have found there is no improvement if we use more clusters. This could be explained by the fact that if we increase the number of clusters, they become more similar among them. Considering that cluster membership is estimated using only the reliable components of the spectrogram, it becomes more difficult to select precisely the best cluster as the number of clusters rises.

We have repeated the experiment using signals contaminated with SNR between 5 and 20 dB. This time we have used only 256 clusters, what seems a good choice given the previous results. In Table 2, we give a summary of the obtained results. We have got interesting improvements for all SNR tested. Something curious we note is that with no feature normalization and a SNR of 20 dB EER outperforms clean signal one. We expected a more important decrease of the improvement with low SNR due to the fact that we have less reliable components to make the spectral reconstruction but results are quite good.

Table 1: *EER and Improvements to SNR = 10 dB.*

	No Feat. Norm.		Feat. Warp.	
	EER(%)	Δ(%)	EER(%)	Δ(%)
clean	22.3		8.7	
No MF	42.9	0	21	0
CBR 64	25.2	41.2	17.7	15.4
CBR 128	24.7	42.4	17.4	17.1
CBR 256	24.2	43.6	17.1	18.6
CBR 512	24.9	41.9	16.8	20
CBR 1024	24.3	43.5	17.3	17.6

Table 2: *EER and Improvements to SNR = 5-20 dB.*

SNR(dB)	20	15	10	5
EER(%) No MF	29.8	36.9	42.9	46.8
EER(%) MF	21.8	22.5	24.2	29.5
Δ(%)	26.8	39	43.6	36.9
Feature Norm.				
EER(%) No MF	13.37	16.95	21	27.2
EER(%) MF	11.5	13.5	17.1	22.5
Δ(%)	14.5	20.3	18.6	17.28

5. Conclusions and future work

In this paper the proposed MF techniques combination has shown its potentiality in providing robustness for speaker recognition systems. The results obtained with MF alone or in combination with feature normalization produced an important increase of verification performance. It is convenient to highlight some ideas:

Improvement obtained in speaker verification results show that SNR criterion is an effective method when trying to obtain the reliability of the corrupted speech spectral components. However the enhancement of SNR contributes to increase speech quality, but does not necessarily ensure the improvement of recognition performance, so in the future we will focus on criteria that use representative speaker features. We will evaluate mask estimation methods based on spectral features classification such as Seltzer et. al work [11].

Since mask estimation is the prior step in MF approach, we do not lose sight of the MF compensation step. In this work we have used a reconstruction technique originally designed for speech recognition. We must take into account the fact that we have used speaker independent cluster models. This means that reconstructed features will be made more speaker independent too. In speaker recognition applications this is a great drawback. Despite that, results show improvements since noise compensation is more important than the effect of using speaker independent models. Nevertheless, we think we could get even better results using cluster models adapted to the test signal. Future work will be oriented in that direction.

On the other side, we must take into account the fact that GMM distributions with diagonal covariance matrices have limited correlation information between features. In future work, we plan to perform MF reconstruction using more complex distributions that should be able to perform a more precise estimation of the U components values. Examples of these models are GMM with full covariance matrices or graphical models [17]. Graphical models have the capacity of modeling correlations between features or groups of features at any level of complexity, what can be very promising for the MF approach.

6. Acknowledgements

This work has been supported by the Spanish Government through national project TIN2008-06856-C0504.

7. References

[1] Benesty J., J.C., Huang Y. and Cohen I., "Noise Reduction in Speech Processing", Springer Topics in Signal Processing, 2, 2009.

[2] Raj B. and Stern R., "Missing-Feature Approaches in Speech Recognition", IEEE Signal Proc. Magazine, 2005.

[3] Padilla M, Quatieri T, and Reynolds D, "MF Theory with Soft Spectral Subtraction for Speaker Verification", 2006.

[4] Ming J., Hazen T., Glass J.R. and Reynolds D.A., "Robust Speaker Recognition in Noisy Conditions", IEEE Trans. on Speech and Audio Proc., 15: 1711-1723, 2007.

[5] Pullella D., Kuhne M. and Togneri R., "Robust Speaker Identification Using Combined Feature Selection and Missing Data Recognition", ICASSP, 2008.

[6] Kuhne M., Pullella D., Togneri R. and Nordholm, S., "Towards the use of full covariance models for missing data speaker recognition", ICASSP, 2008.

[7] Cerisara C., Demange S. and Haton J-P., "On noise masking for automatic missing data speech recognition: a survey and discussion", Computer Speech and Language, 21(3): p. 443-457, 2007

[8] Drygajlo A. and El-Maliki M., "Speaker Verification in Noisy Enviroments with Combined Spectral Subtraction and MF Theory", Signal Proc. Laboratory, Swiss Federal Institute of Technology at Lausanne, 1998.

[9] Pelecanos J. and Sridharan S., "Feature warping for robust speaker verification", Speaker Odyssey, 2001.

[10] Raj B., Seltzer M. and Stern R.M, "Reconstruction of MFs for robust speech recognition, Speech Communication, 43, 2004.

[11] Seltzer M., Raj B. and Stern R.M., "A Bayesian classifier for spectrographic mask estimation for MF speech recognition", Speech Communication, 43, 2004.

[12] Dempster A.P., Laird N.M. and Rubin D.B., "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, 1977.

[13] "The NIST year 2006 speaker recognition evaluation plan", [Online] Available:
http://www.nist.gov/speech/tests/spk/2006/index.htm.

[14] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", IEEE ICASSP, 1979.

[15] Martin R., "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", IEEE Trans. on Speech and Audio Proc., 9, 2001.

[16] Reynolds D., Quatieri T. and Dunn R., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Proc., 10, 2000.

[17] Bilmes, J., "Graphical Models and Automatic Speech Recognition", Mathematical Foundations of Speech and Language Proc., pp. 191-235, 2004.

FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

On Line Vocal Tract Length Estimation for Speaker Normalization in Speech Recognition

William R. Rodríguez, Oscar Saz, Antonio Miguel and Eduardo Lleida

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Zaragoza, Spain
{wricardo,oskarsaz,amiguel,lleida}@unizar.es

Abstract

This paper presents the results on an Automatic Speech Recognition (ASR) framework that takes advantage of robust vocal tract length estimation methods for improving the performance of speech recognition in the presence of speakers with different conditions in age and gender. Well known techniques for Vocal Tract Length Normalization (VTLN) usually require previous stages for the estimation of the best warping factor for a given speaker, either by Maximum Likelihood (ML) estimates or by the calculation of acoustic features from the speakers like formant frecuencies through several utterances. This paper will show how to use robust framewise estimations of the vocal tract length to obtain a speaker dependent warping factor for achieving major improvements over all conditions of the TIDigits database. In the end, an updating function will be used to calculate an on-line estimate of the vocal tract length and the warping factor to use real time VTLN in speech recognition with similar results to the off-line strategies.

Index Terms: vocal tract length estimation, speech recognition, speaker normalization

1. Introduction

The mismatch between the set of speakers used to train a given acoustic model for Automatic Speech Recognition (ASR) and the set of speakers which are recognized in that ASR system can seriously degrade the performance of the recognition results. A well known source of mismatch are the anatomical features that different speakers may have regarding the structure or their vocal tract. The vocal tract, and more precisely its length, varies largely from one speaker to another, especially if the range of speakers gathers males, females, adults or children. All this can make an ASR system trained on adults perform poorly in the presence of children speech and vice versa.

Different possibilities have arisen for the reduction of this mismatch between training data and recognition data. Some of them require a re-training of the acoustic models, as in speaker adaptation techniques like Maximum A Posteriori (MAP) [1] or Maximum Likelihood Linear Regression (MLLR) [2]; while some others act on the speech signal and keep the models unchanged. Vocal Tract Length Normalization (VTLN) is a well known technique for reducing this mismatch without modifying the initial acoustic modeling [3]. It considers that the main difference between two speakers is the change in the frequency axis due to the difference in vocal tract length between the speakers.

However, VTLN techniques usually require of large computational delays, as they need to process speech data from the speaker in advance to estimate which is the best transformation from the speaker's frequency axis to the target speakers' frequency axis. This makes that most of VTLN-based techniques can not provide their improvement in the ASR performance in a real-time situation. The proposal in this work wants to advance in the field of providing this improvement in a real-time on-line framework. It makes use of robust speech processing algorithms to give a frame-by-frame estimation of the vocal tract length of the speaker, in such a way that it allows for providing a transformation factor for a given speaker without requiring more information that the current frame and previous frames.

The paper is organized as follows: Section 2 will review the basis of VTLN techniques and the different existing approaches. Section 3 will present the signal processing techniques which lead to a robust framewise vocal tract length calculation for all speakers and Section 4 will present the three VTLN techniques evaluated in this paper, including the on-line real-time approach. Next, in Section 5 the results and improvements achieved with the evaluated methods over the different conditions of the TIDigits database will be shown. Finally, Section 6 will provide the conclusions to this work.

2. Vocal Tract Length Normalization

The aim of VTLN is to provide a warping function that transforms the frequency axis from a given speaker (f) to the frequency axis of a target speaker or a target group (f'). Many different possibilities have been researched in the literature to provide the function which reflects this transformation, from piecewise linear approaches to exponential functions. All of them depend on a warping factor, α like in Equation 1, which contracts or expands the spectrum of the speech signal in the desired way [3].

$$S_{warped}(f) = S_{unwarped}(f'(\alpha, f)) \tag{1}$$

A warping factor that contracts the frequency axis is used to transform speakers with shorter vocal tracts (mainly children and women) towards speakers with the longest vocal tracts (i.e. men), and a warping factor that expands this axis is used for warping speakers with longer vocal tracts (men) towards the shortest ones (children or women). A more efficient proposal has been proposed for ASR consisting in transforming and warping the Mel-scale filter banks during the Mel-Frequency

This work was supported by national project TIN2008-06856-C05-04 from the Spanish government and Santander Bank scholarships

Cepstral Coefficients (MFCC) calculation, instead of warping all the input frames from the speaker.

2.1. Estimation of the Warping Factor

The estimation of the warping factor α for a given speaker or utterance is the most delicate part in the use of VTLN. An inadequate factor may reduce the potential improvement provided by VTLN, or even produce a loss in performance.

Two trends in the proposals for estimating the warping factor can be observed in the literature: On one hand, Maximum Likelihood (ML) based proposals select that warping factor which achieves the highest score by forced aligning several versions of the input utterances, warped with different factors, to the acoustic model for recognition [3]; on the other hand, feature-based proposals use acoustic features from the speaker like formants, or a combination of them, to estimate the warping factor, as it is known that the formant frequencies correlate to the vocal tract length of the speaker [4].

3. Vocal Tract Length Estimation

Although many methods rely on the estimation of speaker features which can be correlated to the vocal tract length to calculate the optimum warping factor, there have been little efforts to estimate the actual vocal tract length. Difficulties in the estimation of this anatomical measure, especially in the presence of voices with a high fundamental frequency, have limited the development of methods based on direct vocal tract length estimation. This Section will describe a robust method to estimate this value for all possible speakers with the aim of using it as estimator of the optimal warping factor.

Modeling the vocal tract as a uniform lossless acoustic tube, its resonant frequencies given by Equation (2) are uniformly spaced, where $v = 35300$ cm/s is the speed of sound at $35\,^{\circ}$C, and l is the length of the uniform tube in cm.

$$F_k = \frac{v}{4l}(2k-1), k = 1, 2, 3, \ldots \qquad (2)$$

The estimation of the length was proposed in [5], and it can be reduced to fitting the set of resonance frequencies of a uniform tube, which are determined solely by its length l. Therefore, the problem can be approximated to minimizing Equation 3, where $D(\tilde{F}_k, (2k-1)F1)$ is a function that express the difference between the measured formants $\tilde{F}_k(k = 1, ..., M)$ and the resonance of the uniform tube.

$$\varepsilon = \sum_{k=1}^{M} D(\tilde{F}_k, (2k-1)F_1) = \sum_{k=1}^{M} D(\tilde{F}_k, (2k-1)\frac{v}{4l}) \qquad (3)$$

From [5], the error measure given in equation (3) can be turned in Equation 4 using the distance function between the measured formants$(\tilde{F}_k)$ and the odd resonances of a uniform tube, $(2k-1)F_1$.

$$\varepsilon = \sum_{k=1}^{M} \frac{(\frac{\tilde{F}_k}{2k-1} - F_1)^2}{F_1} \qquad (4)$$

The formant frequencies $\tilde{F}_k$ are extracted using traditional Linear Prediction Coefficients (LPC) method with order $p = 8$, over a 25 ms long speech frame. The filter coefficients for the all-pole vocal tract model are obtained through Durbin's recursion using the autocorrelation method, after Hamming-windowing the pre-emphasized speech frame.

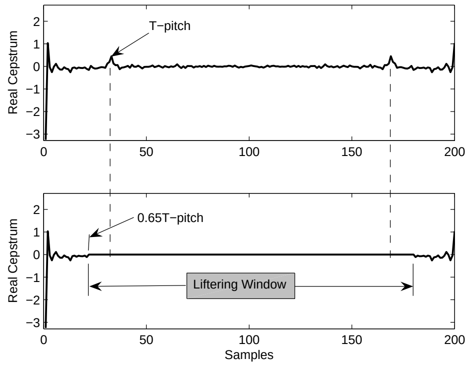


Figure 1: *Effect of liftering in the real cepstrum domain*

Finally, the vocal tract length can be obtained with the expression in Equation 5 which makes use of the estimated resonance frequency of the uniform tube (F_1), calculated from Equation 4 as in Equation 6.

$$VTL = \frac{v}{4F_1} \qquad (5)$$

$$F_1 = \left(\frac{1}{M} \sum_k \left(\frac{\tilde{F}_k}{2k-1} \right)^2 \right)^{1/2} \qquad (6)$$

3.1. Robust Formant Estimation in High Pitch Voices

The formant measurement is technically difficult. The situation is less severe in male adult cases in which the fundamental frequencies (F0) are low [6]. In women and children $F0$ increases, so $F0$ and its harmonics could get closer to the range of the formant values affecting the estimation [7]. The conventional autocorrelation method with the LPC parameters works well in signals with a long pitch period (low-pitched), but as the pitch period in high-pitched speech is small, the periodic replicas cause aliasing in the autocorrelation sequence. In that case it is required to separate these effects in order to obtain formants not contaminated by $F0$ by means of homomorphic analysis.

The main idea within the homomorphic analysis is the deconvolution of a segment of speech $x[n]$ into a component representing the vocal tract impulse response $e[n]$, and a component representing the excitation source $h[n]$ as in Equation 7.

$$x[n] = e[n] * h[n] \qquad (7)$$

The way in which such separation is achieved is through linear filtering of the cepstrum, defined as the inverse Fourier transform of the log spectrum of the signal. As the cepstrum in the complex domain is not suitable to be used because of its high sensitivity to phase[8], the real-domain cepstrum $c[n]$ defined by Equation (8) is used, where $X(k)$ is the N-point Fourier transform of the speech signal $x[n]$.

$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} ln\left|X(k)\right| e^{j\frac{2\pi}{N}kn}, 0 \leq n \leq N-1 \qquad (8)$$

The values of $c[n]$ around the origin correspond primarily to the vocal tract impulse information, while the farthest values are affected mostly by the excitation. Knowing previously the value of the pitch period T_{pitch} from the LPC analysis using the autocorrelation method it is possible to filter the cepstrum signal (liftering) and use the liftered signal to find the formant frequencies, once the signal is back in the time-domain.

A liftering window with the length of $0.5T_{pitch}$ has been proposed in [9] or $0.6 - 0.7T_{pitch}$ in [10]. In this work, the liftering window $w[n]$ is $0.65T_{pitch}$ and the effect of appliying $w[n]$ in the real cepstrum domain can be observed in Figure 1.

4. VTLN Techniques Applied in ASR

Two VTLN techniques were initially compared in this work, first one was a state of the art ML-based approach, while second one was based on the proposal for robust vocal tract length calculation in Section 3. An exponential function was used for the warping of the Mel-scale bank filters in the MFCC calculation. No model adaptation was performed in the experiments and, hence, no Jacobian compensation was done as previous results reported how this feature degraded performance in the presence of unadapted models [11, 12].

The ML-based technique was based on the diagram seen in Figure 2 [3], where an initial ASR stage obtained an estimate of the transcription of the uttered sentence. A set of n Viterbi alignment decoders using a set of warping factors $\{\alpha_1 ... \alpha_n\}$ decided the most likely of those warping factors according to the score achieved by each decoder. Finally, that warping factor was used in a second ASR stage which made use of VTLN to improve the estimation of the output utterance and provide a final result. This work used 11 warping factors in the Viterbi decoding phase, ranging from 0.9 to 1.1 in 0.02 intervals.

The feature-based technique to be evaluated in this work used the framewise estimation of the vocal tract length as seen in Section 3. This estimation provided a value of the vocal tract length in the sonorant frames and a void output in the rest of frames (silence and unvoiced sounds). From all the valid estimations of one speaker, the mean of the vocal tract length for the speaker was calculated (VTL_{spk}) and the warping factor was obtained as in Equation 9, where $\overline{VTL}_{model}$ was the mean of the vocal tract lengths calculated for all the speakers used in the training of the acoustic model, which could be easily done in the prior training phase. The factor λ was used to moderate the amount of warping applied, and was set to $\lambda = 0.5$ after a prior development set of experiments on smaller databases.

$$\alpha = 1 + \lambda \frac{\overline{VTL}_{model} - VTL_{spk}}{\overline{VTL}_{model}} \quad (9)$$

4.1. On-line Vocal Tract Length Estimation

The main drawback of the VTLN techniques shown previously was that they required previous stages to estimate the warping factor. The ML-based strategy required three stages for each utterance (initial recognition, decision of the warping factor and final recognition), while the feature-based strategy required that several utterances from a speaker were available to estimate robustly the mean of the vocal tract length from that speaker. These approaches were not feasible for real world applications which should provide a real-time decoding of the input speech utterance.

The proposal in this work was, as the calculation of the vocal tract length was robust in a framewise approach, to re-estimate the vocal tract length for each frame as in Equation 10, where β is the memory factor of the system. The value estimated for the vocal tract length in a given frame i only depended on its value in the previous frame $i - 1$ and the actual

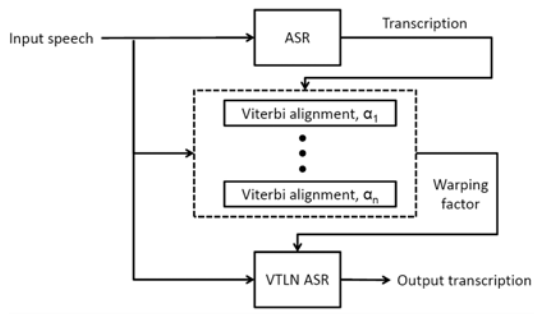


Figure 2: *ML-based VTLN diagram*

value of vocal tract length estimated for frame i: $VTL(i)$. This approach avoided the influence of local variations of the vocal tract length, while tending to the mean value of the speaker when sufficient frames were analyzed. A memory factor of $\beta = 0.99$ was used for the experimentation.

$$VTL_{spk}(i) = \beta * VTL_{spk}(i-1) + (1-\beta) * VTL(i) \quad (10)$$

This way, when a speaker accessed for the first time to the ASR system, the vocal tract length was initialized with the vocal tract length of the target model as in Equation 11; as every new sonorant frame was available and a value of the vocal tract length was provided for that frame, the vocal tract length for the speaker was updated according to Equation 10 (again with $\lambda = 0.5$), and the warping factor for that frame calculated as in Equation 12.

$$VTL_{spk}(0) = \overline{VTL}_{model} \quad (11)$$

$$\alpha(i) = 1 + \lambda \frac{\overline{VTL}_{model} - VTL_{spk}(i)}{\overline{VTL}_{model}} \quad (12)$$

5. Experimental Framework and Results

The evaluation of the techniques proposed here was made over the TIDigits database [13]. This corpus contains 25 boys, 26 girls, 55 men and 57 women for the training of models and 25 boys, 25 girls, 56 men and 57 women for the recognition evaluation. Seven conditions were designed with seven different acoustic models trained for each condition: Boys, girls, men, women, adults (men and women), child (boys and girls) and all speakers. Recognition was performed over all the 163 speakers available for evaluation.

A set of 11 word Hidden Markov Models (HMM) representing digits in English were trained for each condition. An ETSI-like front end was used to extract the MFCC parameters from each signal, using the 12 first static parameters ($c1$-$c12$) plus log-energy and their first and second derivatives for the final 39 dimension feature vectors. The ASR system used for the experiments was a state of the art Viterbi decoder.

The baseline results in first row of Table 1 in terms of Word Error Rate (WER) showed the big influence of acoustic mismatch between models and speakers for recognition. The recognition results provided are the ones obtained for the recognition of the test data from boys, girls, men and women. Worst results were achieved with models from men speech, as men presented the longest vocal tracts, which separates greatly from the rest of the speakers. On the other edge, girls have the shortest vocal

Table 1: *Results in WER for the TIDigits database: Models trained on boy, girl, man, woman, adult, child and all speech*

	Boy	Girl	Man	Woman	Adult	Child	All
Baseline	7.37%	19.21%	25.17%	5.32%	2.01%	8.20%	0.65%
Off-line ML-based VTLN	2.47%	5.26%	8.58%	1.28%	1.05%	2.40%	0.57%
Off-line vocal tract estimated VTLN	2.84%	5.37%	11.25%	1.94%	1.19%	2.81%	0.66%
Off-line vocal tract estimated VTLN (liftering)	2.35%	3.92%	10.15%	1.57%	1.07%	2.18%	0.65%
On-line vocal tract estimated VTLN	2.61%	4.78%	10.48%	1.82%	1.18%	2.49%	0.65%

Table 2: *Mean vocal tract length in cms with standard deviation intervals estimated for the speaker groups in the TIDigits database*

Train speakers				
	Boy	Girl	Man	Woman
VTL	16.0±0.64	15.5±0.65	18.8±0.67	16.6±0.64
Test speakers				
	Boy	Girl	Man	Woman
VTL	15.9±0.74	15.4±0.58	18.8±0.71	16.6±0.63

tract and the models trained from girls speech also performed poorly. The model trained with all the speech matched perfectly the recognition speakers and achieved a 0.65% of WER.

The off-line techniques to be evaluated in this work achieved the results in the second, third and fourth rows of Table 1 for the ML-based VTLN and two versions of the feature-based technique respectively. The performance of both techniques was similar, with some differences across all the conditions which indicated that the vocal tract length estimation in Section 3 was as good as the state of the art techniques for speaker normalization in ASR. It was noticed the improvement that the use of liftering, seen in Section 3.1, produced in reduction of the WER; showing the need for using robust formant estimation techniques when dealing with variable speech.

More precisely, the VTLN based on direct estimation of the vocal tract length with liftering achieved better results with those models trained on boys, girls and children altogether; while the ML-based technique performed better with models trained on men, women and adults. Performance on the model trained with all speakers was similar for both techniques. Table 2 shows the mean values with their standard deviation of the vocal tract lengths for all the speakers in the TIDigits database with the estimation method of Section 3. These values confirm the big mismatch between the different group, especially men versus the rest of the groups; and confirmed the need of applying the speaker normalization techniques studied in this work. The robustness of the estimation of these values was seen in their good statistical properties across speakers.

Finally, the on-line technique proposed in Section 4.1 achieved the results shown in fifth row of Table 1. Although a certain decrease of performance was noticed throughout most of the conditions, the results were comparable to those of the off-line version of the algorithm and confirmed the possibilities of performing real-time VTLN in ASR with results similar to the state of the art off-line techniques.

6. Conclusions

The main result and conclusion of this work is the development of a method for applying successful on-line speaker normalization in ASR. This method relies in a robust estimation of the vocal tract length from the speaker at the frame level, which allows to apply a warping factor which can be updated and improved as more data from the user is available. This overcomes the drawback from traditional VTLN techniques where speech utterances are processed in several stages to obtain the desired improvement. Furthermore, the proposed method relies only in voice features from the speaker and is not based in models and likelihood measures as previous approaches.

The use of techniques for enhancing the formant estimation methods and thus, improving the vocal tract length calculation, makes suitable this speaker normalization method for all types of speakers (children or adults), independently of the value in fundamental frequency of the speaker. This is a relevant improvement over other techniques for the estimation of acoustic features like formants which may face difficulties in the presence of speech with a high fundamental frequency.

7. References

[1] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[2] C.-J. Legetter and P.-C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of the parameters of continous density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[4] E.-B. Gouvea and R.-M. Stern, "Speaker normalization through formant-based warping of the frequency scale," in *Proceedings of Eurospeech*, Rhodes, Greece, 1997, pp. 1139–1142.

[5] F.-N. Burhan, A.-C. Mark, and P.-B. Thomas, "Unsupervised estimation of the human vocal tract length over sentence level utterances," in *Proceedings of ICASSP'00*, 2000, pp. 1319–1322.

[6] H. Traunmuller and A. Eriksson, "A method of measuring formant frequencies at high fundamental frequencies," in *Proceedings of Eurospeech)*, 1997, pp. 477–480.

[7] W. Rodríguez and E. Lleida, "Formant estimation in children's speech and its application for a spanish speech therapy tool," in *Proceedings of SLaTE*, Wroxhall Abbey, UK, 2009.

[8] R. Schafer and L. Rabiner, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[9] W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 2, pp. 43–51, 1986.

[10] M. Shahidur and T. Shimamura, "Formant frequency estimation of high-pitched speech by homomorphic prediction," *Acoustic Sci. and Tech.*, vol. 26, no. 6, pp. 502–510, June 2005.

[11] A.-N. Harish, D.-R. Sanand, and S. Umesh, "Characterizing speaker variability using spectral envelopes of vowel sounds," in *Proceedings of Interspeech'07*, Brighton, UK, 2009, pp. 1107–1110.

[12] S. Panchapagesan and A. Alwan, "Frequency warping for vtln and speaker adaptation by linear transformation of standard mfcc," *Computer, Speech and Language*, vol. 1, no. 23, pp. 42–64, 2009.

[13] R.-G. Leonard, "A database for speaker independent digit recognition," in *Proceedings of ICASSP'84*, San Diego, CA (USA), 1984, pp. 328–331.

Automatic Phonetic Segmentation

Jon Ander Gómez Adrián

Departmento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
jon@dsic.upv.es

Abstract

This paper presents an approach to automatic segmentation of speech corpora. The availability of sufficiently precise labelled sentences can avoid the need for a segmentation by human experts. The goal of this process is to prepare speech corpora both for training acoustic models and for concatenative text to speech synthesis.

Our system only needs the speech signal and the phonetic sequence for each sentence of a corpus. It estimates a GMM by using all sentences, where each Gaussian distribution represents an acoustic class. Then it combines the probability densities of each acoustic class with a set of conditional probabilities in order to estimate the probability densities of the states of each phonetic unit. A DTW algorithm fixes the phonetic boundaries using the known phonetic sequence. This DTW is a step inside an iterative process which aims to segment the corpus and re-estimate the conditional probabilities. A flat start setup is used to give initial values to the conditional probabilities.

Index Terms: automatic speech segmentation, phoneme boundaries detection, phoneme alignment

1. Introduction

The two main applications of phonetic level segmentation are text to speech synthesis and acoustic models training. In both cases it is useful to have as many labelled sentences as possible. Doing this labelling task by hand implies a great effort that can be very expensive. Furthermore, as some authors point, manual segmentations of a single corpus carried out by different experts can have significant differences, thus it is reasonable to use automatic segmentations. As an example, some researchers have given the same speech database to different human experts to segment it. Then, they evaluated the differences between the manual segmentations obtained. In [1], 97% of the boundaries within a tolerance interval of 20 ms were found, and 93% in [2].

There are several different approaches to the automatic segmentation of sentences when the phonetic sequence is available. Most of them are systems in two stages: the first one is performed by a Hidden Markov Model (HMM) based phonetic recognizer using the Viterbi forced alignment, and the second one adjusts the phonetic boundaries. In [1, 3, 4] different pattern recognition approaches are proposed for the local adjustment of boundaries. [5] presents a HMM based approach where pronunciation variation rules are applied and a recognition network is generated for each sentence. Then a Viterbi search determines the most likely path and obtains an adapted phonetic transcription for each sentence. This process is repeated until the adapted phonetic transcriptions do not change any more. Initial phone HMMs are generated with flat-start training using the canonical transcriptions of the sentences.

A Dynamic Time Warping (DTW) based method which aligns the spoken utterance with a reference synthetic signal produced by waveform concatenation is proposed in [6]. The known phonetic sequence of each sentence is used to generate the synthetic signal. The alignment cost function is computed using a combination of acoustic features depending on the pair of phonetic segment classes being aligned. In [7] a set of automatic segmentation machines are simultaneously applied to draw the final boundary time marks from the multiple segmentation results. Then, a candidate selector trained over a manually-segmented speech database is applied to identify the best time marks.

An approach inspired by the minimum phone error training algorithm for automatic speech recognition [8] is presented in [9]. The objective of this approach is to minimize the expected boundary errors over a set of phonetic alignments represented as a phonetic lattice.

A quite different approach is presented in [10], which uses an extension of the Baum-Welch algorithm for training HMM that uses explicit phoneme segmentation to constrain the forward and backward lattice. This approach improves the accuracy of automatic phoneme segmentation and is even more computationally efficient than the original Baum-Welch.

A technique which modifies the topology of the HMM to control for duration is presented in [11]. The prototype for all phones is defined as a 5-state left-right topology with duration control states at each end. This topology improves segmentation accuracy by reducing the probability of remaining in the beginning and end states as these states model the boundaries between phonetic units. The acoustic vectors at the transition from one phonetic unit to the other are clustered at these states.

In this paper we present a phonetic level automatic speech segmentation technique based on the same idea of altering the topology of the HMM. Nevertheless, we calculate the emission probabilities in a different way, the forced alignment is performed by means of a DTW algorithm and we do not use any manually segmented sentences. Emission probabilities are computed by combining acoustic probabilities with conditional probabilities estimated *ad hoc*. The conditional probabilities reflect the relation between the acoustic and the phonetic probability densities. The estimation of these conditional probabilities is done by means of an iterative process of progressive refinement which segments all sentences of the training set at every step. The initial values given to the conditional probabilities are calculated using a flat start setup, and the acoustic probability densities are computed from a GMM (Gaussian Mixture Model) obtained as a result of a clustering process.

Next, we describe in Section 2 the proposed approach for automatic speech segmentation. Then, in Section 3, we show and comment the experimentation results. Finally, we conclude in Section 4.

2. System description

A DTW algorithm to automatically segment each sentence is used. This algorithm aligns the sequence of states with respect to the sequence of acoustic frames. The sequence of states is composed by concatenating the model of each phonetic unit from the known phonetic sequence. There are two relevant constraints on the topology of the models, that are the number of states and the number of duration control states. Figure 1 shows a model with 8 emitting states and 3 duration control states at both sides, similar to the ones proposed in [11]. The number of states sets the minimum number of frames assigned to each phonetic unit.

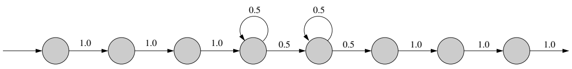


Figure 1: *An 8 emitting states HMM with 3 duration control states at each side.*

The alignment cost function used in the DTW algorithm uses $\Pr(e_i^u|x_t)$ as the emission probability, which represents the posterior probability of each state given an acoustic vector, where e_i^u is the i-th state of the phonetic unit u. For each acoustic frame x_t we obtain another vector with the phonetic level probabilities $\{\Pr(e_i^u|x_t)\}\ \forall u \in U,\ i = 1..E(u)$, where U is the set of phonetic units and $E(u)$ is the number of states of the phonetic unit u. Applying this process to each frame of an utterance we obtain as a result a sequence of vectors with the probability of each state of each phonetic unit.

2.1. Acoustic probabilities

The acoustic probabilities are computed from a GMM where each Gaussian distribution represents an acoustic class. The GMM is estimated by means of a clustering procedure using as training samples all the acoustic vectors from every training sentence. The unsupervised learning of the means and diagonal covariances for each acoustic class has been done by maximum likelihood estimation as described in [12].

The underlying idea of our approach is based on the fact that, once we transform the waveform into frames (d-dimensional acoustic vectors), they are distributed into a region of $\mathbb{R}^d$, in such a way that more dense subregions are formed according to similar acoustic phonetic features. The dense subregions can be related to different acoustical manifestations of speech. Each phonetic unit can have many acoustically different ways of being pronounced due to, among other phenomena, the mood and the accent of the speaker. Context is also an important factor that affects to the way a phoneme is pronounced, previous and following phonemes influence the one being uttered. In addition, not all the possible acoustic manifestations are related to an only phoneme, but many of them fall in the intersection of two or more phonemes. So, we can conclude that the subregions of each phoneme are neither isolated nor continuous.

In short, we can consider that the phonetic units are distributed in overlapped subregions inside $\mathbb{R}^d$, and that the natural acoustic classes allow us to model more precisely the region of this space where acoustic frames are distributed. It is easy to see that the number of acoustic classes will be much larger than the number of phonetic units.

2.2. Phonetic probabilities

To take into account the different degrees of relation between a phonetic class and a phonetic unit we have used conditional probabilities estimated for this goal. So, for each state we have $\Pr(a|e_i^u)$, which represents the conditional probability that the acoustic class a has occurred having that the phonetic unit u has been pronounced and its internal state e_i^u is active.

We can compute the class-conditional probability density function of observing the acoustic frame x_t assuming that the acoustic class a has been manifested, $p(x_t|a)$, according to the GMM. Nevertheless, we need the phonetic-conditional probability density function of observing the acoustic frame x_t given the state e_i^u, that can be calculated as follows:

$$p(x_t|e_i^u) = \sum_{a \in A} p(x_t|a) \cdot \Pr(a|e_i^u) \tag{1}$$

Applying the Bayes rule we can obtain the posterior probability of each phonetic state given a frame:

$$\Pr(e_i^u|x_t) = \frac{p(x_t|e_i^u)\pi(e_i^u)}{\sum\limits_{v \in U} \sum\limits_{j=1}^{E(v)} p(x_t|e_j^v)\pi(e_j^v)} \tag{2}$$

where $\pi(\cdot)$ is the prior probability of each state of each phonetic unit. In our approach we consider that all the prior probabilities are the same, so we can eliminate them. Taking this into account and expanding Equation 2 according to Equation 1 we have:

$$\Pr(e_i^u|x_t) = \frac{\sum_{a \in A} p(x_t|a) \cdot \Pr(a|e_i^u)}{\sum\limits_{v \in U} \sum\limits_{j=1}^{E(v)} \left(\sum_{a \in A} p(x_t|a) \cdot \Pr(a|e_j^v)\right)} \tag{3}$$

The conditional probabilities $\Pr(a|e_i^u)$ can be estimated either in a supervised way, using a manually segmented and labelled corpus, or in an unsupervised way, using an iterative process of progressive refinement like the one proposed here.

The initial values of the conditional probabilities are calculated using a flat start setup. Then the iterative process that re-estimates the conditional probabilities starts and goes on until there are no significant changes.

2.3. DTW state alignment

The DTW algorithm used to align the sequence of states against the sequence of acoustic frames obtains the phonetic level segmentation. For each sentence we can build the sequence of states by concatenating the models of the phonetic units that were pronounced. It is important to highlight that each phonetic unit can have a different number of states according to its nature. Figure 2 shows the allowed movements inside the DTW matrix for an example corresponding to the join between two phonetic unit models, with one duration control state at each end. We can observe that horizontal movements are forbidden for the duration control states, which only allow diagonal movements. Vertical movements are not allowed, since it would imply that an only frame is assigned to more than one state.

3. Experimentation results

In this section we describe the performed experiments and the obtained results. First, we present the speech corpus used for testing our system and comment the evaluation criteria. Next, the results for different combinations of the total number of

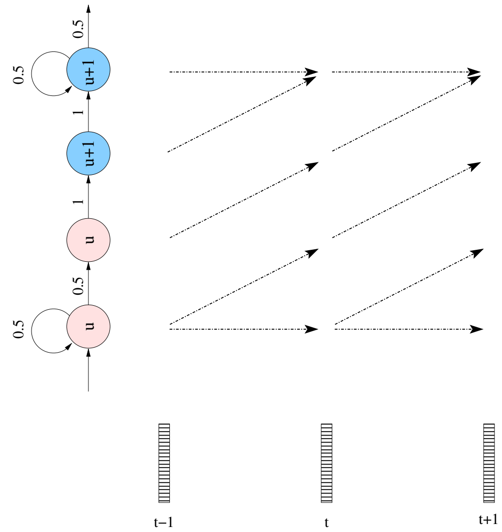


Figure 2: *Example of possible movements in our DTW focused on the join between two phonetic units.*

states and the number of duration control states at each end are presented. Our goal was to find the optimal topology for all phonetic units, so we repeated the training and testing processes using different configurations. When the best configurations were detected, we focused on those phonetic units we considered should have a different number of states, so we tried to achieve better results modifying their particular topology.

We also tested what happened when we did not allow the re-estimation of the transition probabilities of central states. The obtained results show that re-estimating the transition probabilities runs worse than not re-estimating them for every topology tested. Only results obtained with the best modality are presented here.

3.1. Speech corpus

The phonetic subcorpus from *Albayzin* database [14] that was used for the experiments is composed by 6,800 utterances (around six hours of speech) obtained by making groups from a set of 700 different sentences uttered by 40 different speakers. 1,200 sentences manually segmented and labelled were used for testing and the remaining 5,600 sentences were used for training. There are no common speakers between the training and test subcorpora.

Each acoustic frame is a 39-dimensional vector composed by the normalized energy, the first 12 Mel frequency cepstral coefficients and their first and second time derivatives. Each acoustic frame was obtained using a 20 ms Hamming window every 5 ms.

It is worth to say that we did not use the original training and test subsets that had the database because all the manually segmented and labelled sentences were included in the training subcorpus. So, we used the subset of 1,200 manually segmented and labelled sentences for testing and the 5,600 remaining sentences for training.

3.2. Evaluation criteria

The evaluation criteria most widely used in the literature is to measure the agreement of the obtained segmentation with respect to a manual segmentation. Usually the percentage of boundaries whose error is within a tolerance is calculated for a range of tolerances [1, 2, 13].

As discussed in the introduction, some researchers have wondered whether or not a manual segmentation is a valid reference [1, 2]. To evaluate it, they gave the same speech database to different human experts to segment it, and then evaluated the differences between them. In the study presented in [1], 97% of the boundaries within a tolerance of 20 ms were found and in [2] 93%. We interpret this agreement as the maximum accuracy for a segmentation system, since a system that reaches 100% compared with a manual segmentation will at least differ around 95% with another manual segmentation for the same speech database.

3.3. Experimental results

Our system has been evaluated for different combinations of the number of emitting states and duration control states. Table 1 presents the results obtained using different $E \times B$ topologies, where E represents the number of emitting states and B the number of duration control ones. Furthermore, Figure 3 shows a graphic representation of the same results, where a significant improvement is easily observed when tolerance increases from 10 to 20 ms.

Table 1: *Percentage of correctly fixed phonetic boundaries for a range of tolerances.*

	Tolerance en ms					
Topology	5	10	15	20	30	50
3x0	23.8	47.2	66.9	80.6	92.0	97.9
3x1	26.6	51.9	70.0	82.0	92.4	97.8
5x0	28.4	52.5	71.3	83.3	93.4	98.3
5x1	32.3	58.6	76.5	87.0	94.7	98.6
5x2	36.7	62.6	78.7	88.0	94.8	98.5
6x2	37.0	63.1	79.2	88.5	95.2	98.8
7x0	32.6	58.8	75.9	85.9	94.9	98.6
7x1	34.0	61.3	79.0	88.5	95.5	98.8
7x2	34.6	62.1	79.5	88.7	95.4	98.7
7x3	35.7	63.6	80.5	89.1	95.4	98.8
8x3	40.3	67.2	81.9	89.0	95.8	99.0
9x2	37.4	65.7	81.2	88.5	92.5	95.3
9x3	39.4	67.6	82.2	89.0	95.7	98.9
9x4	42.4	69.2	82.0	88.6	95.3	98.8

Results show a significant improvement when duration control states at each end are used. Also we can observe that the more restrictive a tolerance interval is, the more relevant is the improvement we achieve. For example, if $E = 7$ then the segmentation accuracy improves from 58.8% to 63.6% as B increases, for a tolerance error of 10 ms, and from 85.9% to 89.1% for 20 ms. By observing the results for different values of E we can detect a better performance when all the states except the central ones are duration control states.

As mentioned above, our system begins the learning process from a flat start setup and then iterates to re-estimate the conditional probabilities which relate the acoustic probability densities to the phonetic ones. Figure 4 shows the evolution of segmentation accuracy for several topologies within a tolerance interval of 20 ms. No significant improvements are obtained from 15th step, and we can clearly see the difference of segmentation accuracy when the 7×0 topology was used.

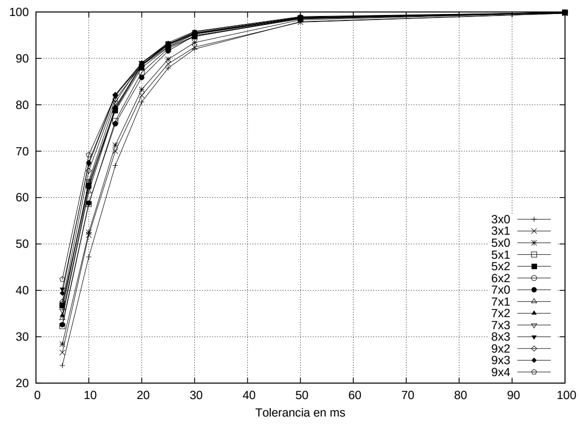


Figure 3: *Evolution of segmentation accuracy in function of tolerance error.*

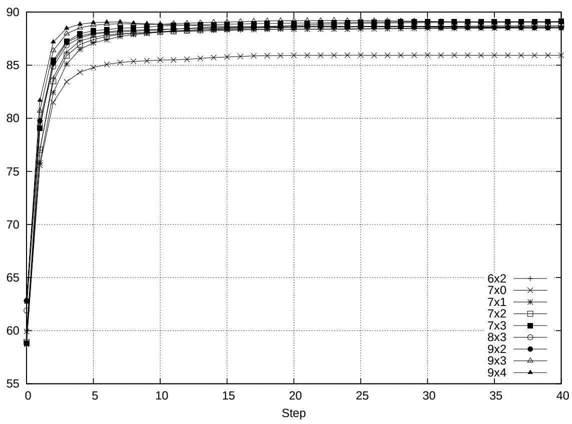


Figure 4: *Evolution of segmentation accuracy in function of iterative steps with a tolerance error of 20 ms.*

Taking into account that the subsampling rate is 200 Hz, a HMM with 8 emitting states forces a minimum phone duration of 40 ms, which is longer than usual for some phonetic units. The topologies of voiced plosives /b/, /d/ and /g/ differs from the topologies of the remaining units when using models with more than 7 emitting states. In this particular case, a 5×2 topology was used and the results improved significantly when this change was applied. The topologies of voiceless plosives /p/, /t/ and /k/ were not different from the topologies used for the rest of units. The burst of these plosives is always preceded by a short silence. So, voiceless plosives do not need a special topology because the frames of preceding silence are properly clustered by the HMM states. Finally, the phonetic unit representing silences is considered a special case, for which we used a 3×0 topology.

4. Conclusions

We have presented here an automatic segmentation technique which combines three ideas. The first one consists in using duration control states at each end of each HMM and in increasing the number of emitting states. This idea improves significantly the segmentation accuracy as it was shown by some researchers [11]. The second one, detailed in Section 2, deals with the way emission probabilities are calculated. The third idea consists in using a DTW algorithm to align the sequence of states against the sequence of acoustic frames.

The main goal of our approach is to automatically segment speech corpora for training acoustic models without making use of any subset of manually segmented and labelled sentences. A segmentation accuracy close to 90% within a tolerance of 20 ms enables our system to be used for this purpose. In addition, our system can be useful for concatenative text-to-speech synthesis.

5. Acknowledgments

I am very grateful to Marcos Calvo Lance, he supported me correcting English text.

This work was supported by the Spanish MICINN under contract TIN2008-06856-C05-02.

6. References

[1] Toledano, D. T., Hernández Gómez, L. and Villarrubia Grande, L., “Automatic Phonetic Segmentation”, in IEEE Transactions on Speech and Audio Processing, vol. 11, no. 6, pages 617–625, November 2003.

[2] Kipp, A., Wesenick, M.B. and Schiel F., “Pronunciation modelling applied to automatic segmentation of spontaneous speech”, in Proceedings of Eurospeech, 1997, pages 2013–1026, Rhodes, Greece.

[3] Sethy, A., Narayanan, S., “Refined Speech Segmentation for Concatenative Speech Synthesis”, in Proceedings of ICSLP, 2002, pages 149–152, Denver, Colorado, USA.

[4] Jarify, S., Pastor, D., Rosec, O., “Cooperation between global and local methods for the automatic segmentation of speech synthesis corpora”, in Proceedings of Interspeech, 2006, pages 1666–1669, Pittsburgh, Pennsylvania, USA.

[5] Romsdorfer, H., Pfister, B., “Phonetic Labeling and Segmentation of Mixed-Lingual Prosody Databases”, in Proceedings of Interspeech, 2005, pages 3281–3284, Lisbon, Portual.

[6] Paulo, S., Oliveira, L.C., “DTW-based Phonetic Alignment Using Multiple Acoustic Features”, in Proceedings of Eurospeech, 2003, pages 309–312, Geneva, Switzerland.

[7] Park, S.S., Shin, J.W., Kim, N.S., “Automatic Speech Segmentation with Multiple Statistical Models”, in Proceedings of Interspeech, 2006, pages 2066–2069, Pittsburgh, Pennsylvania, USA.

[8] Povey, D., Woodland, P.C., “Minimum Phone Error and I-smoothing for improved discriminative training”, in Proceedings of ICASSP, 2002, pages 105–108, Orlando, Florida, USA.

[9] Kuo, J.W., Wang, H.M., “Minimum Boundary Error Training for Automatic Phonetic Segmentation”, in Proceedings of Interspeech, 2006, pages 1217–1220, Pittsburgh, Pennsylvania, USA.

[10] Huggins-Daines, D., Rudnicky, A.I., “A Constrained Baum-Welch Algorithm for Improved Phoneme Segmentation and Efficient Training”, in Proceedings of Interspeech, 2006, pages 1205–1208, Pittsburgh, Pennsylvania, USA.

[11] Ogbureke, Kalu U., Carson-Berndsen, Julie, “Improving initial boundary estimation for HMM-based automatic phonetic segmentation”, in Proceedings of Interspeech, 2009, pages 884–887, Brighton, U.K.

[12] Duda, R. O., Hart, P. E. and Stork, D. G., *“Pattern Classification”*, John Wiley and Sons, second edition, 2001.

[13] Mporas, I., Ganchev, T., and Fakotakis, N., “A Hybrid Architecture for Automatic Segmentation of Speech Waveforms”, IEEE ICASSP 2008, pages 4457–4460, Las Vegas, USA.

[14] Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B. and Nadeu, C., “Albayzin Speech Database: Design of the Phonetic Corpus”, in Proceedings of Eurospeech, 1993, volume 1, pages 653–656. Berlin (Germany), September 1993.

Some issues on the Expectation-Maximisation process for Maximum Likelihood Linear Regression

Míriam Luján-Mares, Carlos-D. Martínez-Hinarejos, Vicente Alabau, Alberto Sanchis

Institut Tecnològic d'Informàtica, Universitat Politècnica de València
Camí de Vera, s/n. 46071 València, Spain
{mlujan, cmartine, valabau, josanna}@dsic.upv.es

Abstract

The Maximum Likelihood Linear Regression (MLLR) technique has commonly been used in speaker adaptation. In the computation of the transformation matrix usually only one iteration of the Expectation-Maximisation (EM) algorithm is used, but there is not a complete study about results with a different number of iterations. We analyze how the number of iterations affects to the adaptation. The obtained results lead us to suggest a new method to accelerate the convergence of adaptation. Additionally, we propose a way to verify the contribution of the different adaptation matrices obtained in the EM process. We present experiments with the Wall Street Journal corpus whose aim is to determine the best option for the MLLR technique with respect to the number of EM iterations and the quality of the new convergence criterion.

Index Terms: speaker adaptation, speech recognition

1. Introduction

We can find Automatic Speech Recognition (ASR) systems all around the world. Current state-of-the-art ASR systems are based on Hidden Markov Models (HMM) to model the acoustic knowledge and n-grams to model the syntactical knowledge [1]. A robust ASR system needs to perform well in different environments and with different speakers. However, many speech recognition systems are for personal use, as only one speaker does usually use them (e.g., in a mobile phone or in a car). Consequently, it is interesting to guarantee an optimal performance for a particular speaker of an ASR system initially designed for multiple speakers. For this reason, speaker adaptation has become an essential part of a state-of-the-art ASR system.

An initial speaker-independent system can be adapted by using the Maximum Likelihood Linear Regression (MLLR) technique [2]. MLLR computes a set of transformations that reduces the mismatch between the initial model set and the adaptation data. These transformations are obtained by solving an optimisation problem using the Expectation-Maximisation (EM) technique [3]. The EM algorithm is used for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.

In the MLLR technique every iteration of the EM process provides a complete set of transformations that can be used to adapt the model. However, the MLLR technique has commonly been used with only one iteration of adaptation [2]. Some works study the performance of the adaptation with only the first iterations of the EM algorithm [4, 5, 6].

In this work we study the performance of the adaptation with respect to the number of iterations of the EM algorithm, from the first iteration to the convergence of the EM process. Since in each iteration the transformation matrix is closer to the identity matrix (because models are closer to adaptation data), we define a new stop criterion based on the distance of the transformation matrix to the identity matrix. Determining the contributions of the different matrices computed in the EM process can be used to verify whether the main contribution is given by the matrix of the first iteration. To determine these contributions we present a way to compute a general transformation matrix. We present experimental results on 8 speakers from the Wall Street Journal corpus to study this influence.

2. The MLLR adaptation technique

The aim of speaker adaptation techniques is to obtain a speaker-dependent recognition system by using a combination of general speech knowledge from well-trained HMM and speaker-specific information from a new speaker's data.

MLLR is a technique to adapt a set of speaker-independent acoustic models to a speaker by using small amounts of adaptation material. The MLLR approach requires an initial independent continuous density HMM system. MLLR adapts the acoustic models and updates the model mean parameters to maximise the likelihood of the adaptation data by using a transformation matrix, which is estimated from the adaptation data.

The theory is based on the concept of regression classes. A regression class is a set of mixture components that share the same transformation matrix $\vec{W}$. This matrix is applied to the extended mean vector of all the mixtures pertaining to the regression class to obtain an adapted mean vector. Given a state q in a HMM, for the ith gaussian of its output distribution, we denote its mean vector as $\vec{\mu}_{qi}$. The adapted mean vector $\widehat{\vec{\mu}}_{qi}$ is obtained by:

$$\widehat{\vec{\mu}}_{qi} = \vec{W} \cdot \vec{\xi}_{qi}$$

where $\widehat{\vec{\mu}}_{qi}$ is the adapted mean and $\vec{\xi}_{qi}$ is the extended mean vector defined as $\vec{\xi}_{qi} = [w, \mu^1_{qi}, \ldots, \mu^n_{qi}]' = [w : \vec{\mu}_{qi}]$, where n is the number of features, $\vec{\mu}_{qi}$ is the original mean vector and w is an offset term.

If we have a set of adaptation data denoted by the sequence of acoustic feature vectors $\vec{X} = \vec{x}_1\vec{x}_2 ... \vec{x}_T, \vec{x}_t \in \mathbb{R}^D, t = 1, ..., T$, we can estimate the transformation matrix $\widehat{\vec{W}}$ using the maximum likelihood approach as:

$$\widehat{\vec{W}} = \max_{\vec{W}} p_{\widehat{\lambda}}(\vec{X})$$

where $\widehat{\lambda}$ defines the parameters of the adapted model.

The problem is solved by using the EM algorithm [3]. EM is an iterative method which alternates between performing:

1. An expectation (E) step: which computes an expectation of the log likelihood with respect to the current estimate of the distribution for the latent variables.
2. A maximization (M) step: which computes the parameters which maximise the expected log likelihood found on the E step. These parameters are used to determine the distribution of the latent variables in the next E step.

The main idea is to define an auxiliary function $Q(\vec{\lambda}, \widehat{\vec{\lambda}})$ in the step E as: $Q(\vec{\lambda}, \widehat{\vec{\lambda}}) = \sum_{\vec{\theta} \in \vec{\Theta}} p_{\vec{\lambda}}(\vec{X}, \vec{\theta}) \cdot \log(p_{\widehat{\vec{\lambda}}}(\vec{X}, \vec{\theta}))$ where $\vec{\theta}$ is a state sequence and $\vec{\Theta}$ is the set of all possible state sequences with length T. The auxiliary function depends on both the initial model $\vec{\lambda}$ and the adapted model $\widehat{\vec{\lambda}}$. Using this definition it can be shown that by successively defining a new model $\widehat{\vec{\lambda}}$ which maximises Q (in the step M), the auxiliary function has the property that the value of $p_{\widehat{\vec{\lambda}}}(\vec{X})$ will not decrease, which was the original objective. The auxiliary function is maximised in the standard way by differentiating and equating to zero. The solution of this equation is the set of transformation matrices $\vec{W}$, that are applied on $\vec{\lambda}$ to obtain $\widehat{\vec{\lambda}}$. Therefore, it is necessary to compute a transformation matrix for every iteration of the EM algorithm until the algorithm converges. Our convergence criterion is based on the difference between the values of Q in the current and the previous iteration. We assume convergence when this difference is 0 (for the used numeric precision).

To compute the transformation matrix (M step), we can use several variants. Details on the estimation of these variants can be consulted in [2]. In our case, we suppose different covariances for each distribution and full adaptation matrices.

3. MLLR issues

One important problem in MLLR is the large number of iterations required for the EM convergence with a full transformation matrix. We know that in each iteration the transformation matrix is closer to the identity matrix because models are closer to adaptation data. If a transformation matrix is not closer to the identity matrix that the matrix of the previous iteration it is possible that models are overfitted. With this idea, we propose a new method to stop the EM estimation: in each iteration, we calculate the distance between the transformation matrix and the identity matrix using the Euclidean Distance Matrix [7]:

$$\delta(\vec{W}, \vec{I}) = ||\vec{W} - \vec{I}||$$

When this distance is higher than the distance obtained with the matrix of the previous iteration we stop the adaptation. Since the adaptation of the models depends directly on the transformation matrix, this seems a good criterion to evaluate when is worth applying the transformation matrix. Therefore, we use this idea as a new method to stop the estimation. With this option, we reduce the computing time to obtain good adapted models.

Since we want to study the performance of the adaptation with respect to the number of iterations of the EM algorithm, we need to determine the contribution of all adaptation matrices obtained in the EM process. We can define that the transformations define a path in the representation space between the initial and final models, and this path covers a certain distance. Therefore, to determine the contribution of each EM step, the idea is to determine how much distance is covered in each iteration. To determine the contribution of a sequence of EM steps we present a way to compute a general transformation matrix that reflects the whole effect of the EM steps (i.e., we calculate a general matrix that applied to the initial models allows to obtain the adapted model in any iteration).

The calculation of a general matrix was not defined previously, as far as we know. This matrix can not be calculated as the product of the different matrices obtained in the EM process (because dimensions do not match). We have obtained a way to compute a general matrix to obtain the adapted models from the initial models. The method solves a linear equation system where the unknown variables are the matrix coefficients. These coefficients are calculated using only n means (where n is the number of rows of $\vec{W}$) to be transformed from the initial model to the adapted model. For example, for $n = 2$, if we have the original means $\vec{\varepsilon_1} = [w\ \varepsilon_{11}\ \varepsilon_{12}]$, $\vec{\varepsilon_2} = [w\ \varepsilon_{21}\ \varepsilon_{22}]$ and the corresponding adapted means $\vec{\mu_1} = [\mu_{11}\ \mu_{12}]$, $\vec{\mu_2} = [\mu_{21}\ \mu_{22}]$, the coefficients of $\vec{W} = [w_{10}\ w_{11}\ w_{12}; w_{20}\ w_{21}\ w_{22}]$ are obtained by solving:

$$\begin{bmatrix} w & \varepsilon_{11} & \varepsilon_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & w & \varepsilon_{11} & \varepsilon_{12} \\ w & \varepsilon_{21} & \varepsilon_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & w & \varepsilon_{21} & \varepsilon_{22} \end{bmatrix} \begin{bmatrix} w_{10} \\ w_{11} \\ w_{12} \\ w_{20} \\ w_{21} \\ w_{22} \end{bmatrix} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix}$$

where w_{ij} are the unknown variables.

Using the definition of the general matrix, we compute the distance covered by the whole transformation process as the distance between the general matrix of the models obtained in the convergence ($\vec{W}_c$) and the identity matrix. The contribution of each step is calculated as the projection on this path, using the distance between the general matrix of step i ($\vec{W}_i$) and $\vec{W}_c$ and $\vec{I}$ to define the projection.

Furthermore, with the use of the general matrix it is not necessary to keep the models of each iteration (since these models can be obtained by applying the general transformation matrix on the original models), with the consequent space saving.

4. Corpus

The experiments were performed using the Wall Street Journal (WSJ) database [8]. The ARPA WSJ corpus consists of samples of read texts drawn from WSJ publications recorded under high-quality conditions. We have used the Nov'92 (WSJ0) and Nov'93 (WSJ1) training data.

The initial HMMs have been trained with HTK [9] using just the WSJ0 training database composed of 84 speakers with a duration of 15 hours. The HMMs are word-internal triphones and gender independent. They are composed of $2.3k$ tied-states and the topology is left-to-right with loops. The number of gaussians per state is 16. We have also trained silence and interword silence models.

For adaptation experiments, we have used a set of eight speakers selected from the WSJ1 (Nov'93) training material [1]. We have only one regression class (a global transformation matrix). There are about 150 utterances available for each speaker. 50 sentences were used for adaptation and the remaining were used for testing. As language model, we used the standard $20k$ trigram grammar.

5. Experiments and Results

To analyze the results, we used the Word Error Rate (WER) as the evaluation measure. This measure computes the edit dis-

[1]Notice that this subcorpus is not the usual WSJ benchmark and baseline results might not be comparable with other works.

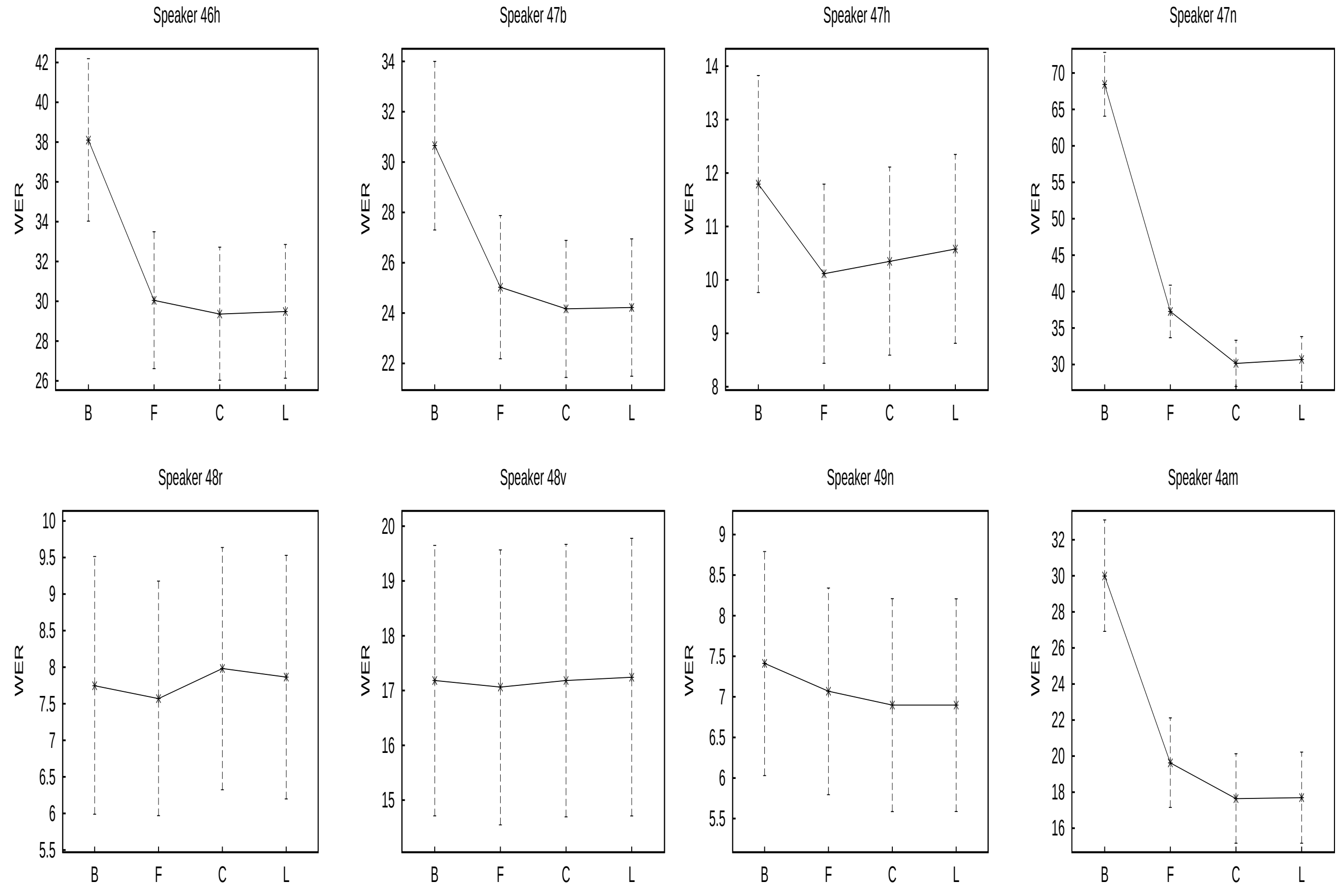


Figure 1: Comparison among full adaptation matrices with a different number of EM iterations for each speaker. B is baseline result. F is the result for the first iteration. C is the result for the matrix convergence. L is the result for the last iteration.

tance between a reference sentece and the recognized sentence.

We performed some experiments with the MLLR technique to determine its behavior with respect to the number of EM iterations. These experiments included a comparison between the recognition results obtained when applying the different transformations obtained in each EM iteration (from first iteration till convergence and the new convergence criterion). Recognition was carried out with the iATROS recogniser [10].

In Figure 1 we can see the results for each speaker with confidence intervals that show whether the differences among the results are statistically significant [11]. Every graphic shows four results:

- **Baseline**: It is the WER obtained when using the models without adaptation.
- **First iteration**: It is the WER obtained when using an adapted model with only one iteration of the EM algorithm.
- **Matrix convergence**: It is the WER obtained with an adapted model with the matrix convergence that we defined with the distance between the transformation matrix and the identity matrix.
- **Last iteration**: It is the WER obtained with a model adapted with the transformations obtained in the EM convergence.

According to these results, we can distinguish two groups of speakers:

- 46h, 47b, 47n, 4am: these speakers have a bad WER (above 30) when they use models without adaptation. These speakers have a better statistically significant WER when they use adapted models (47b has a statistically significant WER only for the matrix convergence or last iteration).
- 47h, 48r, 48v, 49n: these speakers have a better WER (below 20) when they use models without adaptation. In these speakers it is not necessary to adapt the original models because the adaptation does not improve the recognition results (the differences among the results are not statistically significant). With these speakers we can not draw conclusions.

In Figure 2 we show the results that we obtain if we calculate the mean of all speakers. The results are better than baseline because confidence intervals show that the differences among the results are statistically significant. Although the results when using EM till convergence (with matrix or EM convergence criteria) are not significantly improved with respect to the results of the first iteration, absolute results are slightly improved and show that it could be convenient to use more than one EM iteration. The matrix convergence is a good option because the number of iterations is quite lower than in the EM convergence (for example, the speaker 46h has 606 iterations with the EM convergence and 19 iterations with the matrix convergence) but results are very similar and overfitting is possibly avoided.

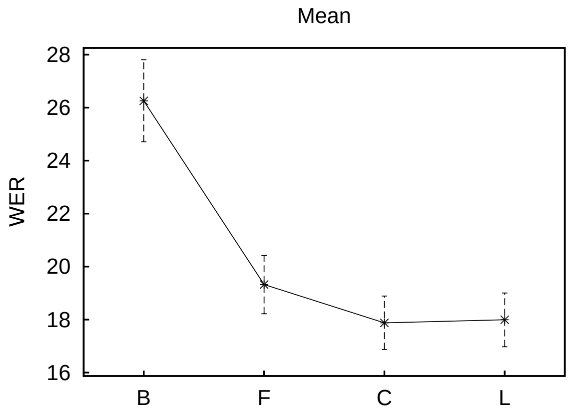


Figure 2: Mean WER for all the 8 speakers. B, F, C, and L have the same meaning than in Figure 1.

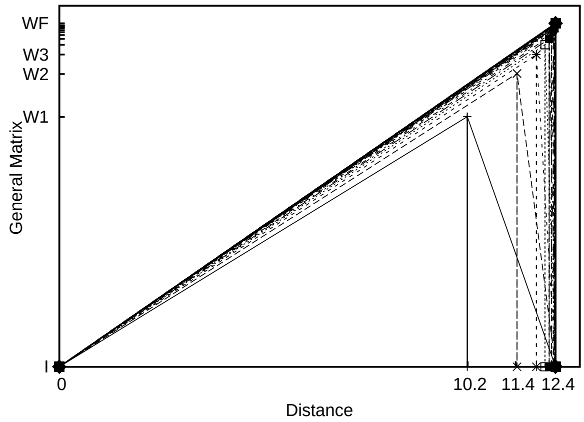


Figure 3: Distance and projections of each general adaptation matrix with respect to the identity matrix and the convergence adaptation matrix for speaker 46h.

With respect to the contributions of each EM step to the adaptation, in Figure 3 we show the calculated projections for the general matrices obtained in each EM step for speaker 46h. We can see that the first distance is the highest distance of all distances. Therefore, we think that the matrix of the first iteration is the matrix that produces a more important transformation in the data and it is the matrix with a greater contribution in the adaptation process. Since the distance decreases in every iteration, the contribution is lower and the first iterations are the most important in the adaptation process.

6. Conclusions and Future Work

The results show that MLLR speaker adaptation significantly improves the performance in speakers with bad performance with speaker independent acoustic models. The results show that there is no significant improvement between using only one EM iteration or more iterations in the recognition performance. We presented a new EM convergence criterion that obtains adapted models with similar performance to those obtained with the usual EM stop criterion and that are possibly not overfitted. Moreover, the number of iterations of EM is lower when using this new criterion. Therefore, we reduced the time of computation.

We studied the contribution of the adaptation matrices to confirm that the transformation matrix of the first iteration is that which produces the major contribution. We demonstrated it with an empirical method, using the recognition results and the distance between matrices. To perform this analisys we provided a method to calculate a general matrix that computes the adapted models from initial models. Furthermore, with the general transformation matrix we reduce the memory to apply adaptation, since only the initial models and a transformation matrix are needed to obtain the adapted models.

Future work is directed towards using more regression classes and automatic building of regression classes. Moreover, other convergence criteria can be defined and experiments can be performed with other corpora to confirm these conclusions. The use of MLLR on handwritten text recognition in another interesting work.

7. Acknowledgements

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018) and MITTRAL (TIN2009-14633-C03-01) projects. Also supported by the EC (FEDER) and the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project, by the Generalitat Valenciana under grants Prometeo/2009/014, GV/2010/067, ACOMP/2010/051 and by VIDI-UPV under PAID06-20070315 program.

8. References

[1] L. Rabiner and B. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall PTR.

[2] C. J. Leggetter and P. C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9:171–185.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

[4] C. J. Leggetter. 1995. Improved Acoustic Modeling for HMMs using Linear Transformations. *PhD thesis, Cambridge University Engineering Department*.

[5] J. Kleban and Y. Gong 2000. HMM adaptation and microphone array processing for distant speech recognition. In *Proceedings of the 2000 IEEE international Conference on Acoustics, Speech and Signal Processing*, ICASSP. IEEE Computer Society, vol.3.

[6] T. Shinozaki, Y. Kubota and S. Furui. 2009. Unsupervisec cross-validation adaptation algorithms for improved adaptation performance. In *Proceedings of the 2009 IEEE international Conference on Acoustics, Speech and Signal Processing*, ICASSP. IEEE Computer Society, Washington, DC, 4377-4380.

[7] J. Dattorro. 2005. Convex Optimization and Euclidean Distance Geometry. *Meboo Publishing*.

[8] D. S. Pallett, J. G. Fiscus, W. M. Fisher, John S. Garofolo, B. A. Lund, and M. A. Przybocki. 1994. 1993 benchmark tests for the arpa spoken language program. In *HLT '94: Proceedings of the workshop on HLT*, pages 49–74, Morristown, NJ, USA.

[9] Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, July, 2004. *The HTK Book*. CUED, UK, v3.2 edition.

[10] M. Luján-Mares, V. Tamarit, V. Alabau, C. D. Martínez-Hinarejos, M. Pastor, A. Sanchis, and A. Toselli. 2008. iatros: A speech and handwritting recognition system. In *V Jornadas en Tecnologías del Habla (VJTH'2008)*, pages 75–78, Spain.

[11] M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of ICASSP'04*, volume 1, pages 409–412.

Speaker Verification Performance Degradation against Spoofing and Tampering Attacks

Jesús Villalba, Eduardo Lleida

Communications Technology Group (GTC),
Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba,lleida}@unizar.es

Abstract

In this paper, we evaluate the performance of current state of the art speaker verification (SV) systems against some examples of spoofing and tampering attacks. We understand as spoofing the fact of impersonating another person using, for instance, a recording of his voice. On the contrary, we call tampering to the alteration of somebody's voice in order to prevent being detected by a SV system. These techniques can produce important performance degradations. We show that, for the EER operating point, spoofing can produce false aceptance rates of 68% and tampering misses rates of 50%. This is critical in some security applications which makes necessary to develop methods to detect manipulated speech signals.

Index Terms: speaker verification, forgery, disguise, spoofing, tampering, JFA.

1. Introduction

Current state of the art speaker verification systems (SV) have achieved great performance due, mainly, to the appearance of the GMM-UBM [1] and Joint Factor Analysis (JFA) [2] approaches. However, this performance is usually measured in conditions where impostors do not do any effort to disguise their voices to be similar to any true target speaker and where a true target speaker does not try to modify his voice to hide his identity. That is what happens in NIST evaluations [3]. Therefore, the purpose of this paper is to evaluate SV on this kind of adverse situations.

We have classified the possible attacks to SV as spoofing and tampering. Spoofing is the fact of impersonating another person using different techniques like voice transformation or playing of a recording of the victim. On the other side, tampering is the alteration of somebody's voice to prevent being detected by SV. There are multiple techniques for voice disguise, in [4] authors do a study of voice disguise methods and classify them into electronic transformation or conversion, imitation, and mechanical and prosodic alteration. In [5] an impostor voice is transformed into the target speaker voice using a voice encoder and decoder. More recently, in [6] an HMM based speech synthesizer with models adapted from the target speaker is used to deceive a SV system. In [7] the effects of speaking while grasping a pencil in the teeth are studied. In this work, we focus on low technology techniques like replay attack or putting a handkerchief over the mouth.

This paper is organized as follows. Section 2 explains the spoofing and tampering methods that have been studied in this work. Section 3 describes the experiments and databases, used to measure the perfomance degradation due to these attacks, and shows the results we have got. Finally, in section 4 we show some conclusions.

2. Spoofing and tampering methods

2.1. Replay attack spoofing

A replay attack consists of an impostor that try to impersonate another person using a recording of his voice. The impersonator could get this recording by several means. One of them would be surreptitiously doing a far field recording of the victim using the microphone of a smartphone or a laptop. Another option could be even getting it from the internet if the victim is a public person. Then, the impostor just needs to replay the sentence using a loudspeaker. This is one of the main weaknesses of SV. Especially text independent systems, that accept that the users say whatever they want. It has mayor importance for applications such as telephone access to bank accounts or admission to restricted areas in a high security facility.

2.2. Cut and paste spoofing

The usual approach of commercial SV systems to prevent replay attacks is the use of text dependent systems. The user, that wants to be authenticated, is asked to utter a given sentence that is different for every access attempt. In this case, the SV checks both, the speaker identity and whether the uttered sentence is correct. In this manner, the robustness of the system against attacks is highly increased given that it is unlikely that the asked sentence is among the ones previously recorded by the impostor.

However, this method is not unfailing. If the impostor would have access to fair amount of data from the legitimate user he could be able to build the requested sentence using pieces of different recordings. This is what is call cut and paste spoofing attack. Currently, anybody without any particular expertise can do this with the audio editing programs available in the market.

2.3. Handkerchief tampering

What we have called handkerchief tampering consists of covering the speaker's mouth with a handkerchief together with the hand between the mouth and the microphone making a shell. This implies an important distortion on the spectral distribution of the speech signal. Current, state of the art SV use mainly spectral based features (MFCC, PLP, LPCC) so the performance of those systems can be greatly affected. This technique can be applied to cheat the systems of law enforcement agencies that search for criminals into phonecalls.

2.4. Nasalization tampering

This kind of tampering consists of obstructing the nostrils while the user is speaking. In this way, the sound wave is reflected back along the nasal cavity interfering with the wave in the pharynx. At certain frequencies both waves cancel each other introducing anti-resonances in the vocal tract transfer function. Like in the previous case, this can modify the spectrum of the signal in such a way that a person could not be detected.

3. Experiments

3.1. Speaker verification system

We have used a SV system based on JFA [2] to measure the performance degradation. Feature vectors of 20 MFCC (C0-C19) plus first and second derivatives are extracted. After frame selection, features are short time Gaussianized as in [8]. A gender independent Universal Background Model (UBM) of 2048 Gaussians is trained by EM iterations. Then 300 eigenvoices v and 100 eigenchannels u are trained by EM ML+MD iterations. Speakers are enrolled using MAP estimates of their speaker factors (y,z) so the speaker means super vector is given by $M_s = m_{UBM} + vy + dz$. Trial scoring is performed using first order Taylor approximation of the LLR between the target and the UBM Models like in [9]. Scores are ZT Normalized and calibrated to log-likelihood ratios by linear logistic regression using FoCal package [10] and the SRE08 trial lists. We have used telephone data from SRE04, SRE05 and SRE06 for UBM and JFA training, and score normalization.

3.2. replay attack spoofing

3.2.1. Database

We have used a database consisting of 5 speakers. Each speaker have 4 groups of signals:

- Originals: Recorded by a close talk microphone and transmitted by telephone channel. There are 1 train signal and 7 test signals. They are transmitted through different telephone channels: digital (1 train and 3 test signals), analog wired (2 test signals) and analog wireless (2 test signals).
- Microphone: Recorded simultaneously with the originals by a far field microphone.
- Analog Spoof: The microphone test signals are used to do a replay attack on a telephone handset and transmitted by an analog channel.
- Digital Spoof: The microphone test signals with replay attack and transmitted by a digital channel.

We have used these signals to create 35 legitimate target trials, 140 non spoof non target, 35 analog spoofs and 35 digital spoofs. The training signals are 60 seconds long and the test signals 5 seconds approximately.

3.2.2. Results

We have got an EER=0.71% using the non spoofing trials only. In Figure 1 we show the score distribution of each trial dataset. There is an important overlap between the target and the spoof dataset. If we would choose the EER operating point as decision threshold we would accept 68% of the spoofing trials. Table 1 presents the score degradation statistics between a legitimate utterance and the same utterance after the spoofing processing (far field recording, replay attack). The average degradation is only around 30%. However, it has a big dispersion with some spoofing utterances getting a higher score than the original ones.

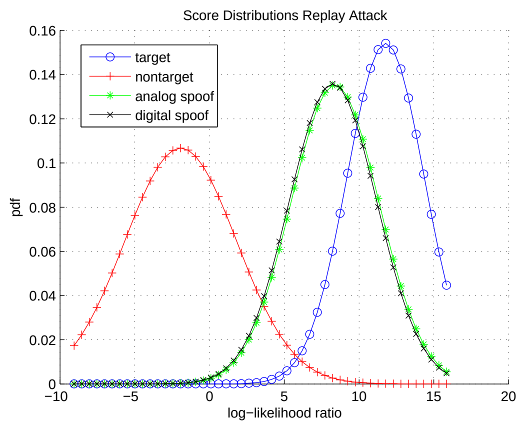


Figure 1: *Score distributions of the replay attack database*

Table 1: *Score degradation due to replay attack*

		Mean	Std	Median	Max	Min
Analog	Δscr	3.38	2.42	3.47	9.70	-1.26
	Δscr/scr (%)	29.00	19.37	28.22	70.43	-10.38
Digital	Δscr	3.52	2.30	3.37	9.87	-1.68
	Δscr/scr (%)	30.29	18.92	29.52	77.06	-16.74

3.3. Cut and Paste spoofing

3.3.1. Database

The cut and paste database consists of three phases:

- Phase 1+Phase2: it has 20 speakers. It includes landline (T) signals for training, non spoof tests and spoofs tests; and GSM (G) for spoofs tests.
- Phase 3: it has 10 speakers. It includes landline and GSM signals for all training and testing sets.

Each phase has three sessions:

- Session 1: it is used for enrolling the speakers into the system. Each speaker has 3 utterances by channel type of 2 different sentences (F1,F2). Each sentence is around 2 seconds long.
- Session 2: it is used for testing non spoofing access trials and has 3 recordings by channel type of each of the F1 and F2 sentences.
- Session 3: it is made of different sentences and a long text that contain words from the sentences F1 and F2. It has been recorded by a far field microphone. From this session several segments are extracted and used to build 6 sentences F1 and F2 that will be used for spoofing trials. After that, the signals are played on a telephone handset and transmitted through a landline or GSM channel. In this manner, these utterances include cut and paste and replay attack processing.

3.3.2. Results

We have done separate experiments using phase1+2 and phase3 datasets. For phase1+2, we train speaker models using 6 landline utterances, and do 120 legitimate target trials, 2280 non

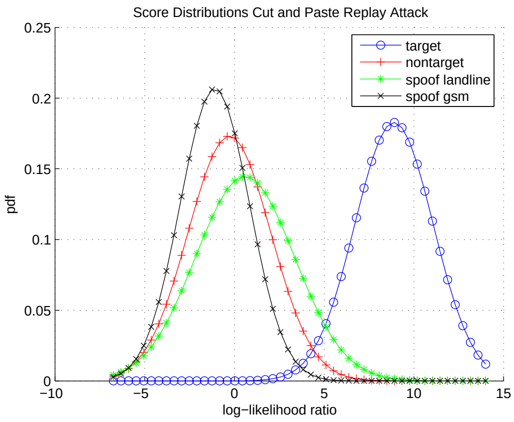


Figure 2: *Score distributions of cut+paste phase 1+2*

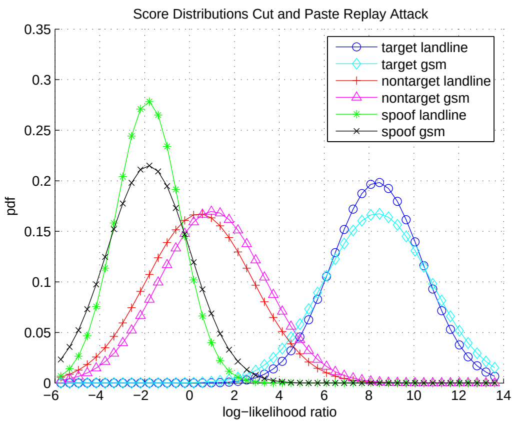


Figure 3: *Score distributions of cut+paste phase 3*

spoof non target, 80 landline spoofs and 80 GSM spoofs. For phase 3, we train speaker models using 12 utterance (6 landline + 6 GSM), and do 120 legitimate target trials (60 landline + 60 GSM), 1080 non spoof non target (540 landline + 540 GSM) and 80 spoofs (40 landline + 40 GSM). Using non spoof trials we have got and EER=1.66% and EER=5.74% for phase1+2 and phase3 respectively. Figures 2 and 3 show the score distributions for each of the databases. Table 2 shows the score degradations statistics due to the spoofing processing. The degradation is calculated by speaker and sentence type, that is, we calculate the difference between the average score of the clean sentence Fx of a given speaker and the average score of the spoofing sentences Fx of the same speaker. We can appreciate that the degradation is more strong in this case than in the database with replay attack only. Even for phase 3, the spoofing scores are lower than the non target scores. This means, the processing used for creating the spoofs can modify the channel conditions in a way that makes the spoofing useless. We think that this is affected too by the length of the utterances. It is known that when the utterances are very short Joint Factor Analysis cannot do proper channel compensation. If the channel component were well estimated the spoofing scores should be higher.

3.4. Handkerchief tampering

3.4.1. Database

This database consists of 10 speakers with 3 sessions:

Table 2: *Score degradation due to cut+paste replay attack*

			Mean	Std	Median	Max	Min
Phase1+2	T	Δscr	8.29	3.87	7.96	17.89	1.41
		Δscr/scr (%)	90.53	31.64	90.72	144.88	27.46
	G	Δscr	9.98	2.96	9.56	18.517535	5.40
		Δscr/scr (%)	111.94	18.03	109.437717	159.69	80.41
Phase3	T	Δscr	10.21	2.51	9.76	17.78	6.86
		Δscr/scr (%)	123.06	18.47	117.54	180.38	95.60
	G	Δscr	10.21	3.32	10.19	18.36	4.65
		Δscr/scr (%)	121.63	19.50	119.39	167.15	92.67

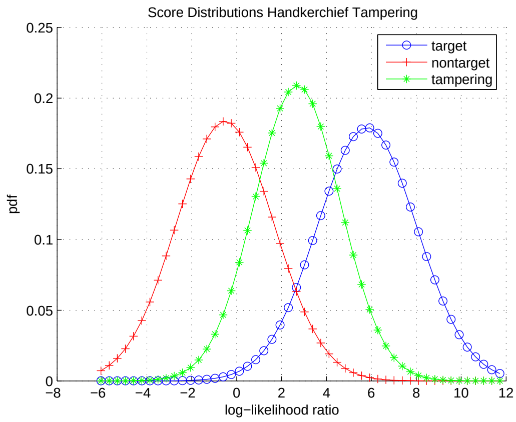


Figure 4: *Score distributions of the handkerchief tampering database*

- Session 1: training speaker models. Around 12 seconds of speech by speaker.
- Session 2: clean test signals. They are 120 short segments of around 3 seconds length.
- Session 3: tampering test signals. Another 120 short segments repeating the sentences of session 2.

With this database, we can do 120 target trials, 120 tampering trials and 1080 non target trials.

3.4.2. Results

We have got an EER=6.66% using non tampering trials only. Figures 4 and 5 show the score distributions of each of the trial subsets and the P_{miss} and P_{fa} versus the decision threshold. These figures evidence the great loss of performance that tampering can produce. For the EER operating point, we would reject 50% of true speakers with tampering. Table 3 presents the score degradation statistics between a clean sentence and itself with tampering.

Table 3: *Score degradation due to handkerchief tampering*

	Mean	Std	Median	Max	Min
Δscr	3.10	2.20	2.90	10.37	-0.88
Δscr/scr (%)	52.80	32.80	56.05	120.19	-31.25

3.5. Nasalization tampering

3.5.1. Database

The database consists of 52 speakers. It includes read and spontaneous speech recorded over a GSM channel. Speech segments can have 60, 90 or 120 seconds. Clean segments of 120 seconds have been used for speaker enrollment and the rest for testing. We have 198 clean targets trials, 165 tampering trials and 10098 non target trials.

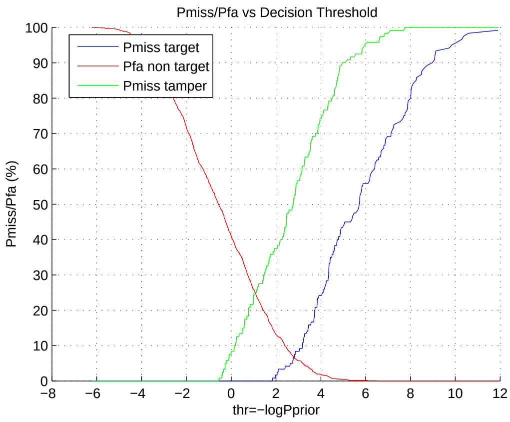


Figure 5: *Pmiss/Pfa vs. decision threshold for the handkerchief tampering database*

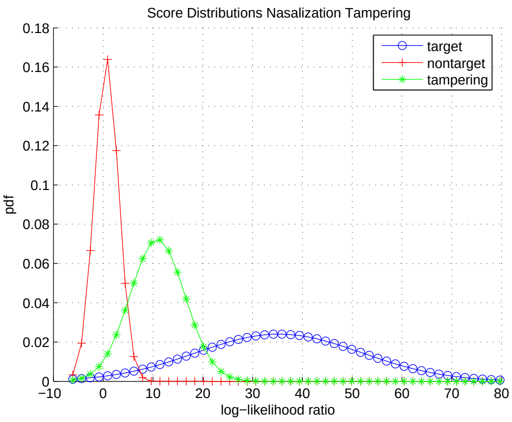


Figure 6: *Score distributions of the nasalization tampering database*

3.5.2. Results

For this database, we have got an EER=4.54% using clean trials only. Figures 6 and 7 show the score distributions of each of the trial subsets and the P_{miss} and P_{fa} versus the decision threshold. Table 4 presents the score degradation statistics due to the tampering. In this case the tampering and clean sentences are different, so we calculate the degradation as the difference between the average score of the clean recordings of a given speaker and the average score of his tampering recordings. The score degradation is quite big, however the error rates seem less affected having 10% of rejection for the EER operating point. Perhaps, this is due to the bigger length of the utterances that allows better intersession compensation.

Table 4: *Score degradation due to nasalization*

	Mean	Std	Median	Max	Min
Δscr	27.03	11.65	28.27	51.72	4.51
Δscr/scr (%)	68.69	13.63	69.21	96.85	11.13

4. Conclusions

In this paper, we have evidenced the vulnerability of state of the art SV systems to several kinds of spoofing and tampering attacks. For this purpose, we have used databases specifically created to evaluate each type of attack. We have seen that spoofing trials, although having lower scores than the legitimate ones, can produce score distributions high enough to get big acceptance rates. This can be a serious threat for security applications such as authenticating a remote client to give him

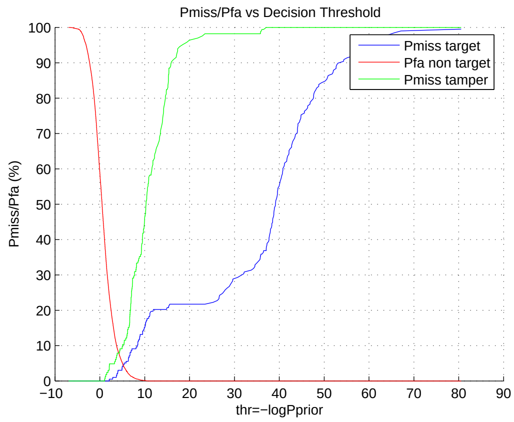


Figure 7: *Pmiss/Pfa vs. decision threshold for the nasalization tampering database*

access to a bank account. On the other side, tampering attacks like nasalization or using a handkerchief to modify your voice can produce low verification scores. This means that applications such as search of criminals in telephone recordings by law enforcement agencies could be easily overcome.

In order to increase speaker verification systems robustness to this kind of attacks, methods to detect when a signal has been manipulated should be investigated. In the future, we will drive our efforts to this task. However, detecting all kind of manipulations that can be done to a signal can be complicated.

5. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[3] "http://www.itl.nist.gov/iad/mig/tests/sre/2010/ NIST_SRE10_evalplan.r6.pdf."

[4] P. Perrot, G. Aversano, and G. Chollet, "Voice disguise and automatic detection:review and perspectives," *Lecture Notes In Computer Science*, pp. 101–117, 2007.

[5] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice Forgery Using ALISP: Indexation in a Client Memory," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE, pp. 17–20.

[6] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[7] R. M. D. Figueiredo and H. D. S. Britto, "A report on the acoustic effects of one type of disguise," *Forensic Linguistics*, vol. 3, no. 1, pp. 168–175, 1996.

[8] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Oddyssey Speaker and Language Recognition Workshop*, Crete, Greece, 2001.

[9] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," in *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 4057–4060.

[10] N. Brummer, "http://sites.google.com/site/nikobrummer/ focalbilinear."

A speaker recognition system based on sufficient-statistics-space channel-compensation and dot-scoring

Mikel Penagarikano, Amparo Varona, Mireia Diez, Luis Javier Rodriguez-Fuentes, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
mikel.penagarikano@ehu.es

Abstract

This paper briefly describes the dot-scoring speaker recognition system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the NIST 2010 Speaker Recognition Evaluation. The system does eigenchannel compensation in the sufficient statistics space and scoring is performed by a simple dot product. An optimized Matlab implementation of of the eigenchannels estimation, the channel compensation and the normalized mean vector computation is provided.

Index Terms: Speaker Recognition, NIST SRE, Dot Scoring, Sufficient Statistics, Eigenchannel Compensation, Matlab

1. Introduction

This paper briefly describes the dot-scoring speaker recognition system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the NIST 2010 Speaker Recognition Evaluation (SRE). This system was built following the SUNSDV system description for SRE08 [1]. The system combines two key technologies: sufficient statistics space eigenchannel compensation and dot scoring.

The rest of the paper is organized as follows. Sufficient statistics equations are described in Section 2. Eigenchannel compensation is discussed in Section 3. The linear scoring technique is introduced in section 4. The experimental setup is outlined in Section 5, including details about the partitioning of previous SRE databases, feature extraction (front-end) and configuration of the eigenchannel computation. Section 6 presents the results of the dot-scoring system in the SRE2010 evaluation. Finally, conclusions are summarized in Section 7.

2. Sufficient statistics

Let $\mathcal{N}\left(\omega, \mu^{ubm}, \Sigma\right)$ be a Gaussian Mixture Model (GMM) representing the Universal Background Model (UBM), consisting of K mixture components of dimension F and diagonal covariance matrix Σ. Let $f(t)$ be the feature vector at time t. Let $\gamma_k(t)$ be the posterior probability of mixture k at time t. Let Σ_k be the covariance matrix of mixture k. Let $repmat(M,i,j)$ be the function that replicates the matrix M $i \times j$ times, and let $vec(M)$ be the function that concatenates all the columns of matrix M in a single vector. We define:

$$n_k = \sum_t \gamma_k(t) \tag{1}$$

$$n = vec\left(repmat\left([n_1\, n_2 \dots n_K], F, 1\right)\right) \tag{2}$$

$$x_k = \sum_t \gamma_k(t)\, \Sigma_k^{-\frac{1}{2}} \left(f(t) - \mu_k^{ubm}\right) \tag{3}$$

$$x = vec\left([x_1, x_2, \dots, x_K]^t\right) \tag{4}$$

Vectors n and x (of size $F \times K$) are the so called zero-order and first-order sufficient statistics, respectively. Once the sufficient statistics are obtained, there is no need to use the UBM again, and therefore all the code is independent of the UBM. For example, the popular one-iteration relevance-MAP adapted and normalized mean vector $m = \frac{\mu_{map} - \mu_{UBM}}{\sigma}$ is obtained by:

$$m = \left(\tau \mathbf{I} + diag(n)\right)^{-1} \cdot x \tag{5}$$

where τ is the relevance factor and $diag(v)$ is a function that returns a diagonal matrix with values from vector v on the diagonal.

3. Eigenchannel compensation

Channel compensation in the space of sufficient statistics is performed using the eigenchannel recipe developed by the Brno University of Technology Speech Group [2]. The first order sufficient statistics are compensated as follows:

$$\hat{x} = x - diag(n) \cdot W L^{-1} W^t x \tag{6}$$

where W is the so called eigenchannel matrix, and matrix L is given by:

$$L = \mathbf{I} + W^t diag(n)\, W = \mathbf{I} + \sum_{k=1}^{K} n_k O_k \tag{7}$$

where $O_k = W_k^t W_k$. Channel compensation on the adapted and normalized mean vector m is performed using equation 5:

$$\hat{m} = \left(\tau \mathbf{I} + diag(n)\right)^{-1} \cdot \hat{x} \tag{8}$$

3.1. Estimation of the eigenchannel matrix

Given a data matrix $M = [m_1, \dots, m_J]$ composed by adapted and normalized mean vectors, the eigenchannel matrix W consists of the D most significant eigenvectors of the data covariance, each eigenvector v weighted by the square roots of the corresponding eigenvalues λ:

$$W = \left[v_1 \cdot \sqrt{\lambda_1} \quad v_2 \cdot \sqrt{\lambda_2} \ \dots \ v_D \cdot \sqrt{\lambda_D}\right] \tag{9}$$

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

Since the matrix M includes, among other sources of variability, speaker variability (which we would not like to compensate), the average speaker model must be subtracted from all sessions of a speaker prior to eigenchannel computation.

It is possible to obtain different eigenvectors from different types of channel variabilities (for example telephone-telephone, microphone-microphone and telephone-microphone variabilities), and then stack all of them in a single matrix. That is:

$$W = [W_{tel} \quad W_{mic} \quad W_{tel-mic}] \tag{10}$$

Note that the data covariance matrix $1/J MM^t$ has dimenssions $FK \times FK$ (F being around 40 and K around 1024), being unfeasible the direct computation of the eigenvectors. A possible solution is to compute the eigenvectors V of matrix $1/J M^t M$ (sized $J \times J$) and then project them by $W = MV$.

3.2. Matlab implementation of eigenchannel compensation

Some comments must be done about the Matlab implementation. Once the eigenchannel matrix W has been obtained, the compensated first order statistics, $\hat{x}$, can be computed from $(\boldsymbol{n}, \boldsymbol{x})$, the non-compensated zero and first order statistics, but note that the calculation of matrix L can be accelerated if products $O_k = \left\{W_k^t \cdot W_k\right\}$ are precomputed. Note also that being L positive definite, $L^{-1}\left(W^t x\right)$ can be solved by Cholesky decomposition. Finally, note that we are only interested in the D largest eigenvectors, which can be efficiently found using the Matlab `eigs` function (instead of the full `eig` version).

Listings 1, 2 and 3 show the Matlab implementations of the eigenchannels estimation, the channel compensation and the normalized mean vector computation, respectively.

4. Linear Scoring

Linear scoring (dot-scoring) makes use of a linearized procedure to score test segments against target models [1]. Given a feature stream f (the target signal) and a speaker spk, the first-order Taylor-series approximation to the GMM log-likelihood is:

$$\log P\left(f|spk\right) \approx \log P\left(f|UBM\right) + m_{spk}^t \cdot \nabla P(f|UBM) \tag{11}$$

where m_{spk} denotes the normalized mean vector of speaker spk and ∇ denotes the gradient vector w.r.t the standard-deviation-normalized means of the UBM, and $\nabla P(f|UBM) = x_f$ is the first order statistics vector of target signal f. The log-likelihood-ratio between the target model and the UBM is used for scoring, as follows:

$$score\left(f, spk\right) = \log \frac{P\left(f|spk\right)}{P\left(f|UBM\right)} \approx m_{spk}^t \cdot x_f \tag{12}$$

When channel compensation is applied, both the normalized mean vector of the speaker and the first order statistics vector of the target signal are compensated:

$$\hat{score}\left(f, spk\right) = \hat{m}_{spk}^t \cdot \hat{x_f} \tag{13}$$

The linear scoring is a very fast and effective method tha has proved to be comparable to (and sometimes even better than) Support Vector Machines (SVM) based scoring methods. Indeed, SVMs require much more computation and an extra set of impostor models.

Listing 1: Eigenchannel estimation function code in Matlab. The functions parametres are M, the data matrix containing normalized mean vectors as rows, K, the dimenssion os the feature vectors, c, a vector containing numeric speaker labels (identities) for each mean vector and D, the desired number of eigenchannels. Output values are the eigenchanel matrix W and the cell array $O = \{O_k\}$.

```
function [W,O] = EigChannEst(M,K,c,D)
  F=size(M,1)/K;
  J=size(M,2);
  % Step 1 - Speaker compensation
  for id=unique(c)
    ii=find(c == id);
    M(:,ii)=M(:,ii)-repmat(mean(M(:,ii),2)
        ,1,length(ii));
  end
  % Step 2 - Eigenchannel estimation
  opts.issym=1;
  opts.isreal=1;
  opts.disp=0;
  opts.tol=1E-3;
  [eigVec,eigVal]=eigs(1/J*M'*M,D,'lm',
     opts);
  V=eigVec*sqrt(eigVal);
  W=M*V;
  % Step 3 - Precompute O{k} matrices
  for k=1:K
    Wk=W(1+F*(k-1):F*k,:);
    O{k}=Wk'*Wk;
  end
```

5. Experimental setup

5.1. Partitioning of the previous SRE databases

To implement the dot-scoring speaker recognition system, the following sets were defined and used:

1. Universal Background Models (UBM)
2. Channel Compensation (CHC)
3. Z-Norm score normalization (SN-ZNorm)
4. T-Norm score normalization (SN-TNorm)
5. Development set

In order to create these sets, SRE04 to SRE08 (including FollowUp SRE08) were used. A study of the databases was carried out to avoid including signals from the same speaker in two different sets. Table 1 shows the speaker distribution in all the databases. The main diagonal shows the number of speakers per database, elements outside the diagonal representing the number of common speakers in each pair of databases.

5.1.1. SRE04 to SRE06

We found 1416 different speakers in the SRE04-06 sets: 180 of them (from SRE05 and SRE06) contained recordings with auxiliary microphones, whereas the remaining 1256 speakers were recorded only through different kind of telephones. Each set of speakers (either containing or not containing mic recordings) was divided into 4 different subsets (UBM, CHC and SN), and SN speakers were further divided into 2 additional sets (ZNorm

Listing 2: Channel compensation function code in Matlab. The functions parametres are vectors n and x, the zero and first order sufficient statistics of the target signal, and matrices W and O, as returner by the eigenchannel estimation function. Output value is vector y, the channel compensated first-order sufficient statistics vector.

```
function y = ChannelComp(n,x,W,O)
  K=length(O);
  L=eye(size(W,2));
  for i=1:K
    L=L+n(i)*O{i};
  end
  y=x-n.*(W*(L\(W'*x)));
```

Listing 3: Normalized means function code in Matlab. The functions parametres are vectors n and x, the zero and first order sufficient statistics of the target signal, and tau, the relevance factor for MAP adaptation. Output value is vector m, adapted and normalized mean vector. Note that if x is channel compensated, them m is channel compensated too.

```
function m = NormalizedMeans(n,x,tau)
  m=x./(tau+n);
```

and TNorm). Those speakers with the greatest number of signals acquired under different conditions where preferably assigned to the CHC set, whereas the remaining speakers were randomly distributed among the three other subsets. Table 2 shows the number of signals for the defined subsets.

5.1.2. SRE08

Unlike previous competitions, SRE08 included in the training and test conditions, for the core test, not only conversational telephone speech data but also conversational telephone speech recorded through microphone channels in an interview scenario. 150 speakers were recorded in this new condition.

The full SRE08 database was used as development set. To avoid interactions with previous databases, the signals of the 112 speakers in common with SRE06 (see Table 1) were not used. The signals of the remaining 1224 speakers, both in train and test, were divided into two well-balanced sets for development.

Table 1: Number of speakers per database (main diagonal) and and number of common speakers in each pair of databases (elements outside the diagonal).

	SRE04	***SRE05***	***SRE06***	***SRE08***	***FU08***
SRE04	310	0	0	0	0
SRE05	0	525	348	0	0
SRE06	0	348	949	112	0
SRE08	0	0	112	1336	150
FU08	0	0	0	150	150

Table 2: Number of signals from SRE04 to SRE06 in the Universal Background Models (UBM), Channel Compensation (CHC) and Score Normalization (ZNorm and TNorm) subsets.

	female	***male***	***Total***
UBM	2804	2119	4923
CHC	4586	3531	8117
TNorm	1479	960	2439
ZNorm	1403	1146	2549

Table 3: Distribution of signals in SRE08 into two balanced sets for development (devA and devB).

	SRE08	***SRE08_reduced***	***devA***	***devB***
train	3263	3149	1621	1528
test	6377	6211	3306	2905

5.1.3. FollowUp SRE08

The FollowUp SRE08 evaluation focused on speaker detection in the context of conversational interview speech. Test segments involved the same interview target speakers and interview sessions used in the SRE08 evaluation. Some involved the same microphone channels used in SRE08, whereas others were recorded through microphones not used previously.

The FollowUp SRE208 set, consisting of 6288 audio signals, was divided into two balanced subsets: CHC and SN, and the SN subset was further divided into two subsets: ZNorm and TNorm (see Table 4).

5.2. Preprocessing and Feature Extraction

The Qualcomm-ICSI-OGI (QIO)[3] noise reduction technique (based on Wiener filtering) was independently applied to the audio streams. The full audio stream was taken as input to estimate noise characteristics, thus avoiding the use of voice activity detectors on which most systems rely to constrain noise estimation to non-voice fragments.

Features were obtained with the Sautrela toolkit [4]. Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features, computed in frames of 25 ms at intervals of 10 ms. The MFCC set comprised 13 coefficients, including the zero (energy) coefficient. Cepstral Mean Subtraction (CMS), RASTA and Feature Warping were applied to cepstral coefficients. Finally, the feature vector was augmented with dynamic coefficients (13 first-order and 13 second-order deltas), resulting in a 39-dimensional feature vector.

Table 4: Distribution of speakers and signals in FollowUp SRE08 database.

		Signals		
	Speakers	***female***	***male***	***Total***
CHC	38 *2	2432	1776	4208
TNorm	18 * 2	1145	848	1993
ZNorm	19 *2	1212	875	2087

5.3. System configuration

Two gender dependent UBMs consisting of 1024 mixture components were trained with the Sautrela toolkit, using binary spliting, ortphan mixture discarting and variance flooring.

Channel compensation was trained for telephone-telephone, microphone-microphone and telephone-microphone variabilities, using 20, 20 and 40 eigenchannels, respectively.

Trials were conditioned on three channel types: no microphone sessions (0MIC), one microphone session (1MIC) and two microphone sessions (2MIC). Gender dependent and channel type condition dependent ZT normalization was performed on trial scores.

Side-info-conditional calibration was performed with FoCal [5], using channel type and gender conditioning. Scores were calibrated to be interpreted as detection log-likelihood-ratios, and the hard accept/reject decisions were made by applying a Bayes threshold of 6,907 (derived from the SRE2010 competition costs, $P_{target} = 0.001$, $C_{miss} = 1$ and $C_{fa} = 1$).

6. Evaluation results

The year 2010 speaker recognition evaluation was part of an ongoing series of evaluations conducted by NIST. The core train and test conditions involved telephone conversational excerpts (of approximately five minutes total duration) and microphone recorded conversational segment (of three to fifteen minutes total duration), with 5460 train segments, 13344 test segments and a total of 610748 trials.

Five main conditions[1] were carried out in the core SRE2010 evaluation, according to train and test recording conditions mismatch:
1 - Interview in train and test, same mic.
2 - Interview in train and test, different mic.
3 - Interview in train and phonecall over tel channel in test.
4 - Interview in train and phonecall over mic channel in test.
5 - Phonecall in train and test, different telephone.

Figure 1 shows the DET curves for the dot-scoring system in the five core conditions. Minimum and actual cost operation points are marked with circles and asterisks, respectively. Whenever the test segment is related to microphone signals (conditions 1, 2 and 4), the DET curves show a calibration error (big distance between minimum and actual cost points). On the other hand, when the test is carried out over the telephon channel, the calibration is really good. A mismatch bewteen the designed development set and the evaluation set could explain this calibration issue.

7. Conclusions

The dot-scoring speaker recognition system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the NIST 2010 Speaker Recognition Evaluation has been described. The system combines two key technologies: sufficient statistics space eigenchannel compensation and dot scoring. An optimized Matlab implementation of of the eigenchannels estimation, the channel compensation and the normalized mean vector computation has been provided.

The dot-scoring system attained competitive results at the NIST SRE 2010, despite being a much simpler approach compared to other methodologies. On the other hand, the calibration errors with microphone test segments suggest a mismatch bewteen the designed development set and the evaluation set.

[1]Another four conditions related to different vocal efforts were also evaluated, but they will be ignored in the current work.

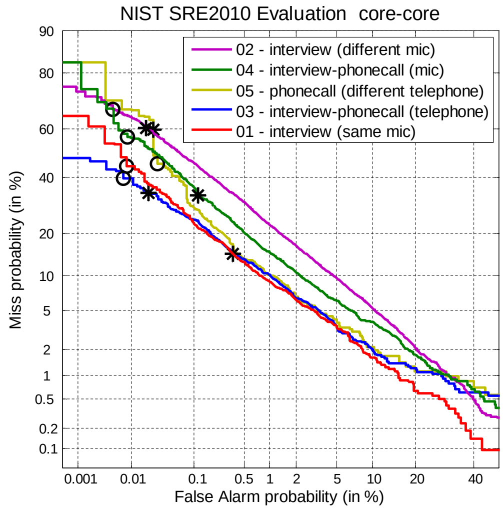


Figure 1: DET curves for the dot-scoring system in the five core conditions. DET curves are ordered in descending Equal Error Rate (ERR), the second condition (interviews with different mic) being the worst and the first one (interviews with same mic) being the best. Minimum cost operation point are marked with circles, and actual operation points with asterisks.

8. References

[1] A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.

[2] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocký, "Analysis of feature extraction and channel compensation in gmm speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, 2007.

[3] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proceedings of ICSLP2002*, 2002.

[4] M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, (San Juan, Puerto Rico), pp. 386–391, December 2005.

[5] *Tools for detector fusion and calibration, with use of side-information*. http://sites.google.com/site/nikobrummer/focalbilinear.

A Fishervoice-based Speaker Identification System

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

Department of Signal Theory and Communications, Universidade de Vigo, Spain
{plopez, ldocio, carmen}@gts.tsc.uvigo.es

Abstract

In this paper, a novel approach for speaker identification called Fishervoice is proposed. It was inspired by the two-dimensional (2D) Fisherface technique, which is a method that combines a two-stage "PCA+LDA" strategy and two-dimensional discrimination techniques. Experimental results on the BANCA database demonstrate that the Fishervoice approach is effective for speaker identification tasks, in particular when there are mismatched conditions. The reduction on the number of parameters needed to describe each speaker model achieved with the Fishervoice technique is remarkable, thereby causing a reduction of computational and memory costs.

Index Terms: fishervoices, speaker identification, dimensionality reduction

1. Introduction

Despite the great advances made recently in the field of speaker recognition systems, they still lack robustness, i.e. their performance degrades dramatically when the acoustic training data differs from the given test conditions. Robustness is currently the major challenge in speaker recognition for real-world applications [1]. For example, in telephone services, users may call in under all kinds of acoustic environments (in the office, on the street, in the car) and use different telephone networks (land-line or cellular). Therefore, mismatched conditions may be found at any time, which makes robustness one of the critical factors that decide the success of speaker recognition technology in these applications.

Most of the state-of-art speaker recognition systems use Gaussian mixture models (GMM) as statistical models to represent the speakers in terms of the probability distribution of low-level acoustic features. These systems achieve very high accuracies on high-quality data when training and test conditions are well controlled, but their performance is significantly degraded under adverse and mismatched conditions. Nowadays, an interesting area of research is the use of discriminant analysis techniques in speaker recognition [2], in order to reduce intra-speaker and channel variability. In this way, speaker recognition techniques based on speaker subspace modeling have been proposed, such as the "eigenvoice approach" [3][4], and more recently kernel learning methods are arising a great interest. The work presented in this paper is inspired by the work described in [5] on face recognition.

Given the similarities between the study of faces and voices, the Fishervoice method (analogous to the Fisherface method) is presented in this paper, which takes a two-stage "PCA+LDA" strategy. It first uses two-dimensional Principal Component Analysis (PCA) to reduce the dimensionality, and then performs Linear Discriminant Analysis (LDA) to extract a discriminative subspace. Thus, in this paper a research to analyze the robustness and effectiveness of this dimensionality reduction and subspace learning technique is performed, to find out if this technique used in face recognition tasks to minimize the computational cost and alleviate the curse of dimensionality can also be helpful in speaker identification (SID) tasks.

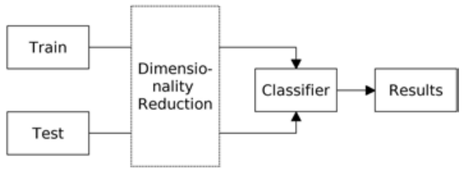


Figure 1: *Fishervoice SID system*

The outline of the paper is as follows. In Section 2, the Fishervoice method and the speaker identification algorithm are presented. In Section 3 the experimental framework is described. In Section 4 experimental results are presented. Finally Section 5 explains the conclusions of the paper.

2. Proposed SID System

The algorithm proposed in this paper to perform speaker identification is very simple, as shown in Fig. 1. There are two data inputs: a train dataset, used to model the speakers in the system, and a test dataset, where each of its elements has to be assigned to a speaker model. To perform this assignment, a classifier is used to decide which speaker segment of the train dataset is more similar to a given speech segment S from the test dataset, thus deciding that the speaker of S is the one that spoke the speech segment in the model that is more similar to S.

It can be seen in figure 1 that a dimensionality reduction step is performed before the classification step. It is at this point when the Fishervoice approach that is proposed in this paper is applied. If this step is skipped, the SID system will be equivalent to a GMM identification system, where a speaker segment is modeled by a GMM and GMMs are compared. This GMM identification system will be used as a baseline to make a comparison between its performance and the performance of the Fishervoice identification system.

2.1. The datasets

Two datasets are needed to perform speaker identification:

- A train dataset (A_{train}) composed by segments of speech spoken by different known speakers. This data is used to model the target speakers.
- A test dataset (A_{test}) composed by segments of speech that have to be assigned to the most likely speaker in the train set.

A_{train} and A_{test} are tridimensional matrices of dimension $m \times n \times L_{train}$ and $m \times n \times L_{test}$, respectively. L_{train} is the number of speech segments in matrix A_{train}, i.e. it is the number of speaker segments that are available to model the different speakers. Consequently, A_{test} is the number of speaker segments that have to be assigned to a speaker in the model.

Each segment in both sets will be represented by the means of a Gaussian Mixture Model (GMM), where the number of gaussians of the model is m and the dimension of the feature space is n. To obtain this GMMs, a Maximum a Posteriori (MAP) adaptation of a universal background model (GMM-UBM) is performed with the available acoustic features. In this work, the acoustic features are 12 Mel-frequency Cepstral Coefficients (MFCC), extracted using a 25ms Hamming window at a rate of 10ms per frame, and augmented with the normalized log-energy and their delta and acceleration coefficients. Thus, the dimension of the feature space (n) used in this paper is 39.

2.2. Dimensionality reduction: the Fishervoice method

The dimensionality reduction strategy proposed in this paper is based on an adaptation of the procedure presented in [5] for face recognition in order to make it suitable for speaker recognition. In [6] a different, but also called, fishervoice approach has been applied to a speaker clustering task, but in the fishervoice approach described in this paper a two-stage "PCA+LDA" strategy combined with a two-dimensional discrimination technique is applied, while in [6] only LDA is used.

Consider a set A_{train} representing speech segments as explained in 2.1. This dataset will be used to compute two transformation matrices X and Y as explained below.

The between-class D_b, within-class D_w and total D_t scatter matrices are defined as:

$$D_b = \sum_{i=1}^{c} P_i (M_i - M)^T (M_i - M) \tag{1}$$

$$D_w = \sum_{i=1}^{c} \sum_{j, j\in i} (A_{train_j} - M_i)^T (A_{train_j} - M_i) \tag{2}$$

$$D_t = D_b + D_w \tag{3}$$

where c is the number of different speakers in A_{train}, P_i is the a priori probability of the ith class, M_i is the mean matrix of the ith class ($i = 1, 2, \cdots, c$), M is the total mean matrix of A_{train}, and A_{train_j} is the $m \times n$ matrix of the jth segment in A_{train}. So, we can understand M as the mean voice of the whole speaker set, and M_i as the mean voice of speaker i.

After computing these $n \times n$ matrices, the eigenvectors and eigenvalues of D_t are computed, finding a matrix X that maximizes $J(X) = X^T D_t X$. To reduce the dimensionality and make the system less time and memory consuming, an automatic strategy for dimensionality reduction is applied. The proposed selection strategy keeps only a percentage of the energy of the subspace (E_X):

$$E_X = \sum_{i=1}^{n} \lambda_i \tag{4}$$

where λ_i is the ith greatest eigenvalue of X. In the end, matrix X keeps a number u of columns (eigenvectors) equal to the number of eigenvalues needed to absorb a given percentage e_1 of the energy E_X. Hence, X is a $n \times u$ matrix.

After obtaining X, the sample set A_{train} is transformed into a new space with a lower dimensionality by doing $B_{train} = A_{train}X$. Then, new between-class and within-class scatter matrices (R_b and R_w, respectively) are computed:

$$R_b = \sum_{i=1}^{c} P_i (L_i - L)(L_i - L)^T \tag{5}$$

$$R_w = \sum_{i=1}^{c} \sum_{j, j\in i} (B_{train_j} - L_i)(B_{train_j} - L_i)^T \tag{6}$$

where L is the mean voice of the set B_{train}, and L_i is the mean voice of the ith speaker in that set.

Applying the Fisher criterion, a matrix Y that maximizes $J(Y) = \frac{Y^T R_b Y}{Y^T R_w Y}$ is obtained. Again, an automatic strategy for dimensionality reduction is applied as before, causing Y to become a $m \times v$ matrix by keeping the $e_2\%$ of the energy of the subspace E_Y.

Finally, performing the transformation $C_{train} = Y^T B_{train}$ a new sample set C_{train} composed by $v \times u$ matrices is obtained. After this procedure, a new representation of the dataset A_{train} with lower dimensionality is obtained.

After computing X and Y, the test data matrix A_{test} is projected to this new low dimensionality subspace by doing:

$$B_{test} = A_{test} X \tag{7}$$

$$C_{test} = Y^T B_{test} \tag{8}$$

2.3. Classifier

After reducing the dimensionality of the datasets, two tridimensional matrices C_{train} and C_{test} of dimensions $v \times u \times L_{train}$ and $v \times u \times L_{test}$ respectively are obtained. A transformation to bi-dimensional matrices is performed, obtaining two matrices C'_{train} and C'_{test} of dimensions $vu \times L_{train}$ and $vu \times L_{test}$ respectively. This transformation consists on stacking the means of the GMMs, i.e. concatenating the rows of each of the L_{train} (L_{test}) matrices in C_{train} (C_{test}) to obtain a matrix of supermean vectors. This transformation is not really necessary, but it makes the classification task easier, because now vectors are compared instead of matrices.

To decide which of the L_{train} speakers spoke one of the segments S in the test set, the following expression is evaluated:

$$T = \min_i d(C'_{test_S}, C'_{train_i}) \tag{9}$$

where $d(.,.)$ is the euclidean distance between two vectors. The speaker of the segment T that minimizes the euclidean distance to the segment S is chosen as the speaker of S. Experiments were performed with different distance measures (for example, Mahalanobis distance), but the best results were achieved with the euclidean distance.

3. Experimental framework

3.1. Description of the database

The speaker identification system proposed in this paper was tested using the BANCA database [7] [8]. This database includes 52 English speakers (26 males and 26 females) each of whom recorded 12 sessions divided into 3 different scenarios: controlled, degraded and adverse.

Each session was recorded using two different-quality microphones. In each session the speaker recorded two different utterances, hence there are eight utterances per speaker in each scenario. A partition of the data in three groups has to

Table 1: Summary of the datasets used in the experiments.

	Experiment	GMM-UBM	Train	Test
GT	1	All	All	All
Matched	1	Controlled	Controlled	Controlled
	2	Degraded	Degraded	Degraded
	3	Adverse	Adverse	Adverse
	4	All	Controlled	Controlled
	5	All	Degraded	Degraded
	6	All	Adverse	Adverse
Mismatched	1	Cont./Deg.	Controlled	Degraded
	2	Cont./Adv.	Controlled	Adverse
	3	Degr./Cont.	Degraded	Controlled
	4	Degr./Adv.	Degraded	Adverse
	5	Adv./Cont.	Adverse	Controlled
	6	Adv./Deg.	Adverse	Degraded
	7	All	Controlled	Degraded
	8	All	Controlled	Adverse
	9	All	Degraded	Controlled
	10	All	Degraded	Adverse
	11	All	Adverse	Controlled
	12	All	Adverse	Degraded

be done, in order to have different data for training the GMM-UBM, computing the matrices X and Y and testing. The eight utterances per speaker are divided as follows: two are used to train the GMM-UBM, three to train the matrices, and three are used for testing.

Three different groups of experiments are described in this paper. The first group consists of only one experiment, and is called Grand Test (GT) because of its similarity to the one with the same name in [7]. This is an experiment that uses data from all the scenarios both for training and for testing. The second group are experiments in matched conditions, i.e. experiments where the data used for training is from the same scenario as the data used for testing. Finally, the third group are experiments in mismatched conditions, where the data used for training is from a different scenario as the data used for testing.

Table 1 describes the different datasets used for the experiments. GT experiment is a global recognition test, using the three different scenarios for training and testing. Experiments 1, 2 and 3 in matched conditions are scenario-based, i.e. training and testing are performed using only data from a given scenario. Experiments 4, 5 and 6 in matched conditions use a scenario-independent GMM-UBM (trained with data from the three scenarios), but matrices X and Y are obtained using data from the same scenario as the Test set.

In the experiments in mismatched conditions, Train comes from scenario i, while Test comes from scenario j, where $i \neq j$. Note that there are two different GMM-UBM sets, separated by a slash (/), i.e. T_i/T_j. This means that T_i is the GMM-UBM used in Train, and T_j is the GMM-UBM used in Test. This means that Train and Test use a GMM-UBM trained with data that belongs to their respective scenarios. GMM-UBMs of 16, 32, 64 and 128 gaussians are going to be used in the experiments. The number of gaussians in the GMM-UBM will be referred to as M.

4. Results

4.1. Baseline

Table 2 shows the accuracies obtained by performing speaker identification without using dimensionality reduction techniques, i.e. comparing GMMs directly. The results obtained for the GT and the experiments in matched conditions show that the baseline achieves acceptable accuracies, but the error rate in mismatched conditions is, in general, excessively high.

For each experiment, Table 2 also indicates which number of gaussians obtained the highest accuracy (M), choosing the lowest one when the same result was obtained with different GMMs. The dimensionality of the data is always $M \times n$.

Table 2: Baseline results.

	Experiment	Accuracy	M
GT	1	91.9872	128
Matched	1	96.1538	32
	2	94.2308	64
	3	94.2308	32
	4	96.1538	64
	5	94.8718	128
	6	97.4359	64
Mismatched	1	10.2564	32
	2	3.8462	32
	3	13.4615	32
	4	4.4872	32
	5	5.1282	32
	6	6.4103	32
	7	94.2308	128
	8	32.6923	128
	9	88.4615	32
	10	31.4103	64
	11	33.3333	64
	12	33.9744	64

4.2. Fishervoice approach

Figure 2 shows the accuracies obtained in the GT experiment, using different values of the energy percentages e_1 and e_2, and GMMs with 16, 32, 64 and 128 gaussians. The values in blue are the lowest, and the ones in red are the highest. The aim is to obtain the highest accuracy with the lowest dimensionality, and in this case it is an accuracy of 99.0385%, with a subspace of dimension 33×39. Nevertheless, an accuracy of 98.7179% can be reached with a subspace of dimension 32×9, which will be computationally better.

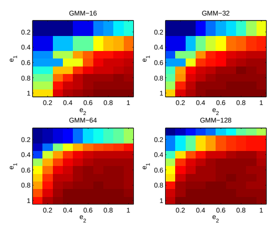


Figure 2: *Results for experiment 1.*

As one of the aims of this system is to obtain high accuracies with low dimensionality subspaces, the accuracies obtained using $e_1 = 100\%$ or $e_2 = 100\%$ will be discarded. Table 3 shows the maximum accuracies achieved in the experiments in Table 1 (using the best-quality microphone) and the lowest dimensionality subspaces that achieve them with a GMM-UBM of M gaussians.

The error rate in the GT experiment, which is the most general, is approximately 1.3%. Experiments in matched conditions obtain an accuracy of 100% with the different GMM-

Table 3: Experimental results with the Fishervoice approach.

	Experiment	Accuracy	Dimension	e_1	e_2	M
GT	1	98.7179	32×9	90	80	64
Matched	1	100	5×8	90	70	16
	2	100	20×6	80	60	64
	3	99.359	37×8	80	70	128
	4	100	3×9	90	30	32
	5	100	10×6	70	50	64
	6	100	18×10	90	50	128
Mismatched	1	28.2051	2×9	90	10	64
	2	9.6154	1×6	70	10	64
	3	39.7436	1×5	60	10	16
	4	10.2564	2×9	90	20	32
	5	30.7692	1×9	90	10	32
	6	31.4103	2×5	60	20	32
	7	100	15×9	90	60	64
	8	70.5128	15×9	90	80	32
	9	100	8×9	90	60	32
	10	74.359	22×9	90	90	32
	11	79.4872	21×9	90	90	32
	12	78.2051	11×9	90	90	16

UBMs, except in experiment 3, which corresponds to a degraded scenario, where the error rate is about 0.7%. It can also be appreciated in table 3 that the dimensionality of the subspace needed to obtain these results is lower in experiments 4,5,6 than in experiments 1,2,3.

In experiments 1 to 6 in mismatched conditions, where Train and Test are adapted using the GMM-UBM corresponding to their own scenario, the error rate is too high. Nevertheless, in experiments 7 to 12, where a global GMM-UBM was used for the adaptation, higher accuracies are obtained, mainly in experiments 7 and 9 where Train and Test belong to the least degraded scenarios.

Comparing tables 3 and 2, it can be observed that the use of the Fishervoice approach achieves higher accuracies in all the experiments. Not only accuracies are higher, but lower dimensionality of data is handled in all cases, making the Fishervoice dimensionality reduction method effective for speaker identification.

Figure 3 compares the results obtained in the experiments using the two available microphones, where the green bars represent the results obtained with microphone 1, and the yellow bars represent the results obtained with microphone 2. It can be seen that, in general, better accuracies are obtained using microphone 1, due to its better quality. Nevertheless, results obtained with microphone 2 are acceptable in matched conditions.

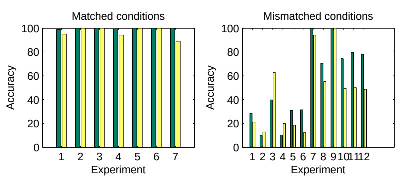


Figure 3: *Results with different microphones.*

5. Conclusions and Future Work

A PCA-LDA based speaker identification system is presented in this paper with two goals: obtain a good performance even in mismatched conditions, and reduce the dimensionality of the data in order to reduce the computational load. Table 3 shows that the speaker identification is almost perfect in matched conditions, and acceptable in mismatched conditions. Moreover, it outperforms the baseline, where no dimensionality reduction techniques are applied, and reduces the dimensionality of the data to be handled. It is also noticeable that the best accuracies are obtained when the GMM-UBM employed is trained with both clean and degraded data, thus helping the recognition system to work with degraded samples. In addition, a substantial reduction of the dimensionality is achieved, allowing the system to be less time and memory consuming, as the vectors and matrices that represent the speaker segments are smaller, and the computing time for classification is reduced.

The main problem of this method is the selection of the best values for e_1, e_2 and M, being necessary to perform some research in the future to find an automatic manner to choose the most suitable values for these parameters.

The GT experiment, which is the most interesting because it does not matter the quality of the samples used for training and testing, has an error rate of 1.3%, making this method useful for real applications.

In future work, the validity of the Fishervoice method for speaker verification will be tested, going into the method in depth to improve it both in speaker identification and verification.

6. Acknowledgements

The work reported here was partially supported by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04 “Búsqueda de Información en Contenidos Audiovisuales plurilingües” and by the Ministry of Science and Innovation under the FPI grant BES-2010-033358.

7. References

[1] X.-H. Ren, Y.-F. Zhang, Y.-J. Xing, M. Li, ”Application of KPCA and PNN for Robust Speaker Identification“, Proceedings of the 2008 Congress on Image and Signal Processing, vol. 4, pp. 533–536, 2008.

[2] A. Errity and J. McKenna, ”A Comparative Study of Linear and Nonlinear Dimensionality Reduction for Speaker Identification“, Proc. of the 15th International Conference on Digital Signal Processing (DSP), pp. 587–590, Cardiff, Wales, 2007

[3] R. Kuhn, J.-C. Junqua, P. Nguyen and N. Niedzielski, ”Rapid Speaker Adaptation in Eigenvoice Space“, IEEE Trans. Speech and Audio Processing, vol. 8, n. 6, pp. 695–707, 2000

[4] O. Thyes, R. Kuhn, P. Nguyen and J.-C. Junqua, ”Speaker Identification and Verification Using Eigenvoices“, International conference on Spoken Language Processing, pp. 242–245, Beijing, China, October 2000.

[5] X.Y. Jing, H.S. Wong and D. Zhang, “Face Recognition Based on 2D Fisherface Approach”, Pattern Recognition, vol. 39, n. 4, pp. 707–710, 2006.

[6] S.M. Chu, H. Tang, T.S. Huang, ”Fishervoice and Semi-supervised Speaker Clustering“, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4089–4092, Los Alamitos, CA, USA, 2009

[7] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Mariethoz, J. Matas, K. Messer, F. Poree, B. Ruiz, “The BANCA Database and Evaluation Protocol”, 2003.

[8] The BANCA Database Website, Online: http://www.ee.surrey.ac.uk/CVSSP/banca/

ATVS-UAM NIST SRE 2010 SYSTEM

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
{javier.gonzalez,ignacio.lopez,javier.franco,daniel.ramos,doroteo.torre,joaquin.gonzalez}@uam.es

Abstract

This paper describes the system submitted by ATVS-UAM to the 2010 edition of NIST Speaker Recognition Evaluation (SRE). Instead of focusing on multiple, complex and heavy systems, our submission is based on a fast, light and efficient single system. Sample development results with English SRE08 data (data used in the previous evaluation in 2008) are 0.53% EER (Equal Error Rate) in tel-tel (telephone data used for training and testing) male data (optimistic evaluation), going up to 3.5% (tel-tel) and 5.1% EER (tel-mic, telephone data for training and microphone data for testing) in pessimistic cross-validation experiments. These results are achieved with an extremely light system in computational resources, running 77 times faster than real time.

Index Terms: speaker recognition, speaker recognition evaluation, factor analysis.

1. Introduction

Our group, ATVS-UAM, has been participating in NIST (National Institute of Standards and Technology) Speaker Recognition Evaluations (SRE) since 2001. In these years speaker recognition technology has evolved dramatically, passing through different phases. During the first years of this period technology has been dominated by the Gaussian Mixture Model – Universal Background Model (GMM-UBM) technique [1]. This technique was fast and accurate but suffered great degradation with inter-session variability. For this reason, it was constantly improved by new channel and in general inter-session variability compensation schemes such as Cepstral Mean Normalization (CMN), RASTA filtering [2], Feature Warping [3], Feature Mapping [4], and so on. Since 2003 and probably up to 2007 there was a generalized trend to fuse GMM-UBM systems with what were known as 'higher-level' systems [5] because they operated on higher levels of information of the speech signal (prosodic, phonotactic, lexical, dialogic, etc.) than the acoustic level used by the GMM-UBM systems. These systems exploited information that was not taken into account by GMM-UBM systems, and therefore provided additional information that tends to fuse well with acoustic-based GMM-UBM systems. However, higher-level systems tend to be computationally expensive and result in a multiplicity of systems that make computational complexity of the overall systems very high and even prohibitive. Since 2005 [6,7] a new inter-session compensation paradigm has appeared for the GMM-UBM framework that has improved so much the performance of this technology that has made it the mainstream again, letting higher-level systems as an interesting option to reduce a few decimals in the scores of the NIST competition, but a not so interesting option for real systems. This paradigm is generally known as Joint Factor Analysis and consists in working in a high-dimensional feature space, the super-vector space, in which the feature vector is composed by the concatenation of the means of the GMM. Provided that we work with diagonal covariance GMMs, with 1024 Gaussians and a speech parameterization that provides a vector of 39 features per frame, the super-vector will include 1024 x 39 = 39936 dimensions. Once an utterance is transformed in a vector in this high-dimensional space, the Joint Factor Analysis approach tries to determine low dimension sub-spaces of this high-dimensional space that cover most of the inter-session variance and most of the inter-speaker variance. Once these sub-spaces are indentified the speaker is identified using the information in the speaker variability sub-space. More recently a new approach called total-variability [8] has been proposed that does not try to disentangle speaker and inter-session variability and rather finds a sub-space (typically of 400 dimensions) that covers most of the variability (both speaker and inter-session) by means of Principal Component Analysis (PCA). The vectors in this sub-space are then compared, after compensation using Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalization (WCCN), with a simple cosine distance function, showing better performance than the more complex Joint Factor Analysis approach [6]. This is the approach that we have used in our system for NIST SRE 2010. The rest of the paper is organized as follows. Section 2 describes gives a brief overview of NIST SRE 2010, focusing in particular in the data used for the evaluation. Section 3 describes feature extraction with particular emphasis on the use of two voice activity detectors, a point that we consider crucial for the success in this evaluation. Then we describe the core of our system (section 4). Finally, we describe the development and evaluation results, including measures of computational complexity (section 5) and conclude the paper in section 6.

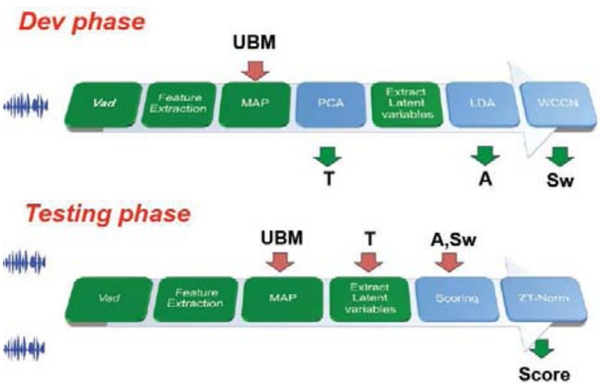


Figure 1. Developing (training) and testing phase of ATVS-UAM NIST SRE 2010 System.

Table1. Development data composition for total space training. (#Utterances/#speakers).

Gender		***Tel-Tel***	***Tel-Mic***
Male	***T/LDA***	5656/824	7868/452
	WCCN	5230/611	7838/437
Female	***T/LDA***	5155/889	10973/610
	WCCN	4521/572	10900/607

2. Overview of NIST SRE 2010

A complete description of the NIST speaker recognition evaluation is available in [9]. In general all these evaluations pose a speaker detection challenge in which the speaker models are trained on training data provided by NIST (and previously unreleased) and, after training the speaker models, these should be used to detect the speakers in test data also provided by NIST and also previously unreleased. The participants must submit their results without knowing the speaker assignments and without hearing the audios. In this paper we are only interested in one of the conditions, the so called core-core condition in which the training and testing material was one two-channel telephone conversational excerpt (we call this type of data *tel* data), of approximately five minutes total duration or a microphone recorded conversational segment (we call this type of data *mic* data) of three to fifteen minutes total duration involving the interviewee (target speaker) and an interviewer, in both cases with the target speaker channel designated. The type of data was known in advance for the systems. The evaluation established a maximum of 6000 speaker models and a maximum of 25000 test segments with a maximum of 750000 trials. The real evaluation was close to those figures.

3. Audio Processing and Feature Extraction

In our system, all audio except that used for tel-tel trials (tel data used for train and test) was first filtered with the QIO (Qualcomm-ICSI-OGI) Wiener filter in order to reduce noise [10]. Feature extraction is performed after noise reduction. It computes 38 coefficients per frame (19 Mel-Frequency Cepstrum Coefficients, MFCC, and deltas) using 20 ms. Hamming windows, overlapped 10 ms and 20 mel-spaced (300-3300 Hz) magnitude filters. Once these features are calculated three channel compensation methods are applied in sequence: CMN, RASTA [2] and Feature Warping [3] with 3 second windows.

Given that the data provided by NIST included speech from conversations, there were long periods in which the target speaker was in silence. In order to avoid processing those segments and achieve better performance we have used two different VAD (Voice Activity Detection) configurations depending on whether the data is *mic* or *tel*. *tel* audios are segmented into speech and non-speech segments combining an energy-based VAD developed by our group, and a VAD tool provided by Sound eXchange (SOX) [11] which uses speech enhancement and dynamic noise modelling. Only segments labelled as speech by both VADs are considered to be valid speech segments. For *mic* audios, we firstly remove the interviewer speech from the audio. In order to detect interviewer activity segments to remove, two different criteria have been used. The first criterion is based on an energy detector applied over the channel corresponding to the interviewer’s microphone. Unfortunately for some recordings, the dynamic range was not enough for detecting any interviewer activity. In those cases, the energy based activity labels were replaced by the ASR (Automatic Speech Recognition) labels also provided by NIST (segments marked as silence was considered silence and segments with any word recognized as speech). After the interviewer speech was removed a VAD scheme equivalent to the one applied for *tel* data is used to detect valid speech segments.

4. Core Speaker Recognition

Figure 1 tries to represent the developing or training phase, and the testing phase of ATVS-UAM system. Our system is a single system based on Gaussian Mixture Models (GMM) where a ‘Total Variability’ modelling strategy [8] was employed in order to model both speaker and session variability. The ‘total variability’ scheme shares the same principles as Joint Factor Analysis (JFA) systems [6, 7], where variability (speaker and session) is supposed to be constrained, and therefore modelled, in a much lower dimensional space than the GMM-supervector space. However, unlike JFA, a *total space* which jointly includes speaker and session variability (represented by a low-rank T matrix) is computed instead of computing two separate subspaces as in JFA (matrices U and V). In our system we trained matrix T (Figure 1) with the development data shown in Table 1. After having the vectors computed in the total variability space defined by T, a session variability compensation stage is applied by means of Linear Discriminant Analysis (LDA), in which we train and use matrix A in Fig. 1, and Within-Class Covariance Normalization (WCCN), in which we train and use matrix Sw in Figure 1.

Instead of using a single total variability subspace, two gender dependent total subspaces of 200 dimensions were generated after applying LDA to a 400 (rank of T) dimensions space calculated via classical eigenanalysis from background data (Table 1). Two different *total spaces* were considered, namely tel-tel (telephone only) and tel_mic. The background, employed to construct the *total spaces* and the Universal Background Model from which GMM-supervectors models were derived (Table 1) contains a subset of data belonging to

Table 2: Breakdown timing for ATVS core system.

	GMM-FA
Testing (per 265s file)	
Total space hidden variables	0.05s
Scoring	1e-6 s
Z-norm	0.02s (~300 test)
T-norm	0.02s (~300 models)
Total (test)	3.66s
xRT test (CPU/speech)	0.013 RT

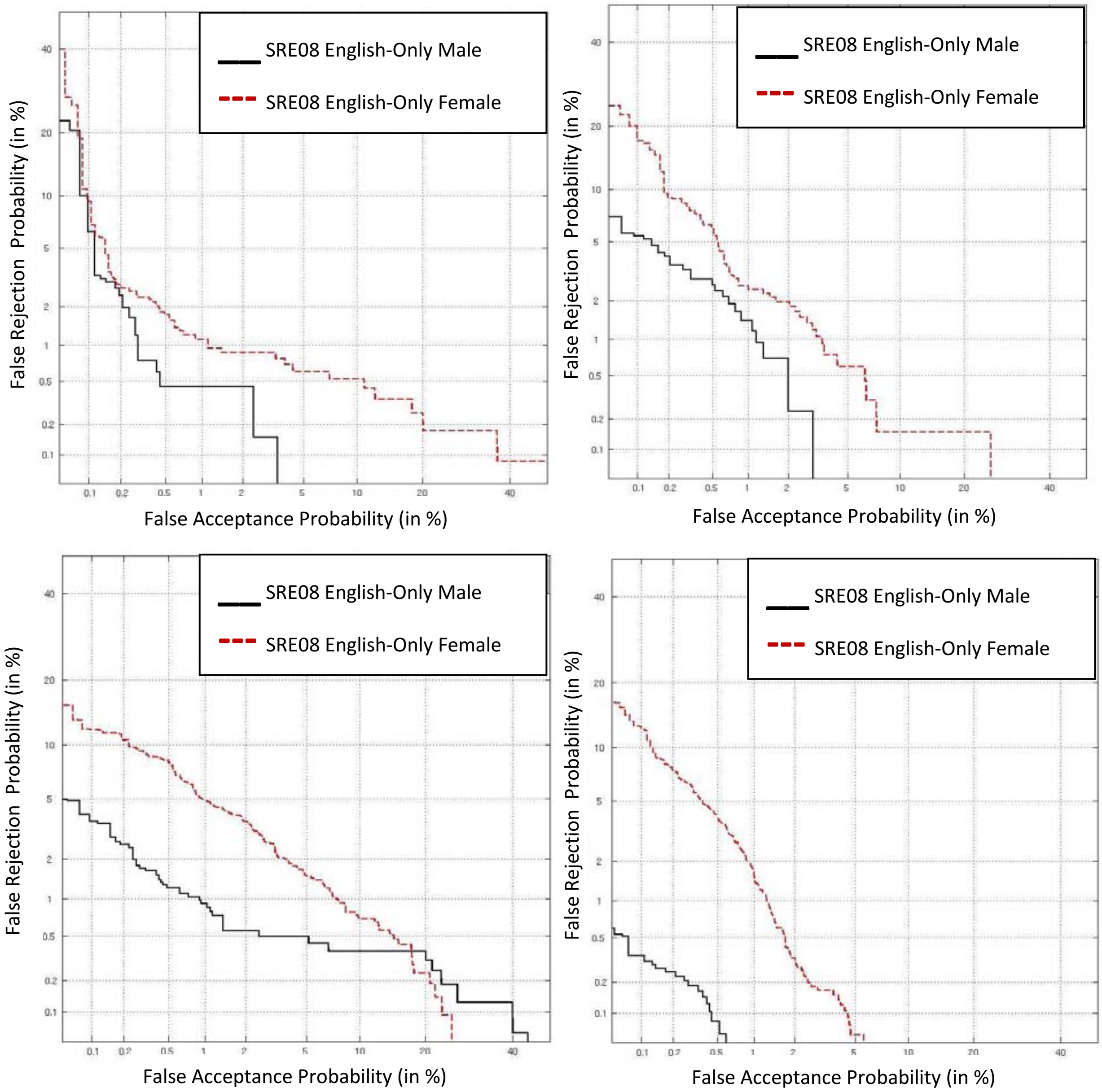


Figure 2: Development results for SRE08 english-only trials in different conditions: tel-tel (top-left), tel-mic (bottom-left), mic-tel(top-right), mic-mic(bottom-right).

Switchboard-I, Switchboard-II phase 2 and 3 and MIXER (from SREs 04, 05, 06 and 08).

The system uses a fast scoring procedure similar to [8]. Scores are then normalized using ZT-norm (Figure 1) and finally calibrated using linear logistic regression with the FoCal toolkit [12]. Calibration has been performed in a gender-independent way using different calibration rules for scores generated using microphone data in training, testing or both and scores generated using just telephone data.

5. Results

Figure 2 shows results obtained in the development phase for optimistic estimation of the T matrix (test data used for estimating it). Results range from a mere 0.53% EER (for tel-tel male and about 3% EER for tel-mic female. In order to have a less optimistic evaluation we used cross-validation excluding 25% of the test files for training 4 different T matrices and testing on the files excluded using the worst case in Figure 2. In this way we obtain Figure 3 in which EER increases up to a 5.13%, which is the result we expected in the real evaluation.

Figure 4 (a figure generated by NIST) shows the results attained by ATVS-UAM system in the real NIST SRE 2010 for the condition using interviews and the same microphone for train and test. This corresponds to our best result, a 3.5% EER. For comparison, best systems in this same condition obtain an EER slightly below 2%. Our results in other conditions can go up to 8.5% EER, which is only slightly worse than the 5.13% EER obtained in our worse development test.

Our emphasis in this evaluation was in developing an accurate and fast system. In this sense, Table 2 summarizes ATVS core system testing timing. All execution times have been obtained in a Red Hat Enterprise 5.0 server on a 2.2 GHz CPU, with cache memory of 1024 kB and RAM of 4GB. The speaker recognition process runs 77 times faster than real time, which makes the system widely applicable in real applications.

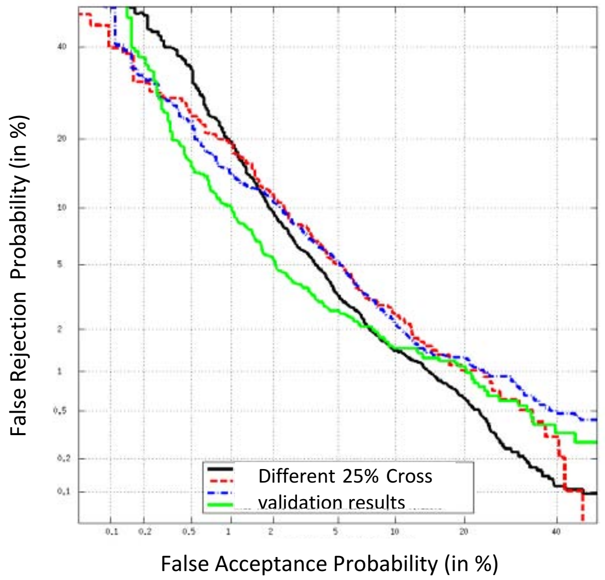


Figure 3. Cross validation development results for all SRE08 conditions, where each cross validation subset totally excludes the 25% of speakers in the subset test from the development.

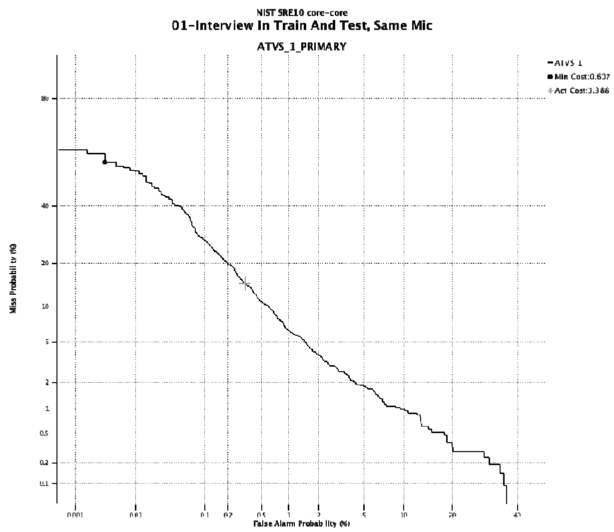


Figure 4. Actual evaluation results achieved at NIST SRE 2010. The figure corresponds to the sub-case using only interview data for train and test with the same microphone

6. Conclusions

This paper has presented the system submitted by ATVS-UAM to NIST SRE 2010. The system is based on a light and effective single system based on Total variability and achieved a 3.5% to 8.5% EER (depending on the condition) on the NIST SRE 2010 real evaluation, working over 75 times faster than real time.

7. Acknowledgements

This work has been partially supported by projects MEC TEC2009-14719-C02-01 and CAM S2009/TIC-1542 MA2VICMR.

8. References

[1] Reynolds D., Quatieri T., and Dunn R., Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10:19-41, 2000.

[2] Hermansky H. and Morgan N. Rasta processing of speech. IEEE Transactions on Speech and Audio Processing, 2(4):578-589, 1984.

[3] Pelecanos J. and Sridharan S., Feature Warping for Robust Speaker Verification, in 2001: A Speaker Odyssey: The Speaker Recognition Workshop, Crete, Greece, June 2001.

[4] Reynolds, D., Channel Robust speaker verification via Feature Mapping, in: IEEE International Conference on Acoustic Speech, and Signal Proccesing, 2003.

[5] D.A. Reynolds, et al. “The SuperSID Project: Exploiting high-level information for high-accuracy speaker recognition,” Proc. ICASSP-03, Hong Kong, Apr 2003.

[6] Kenny, P. and Boulianne, G. and Dumouchel, P., “Eigenvoice Modeling With Sparse Training Data”, IEEE Trans. on Speech and Audio Processing, vol. 13, no. , pp 345-354, 2005.

[7] R. Vogt and S. Sridharan, “Explicit modelling of session variability for speaker verification,” Computer Speech & Language, vol. 22, no. 1, pp. 17–38, 2008.

[8] Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P and Dumouchel, P., Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification In Proc Interspeech 2009, Brighton, UK, September 2009.

[9] NIST SRE 2010 Evaluation Plan, available at http://www.itl.nist.gov/iad/mig//tests/sre/2010/NIST_SRE10_evalplan.r6.pdf (accessed 23/09/2010).

[10] Qualcomm, ICSI, OGI (QIO) Front-End software, available at http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/qio/ (accessed 12/04/2010).

[11] “Sound Exchange” software, Available at http://sox.sourceforge.net/ (accessed 28/06/2010).

[12] Niko Brummer, “FoCal Toolkit”, Available at http://sites.google.com/site/nikobrummer/focal (accessed 12/04/2010).

Oral Session 3: Speaker Characterization

Post-evaluation analysis and improvements with the L^{2}F Language Verification System submitted to NIST LRE 2009

Alberto Abad[1] *and Isabel Trancoso*[1,2]

[1]L^{2}F - Spoken Language Systems Lab, INESC-ID / [2]IST, Lisboa, Portugal
{Alberto.Abad,Isabel.Trancoso}@l2f.inesc-id.pt

Abstract

The INESC-ID's Spoken Language Systems Laboratory (L^{2}F) Language Verification system submitted to the 2009 NIST Language Recognition evaluation is introduced in this paper. Then, as a sequence of the evaluation workshop and post-analysis of the results, the set of modifications that lead to significant performance gains are reported. Main differences between the original *submitted* system and the *post-evaluation* system consist of: 1) the kind and amount of training and development data considered for language model training and calibration and fusion, 2) the improvement of the acoustic based sub-systems and the reduction of the number of sub-systems that compose the whole system, and 3) the application of a better calibration and fusion scheme. Contrastive results of the *submitted* and the *post-evaluation* language recognition system for the different conditions in the evaluation are provided.

1. Introduction

The National Institute of Standards and Technology (NIST) has organized in the last years a series of evaluations in some relevant speech processing topics devoted to encourage language research activities.

In the 2009 NIST Language Recognition Evaluation (LRE09) the objective is to detect whether a target language is in fact spoken in a given speech segment. The number of possible target languages is 23. Three distinct test conditions are proposed depending on the possible set of competitive/non-target languages: "closed-set" (the set of non-target languages is the set of LRE09 target languages, minus the target language), "open-set" (the same as "closed-set", plus other "unknown" languages) and "language-pair" (the non-target language is a single language). Detailed information on the LRE09 campaign can be found in the evaluation plan document [1].

Language recognition (LR) approaches can generally be classified according to the kind of source of information that they rely on. The most successful systems are based on the exploitation of the acoustic phonetics, that is the acoustic characteristics of each language, or the phonotactics which are the rules that govern the phone combinations in a language.

This paper summarizes the LR system developed by the INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the LRE09 campaign and the post-evaluation efforts devoted to improve the LR system. Next Section 2 presents a description of the data used in this work. Section 3 describes the *submitted* language recognition system, starting by the phonotactic modules (subsection 3.1) and the acoustic ones (subsection 3.2). Then, the series of modifications and improvements introduced to the submitted system are described in Section 4. Finally, language verification results are provided for the *submitted* and for the *post-evaluation* systems for the different conditions.

2. Training, calibration and testing data

Data from previous evaluations and new data from Voice of America (VOA) radio broadcast [1] was made available for LR training and development.

2.1. Data for acoustic and phonotactic modeling

Language recognition acoustic models and phonotactic models used for the evaluation have been trained using *only* data from the VOA3 corpus provided for this evaluation.

For all target languages, approximately 15 hours of data from VOA3 automatically labeled as telephone data were extracted. Segments were classified according to their length in sets of approximately 30, 10 and 3 seconds. The number of files of each duration is approximately the same in every language.

A telephone band detector processing was applied to automatically classify the data for which this type of classification was not available. First, speech-non-speech segmentation was applied to the training data [2]. Then two scores were obtained, by averaging frame-based scores over the speech segment. The scores are band-energy ratios around 3400 Hz upper-bound of telephone band (similar to [3]) and 400 Hz lower-bound. Finally, the scores obtained for each speech segment were compared to fixed thresholds.

Notice that VOA3 includes data for all the 23 possible target languages of LRE09, except for the case of American English and Indian English that are not distinguished. We could find around 4.5 hours of Indian English in data sets of previous evaluations, but it was considered insufficient compared to the 15 hours used for all the other languages. Additionally, we were not very sure of the impact of using a different source of data just for one of the target languages. This was the motivation for using a unique set of data for training English models, both American and Indian without distinction.

Finally, an additional data set of approximately 15 hours was also extracted for "other" languages present in VOA3. This "other" languages data set was used to train phonotactic and acoustic models for a general language class corresponding to the "unknown" languages that are not part of the set of 23 possible target languages of this evaluation. Table 1 summarizes the data used for training the L^{2}F language recognition system.

2.2. Calibration and fusion data

Data from three different sources has been used for calibration and fusion of the LR system: VOA2 and VOA3 segments audited by LDC, VOA3 non-audited segments (like the ones of the training set, but different segments) and segments from previous

Lang	30	10	3	Tot
amha	688	685	685	2058 (14.7h)
bosn	647	657	657	1961 (14.1h)
cant	917	894	894	2705 (15.5h)
creo	792	788	788	2368 (14.7h)
croa	336	641	339	1316 (11.3)
dari	902	907	907	2716 (15.1h)
engl(*)	979	977	977	2933 (15.6h)
fars	643	634	634	1911 (14.2h)
fren	794	790	790	2374 (14.9h)
geor	664	2000	664	3328 (14.1h)
haus	764	759	759	2282 (14.7h)
hind	653	654	654	1961 (14.3h)
kore	994	998	998	2990 (15.8h)
mand	1094	1102	1102	3298 (16.1h)
pash	844	844	844	2532 (15.0h)
port	762	749	749	2260 (14.9h)
russ	636	649	649	1934 (14.4h)
span	550	545	545	1640 (13.9h)
turk	619	623	623	1865 (14.3h)
ukra	1085	1088	1088	3261 (16h)
urdu	696	704	704	2104 (14.5h)
viet	985	982	982	2949 (15.6h)
other	679	681	681	2041 (14.3h)
total	17723	19351	17713	54787 (338h)

Table 1: *Number of training speech segments extracted from the VOA3 corpus of each target language and total duration. (*) American English and Indian English are not distinguished.*

LRE evaluation sets. For every target language, approximately 4 hours of data have been selected and also split in 30 seconds, 10 seconds and 3 seconds segment duration.

Distinguished sets were used for American English and Indian English. Additionally, a set of approximately 6.9 hours of "other" languages (including the non-target languages of the training set and some additional ones) has been collected.

The total calibration and fusion corpus is composed of 19346 segments: 7815 of 30 seconds, 5911 of 10 seconds and 5620 of 3 seconds. A summary of this development data set is shown in Table 2.

2.3. Testing data

The LRE09 test set is used for LR assessment. The corpus is composed of 41793 speech segments: 14166 of 30 seconds, 13847 of 10 seconds and 13780 of 3 seconds.

3. The L^{2}F LRE *submitted* system

The complete L^{2}F language recognition system is the result of the fusion of eight language verification scores provided by 8 individual sub-systems: 4 phonotactic and 4 acoustic-based. In this section the 8 sub-systems and the calibration and fusion steps are described.

3.1. The PRLM-LR systems

The PRLM (Phone Recognition followed by Language Modeling) systems used for LRE09 exploit the phonotactic information extracted by four parallel tokenizers: European Portuguese, Brazilian Portuguese, European Spanish (Castilian) and American English. The tokenizers are MultiLayer Perceptrons (MLP) trained to estimate the posterior probabilities of the different phonemes for a given input speech frame (and its context).

Lang	LDC	VOA3	LREold	Tot
amha	1.4h	2.6h	—	4h
bosn	1.6h	2.5h	—	4.1h
cant	—	2.7h	1h	3.7h
creo	1.6h	2.6h	—	4.2h
croa	1.5h	2h	—	3.5h
dari	1.6h	2.6h	—	4.2h
engl.a	—	2h	2h	4h
engl.i	—	—	4h	4h
fars	—	2.5h	2h	4.5h
fren	1.6h	2.6h	—	4.2h
geor	1.1h	2.5h	—	3.6h
haus	1.6h	2.6h	—	4.2h
hind	—	2.5h	2h	4.5h
kore	—	2.9h	2h	4.9h
mand	—	2.8h	3h	5.8h
pash	1.6h	2.6h	—	4.2h
port	1.4h	2.6h	—	4h
russ	—	2.5h	3h	5.5h
span	—	2.5h	4h	6.5h
turk	1.6h	2.5h	—	4.1h
ukra	1.6h	2.8h	—	4.4h
urdu	—	2.5h	1h	3.5h
viet	—	2.8h	3h	5.8h
other	—	4.9h	2h	6.9h
total	18.2h	61.1h	29h	108.3h

Table 2: *Development data set composed of different data sources: audited VOA2 and VOA3 data (LDC), non-audited voa3 data (VOA3) and previous LRE data sets (LREold).*

3.1.1. *Feature extraction*

The system combines four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and the Advanced Font-End from ETSI (ETSI, 13 static + first and second derivatives).

3.1.2. *Phonetic tokenizers/classifiers*

For this evaluation, it was necessary to re-train our phonetic classifiers with Broadcast News (BN) data downsampled at 8kHz, since our original classifiers were developed for BN data at 16 kHz.

The European Portuguese classifier was trained with 57 hours of BN data, and 58 hours of mixed fixed-telephone and mobile-telephone data. The Brazilian Portuguese classifier was trained with around 13 hours of BN data. The Spanish classifier used 14 hours of BN data. Finally, the English system was trained with the HUB-4 96 and HUB-4 97 data sets, that contain around 142 hours of TV and Radio Broadcast data.

The size of the neural networks of each tokenizer differs due to the different amounts of training data. In the case of the output layer, its size corresponds to the number of phonetic units of each language, plus silence (no additional sub-phonetic or context-dependent units have been considered [4]).

3.1.3. *Phonotactics modeling*

For every phonetic tokenizer, the phonotactics of each target language is modeled with a 3-gram model. For that purpose the

SRILM toolkit has been used [5].

In both training and test, the raw phonotactic sequence obtained by each tokenizer was filtered, in order to avoid spurious phone recognitions. Concretely, phones that appeared only once in the middle of long sequences of identical phones were deleted.

3.2. The GSV-LR systems

Acoustic methods for LR are usually preferred to phonotactic approaches since they are not limited by the need of well-trained phonetic tokenizers. Recently, a method generally known as GMM supervectors (GSV) [6] has been shown to be a successful approach for both speaker verification and language verification tasks.

GSV-based approaches consist of a mapping of each speech utterance to a high-dimensional vector and the use of these high-dimensional vectors for training and classification with a support vector machine (SVM). The mapping to the high-dimensional space is the result of stacking in a single supervector the parameters (usually the means) of an adapted GMM to the characteristics of a given speech segment. In language recognition, a binary SVM classifier is trained for each target language with supervectors of the target language as positive examples and supervectors of other non-target languages as negative examples. During test, the supervector of the testing speech utterance must be also obtained and a score for each target language is obtained with the binary classifiers.

The four GSV-LR sub-systems that compose the complete L^2F language recognition system are slight variations of the GSV approach. Concretely, two of the GSV systems differ in the linear kernel considered (different normalization of the Gaussian mixture means in their projection to the high dimensional space). The last two systems are derivations of the previous GSV, where the SVM models parameters are pushed back to the GMM domain as proposed in [7].

3.2.1. *Feature extraction*

The extracted features are Perceptual Linear Prediction static features with log-RelAtive SpecTrAl speech processing (RASTA), and a stacked vector of shifted delta cepstra (SDC) of the same RASTA features. Concretely, 7 RASTA static features and a 7-1-3-7 SDC parameter configuration are computed, resulting in a final feature vector of 56 components.

3.2.2. *GMM UBM and SVM modeling*

A GMM universal background model of 256 mixtures was trained with approximately 20 hours of speech randomly selected from the 30 seconds training speech segments.

Five iterations of Maximum a posteriori (MAP) adaptation are performed for each speech segment to obtain the high-dimensional vector of size 56x256. Then, previously to SVM training (or classification) the high-dimensional vectors are normalized in two different ways resulting in two different GSV-SVM sub-systems.

Linear SVM classifiers are trained for each target language (and for the two different mean normalizations) with the libSVM toolkit [8]. For each target language, all the training segments/supervectors are used as positive examples. The negative examples were randomly selected among the training data from the other languages, in order to achieve approximately 1.2 times the number of positive examples.

3.3. Calibration and fusion

Linear logistic regression (LLR) fusion and calibration of the 8 sub-systems has been done with the FoCal Multiclass Toolkit [9]. For each evaluation condition ("closed-set", "open-set" and different "language-pairs"), a separate calibration and fusion has been trained for the 30, 10 and 3 seconds length segments.

In both the "closed-set" and "open-set" condition, the same score is used for both American and Indian English. However, notice that in the data used for calibration and fusion these varieties are distinguished. Thus, we expected that some discriminative information can be extracted from the relations with the other languages.

In addition to the models trained in the 8 sub-systems for the 23 different target-languages, an additional model for every system was trained with the "other" languages set. The score obtained by these models is used for representing the "unknown" language score in the "open-set" condition.

The scores obtained for the two languages of interest in the "language-pair" test condition were used to train fusion and calibration also with the FoCal Multiclass Toolkit.

4. The L^2F LRE *post-evaluation* system

After the evaluation Workshop and the analysis of the results, we focused on the improvement of the submitted system. In order to do that, we decided to apply simple modifications that did not essentially affect the architecture and the characteristics of the original language recognition system. Thus, the modifications were mainly aimed to correct some erroneous decisions (training data selection), to improve and reduce the number of GSV-LR sub-systems and to modify the calibration stage.

4.1. Training corpora selection

Data management –selection and filtering of the data for training and calibration– was even more important this year than in previous LRE editions due to the characteristics of the VOA corpus. Thus, selecting segments with a large variety of speakers was critical. It was shown during the evaluation that speaker clustering methods for rejecting frequent speakers was very convenient to assure speaker variability. Another issue related with the data was the relatively frequent presence of English in non-audited VOA3 data of any language.

In addition to these common problems, we noticed two errors in our submitted system related with the data. The first and more critical one is that the English data we selected from VOA3 was in fact not American as we expected, but it was English spoken by African speakers. Second, the decision of not training specific language models for Indian English resulted in an error since very poor performance was achieved detecting Indian with the submitted system.

Hence, in order to improve the quality of our training data set after the evaluation we decided to use a modified sub-set of the development data set of Table 2 for training. The reason is that the original development set was expected to be better since a larger amount of audited data was included (reducing miss-labeling errors), the speakers diversity was augmented, data from different sources than VOA3 was included and specific data for Indian English was available for language training. Additionally, the original "American English" segments of VOA3 were replaced with real "American English" speech (instead of English of African speakers). Finally, the *post-evaluation* training set was a random selection of 120 segments of 30 seconds duration per each language. Notice the difference on the amount of training data (1 hour per language) compared to the original set of Table 1.

4.2. GSV-LR systems improvements

The performance of the *submitted* LR system was mainly ruled by the phonotactic systems. It is for that reason that we focused on improving the acoustic based sub-systems.

4.2.1. Silence rejection and normalization

Low-energy frame rejection and mean and variance feature normalization was incorporated to the GSV-LR front-end described in 3.2.1. Silence segmentation is obtained with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment.

4.2.2. Number of mixtures and sub-systems reduction

First, the number of Gaussian mixture components was increased from 256 to 1024. Second, it was verified that the use of the two different kernels (supervector normalizations) did not provide any noticeable performance improvement. It was for that reason that we decided to keep only the GSV systems based on the Kullback-Leibler (KL) divergence [6] in the *post-evaluation* system. Thus, the total number of sub-systems of the *post-evaluation* system is reduced to six.

4.2.3. NAP for channel compensation

The Nuisance Attribute Projection (NAP) approach [10] is a compensation method aimed at removing nuisance attribute-related dimensions in high-dimensional spaces via projections. In the *post-evaluation* system we apply NAP to the conventional GSV approach (not the pushing-back scoring method). NAP projections were trained with a sub-set of the *post-evaluation* training set (50 segments per language). We used a nuisance space of dimension 32.

4.3. Gaussian back-end and development data

In the *post-evaluation* LR system the scores of each individual sub-system are processed by a Gaussian back-end prior to the LLR calibration and fusion. Separate back-ends are also trained with the FoCal toolkit for every evaluation condition and segment length.

Since it was used part of the original development data for language model training, we decided to use the evaluation test set for back-end and LLR training through a random 5-fold cross-validation process.

5. Language recognition results

The L^2F *submitted* system to the LRE09 competition is compared to the *post-evaluation* system described in previous Section 4 . Additionally, for better comparison purposes, a new calibration for the *submitted* system has been trained. Like in the *post-evaluation* system a random 5-fold cross-validation strategy using the test data is applied for training the LLR calibration and fusion of the eight sub-systems. Results for this new calibrated LR system are reported as *submitted**. Average cost LR performances (as defined in [1]) are shown in Table 3.

First it should be noticed that a considerable improvement – ranging from 12% to more than 20% – is achieved due to the *optimistic* calibration process involving the testing data. However, great performance gains are still achieved as a result of the improvements introduced in the *post-evaluation* system in all conditions and categories. These are particularly noticeable for longer segment durations. For instance, relative cost reductions of 46.7% and 36.9% are obtained for 30 seconds duration in the closed and open-set conditions respectively. The use of only 30 seconds segments in the training set of the *post-evaluation* system might partially explain these results. Table 3 also shows a generalized higher relative improvement of the *post-evaluation* system for the closed-set than for the open-set condition.

C	T	submit	submit*	post-eval
closed	30	0.0407	0.0346 (15.0%)	0.0217 (46.7%)
	10	0.0781	0.0618 (20.9%)	0.0517 (33.8%)
	3	0.1692	0.1430 (15.5%)	0.1377 (18.6%)
open	30	0.0582	0.0507 (12.9%)	0.0367 (36.9%)
	10	0.0935	0.0792 (15.3%)	0.0673 (28.0%)
	3	0.1865	0.1590 (14.7%)	0.1513 (18.9%)

Table 3: *Average cost performance for each of the three segment duration categories (T), and for the closed-set and open-set conditions (C). Relative performance improvements with respect to the submitted system are shown in brackets.*

6. Summary and conclusions

Improvements to the L^2F Language Recognition system submitted to the NIST LRE 2009 campaign have permitted remarkable recognition gains for all evaluation categories and conditions, achieving comparable performances to the best present LR systems. Particularly noticeable improvements have been obtained in the closed-set 30 seconds segment duration condition with an average cost performance of 0.0217. It is worth to mention that the *post-evaluation* system makes use of only 24 hours of data for language models training ($\sim$1 hour per target language) in contrast to the 338 hours of the *submitted* system.

7. References

[1] "The 2009 NIST Language Recognition Evaluation Plan (LRE09)", URL: http://www.itl.nist.gov/iad/mig/tests/lre/2009/.

[2] Meinedo, H. and Neto, J., "Audio Segmentation, Classification and Clustering in a Broadcast News Task", in Proc. ICASSP 2003, Hong Kong, Apr 2003.

[3] Plchot, O. et al, "Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition: Technical Report", URL: http://www.nist.gov/speech/tests/lre/2009/radio_broadcasts.pdf.

[4] Abad, A. and Neto, J., "Incorporating Acoustical Modelling of Phone Transitions in an Hybrid ANN/HMM Speech Recognizer", in Proc. INTERSPEECH-08, Brisbane, Australia, Sep 2008.

[5] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", in Proc. ICSLP 2002, Denver, Colorado, Sep 2002.

[6] Campbell, W. M., Sturim, D. E. and Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification" IEEE Signal Process. Letters, vol. 13(5), pp. 308-311, 2006.

[7] Campbell, W. M. , "A covariance kernel for SVM language recognition", in Proc. ICASSP 2008, Las Vegas, USA.

[8] Chang, C.-C. and Lin, C-J, "LIBSVM - A Library for Support Vector Machines", URL: http://www.csie.ntu.edu.tw/ cjlin/libsvm/index.html.

[9] Brummer, N., "FoCal Multiclass Toolkit", URL: http://niko.brummer.googlepages.com/focalmulticlass.

[10] Campbell, W.M. et al, "SVM based Speaker Verification using GMM Supervector Kernel and NAP Variability Compensation", in Proc. ICASSP 2006, Toluse, France, May 2006.

On the Improvement of Speaker Diarization by Detecting Overlapped Speech

Martin Zelenák, Javier Hernando

Universitat Politècnica de Catalunya, Barcelona, Spain
{martin.zelenak,javier.hernando}@upc.edu

Abstract

Simultaneous speech in meeting environment is responsible for a certain amount of errors caused by standard speaker diarization systems. We are presenting an overlap detection system for far-field data based on spectral and spatial features, where the spatial features obtained on different microphone pairs are fused by means of principal component analysis. Detected overlap segments are applied for speaker diarization in order to increase the purity of speaker clusters and to recover missed speech by assigning multiple speaker labels. Investigation on the relationship between overlap detection properties and diarization improvement revealed very distinct behaviour of overlap exclusion and overlap labeling.

Index Terms: speaker overlap detection, speaker diarization

1. Introduction

Spontaneous human conversation very often includes certain amount of simultaneous speech. This naturally occurring phenomenon is typical for meeting environment, where listening people for example interrupt the leading speaker in order to grab floor or give backchannel to encourage his talk. Some speaker overlaps can also be the result of elevated emotions (laughing, arguing). Shribergh [1] observed that the amount of overlapped speech is not necessarily dependant on the amount of people involved in the conversation, ergo, few people can produce significant overlap too. Overlapped speech poses a problem for many automatic human language technologies, including speaker diarization, which, given a recording, strives to answer the question *"Who spoke when?"*. In general, no prior knowledge about the speakers is provided. Conventional diarization systems are able to assign only one speaker label per segment, which, obviously, leads to missed speech for overlapped speakers. Furthermore, including overlapped speech into the model building can be a potential source of speaker error, since the models could be corrupt.

In previous works, the common way to detect speaker overlap in meetings was to segment each of the individual speaker channels with an ergodic hidden Markov model (HMM). Overlapped speech was either one of the decoding classes [2] or was marked in a post-processing algorithm [3]. The solution suggested in [4] eliminated the necessity to train any model.

Some published algorithms focused on distant microphone channels exclusively. Knowing the number of speakers beforehand and assuming their location will not change during a multiparty conversation, the authors in [5] proposed to use microphone pair time delays (TDE) to segment audio according to speakers. They showed the possibility to detect two simultaneous speakers by modeling short-term speaker turns for every pair of the assumed speakers with an HMM.

Explicit modeling of all pairs of speakers after an initial diarization was also explored in [6]. Even though the authors claim to be able to detect overlap, it did not lead to a reduction of the diarization error. Improvement of speaker diarization by handling overlaps detected with an HMM-based detection system was firstly presented in [7] on a subset of the AMI corpus.

In this paper we are presenting an overlapped speech detection system for distant channel microphones, which successfully combines spectral and spatial features. To deal with the high and variable dimensionality of spatial feature space we are suggesting the application of principal component analysis (PCA), which fuses feature vectors across different microphone pairs into one spatial feature set. A similar approach was also chosen for diarization purposes in [8].

Our motivation for detecting overlapped speech is to improve a baseline diarization system with two techniques. In the first, also referred to as overlap exclusion, overlaps shall be discarded from training cluster models, hoping to achieve a more precise segmentation. The second technique makes it possible to assign two speaker labels in segments with simultaneous speech. We examined the behaviour of these two techniques in the context of changing overlap detection properties more in detail and it results to be substantially different for labeling as for exclusion. Experiments were conducted on single- and multi-site recordings from the AMI Meeting corpus.

The remainder of this paper is organized as follows. Speaker overlap detection and speaker diarization system with overlap handling improvements are outlined in Sections 2 and 3, respectively. Experimental results are discussed in Section 4 and conclusions are given in Section 5.

2. Speaker overlap detection

2.1. Spectral features

The overlap detection system uses several spectral-based features which were identified as to be conveying some overlap information. Cepstrum is successfully applied for a handful of speech related tasks and constitutes a good basis of a feature set, for that reason 12 MFCCs were extracted every 10 ms over a window of 30 ms.

Linear predictive coding (LPC) analyzes the speech signal by estimating the formants of a speaker. It is assumed that LPC of a reasonably chosen order can model the spectrum of a single speaker quite well, but will fail for a region with multiple speakers [9]. In the latter case, more energy will be left in the residual signal, which represents prediction error. In this system, residual energy of a 12th-order LPC (LPCRE) was computed over a 25 ms window. The feature set was furthermore extended with first order delta coefficients and all features were mean-variance normalized according to statistics obtained from training data.

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is supported by a grant from the Catalan autonomous government.

Another spectral-based feature is the spectral flatness (SF) extracted over a window of 30 ms. This feature was applied for discrimination between speech and non-speech [10], but can eventually convey information about the number of speakers speaking. It is defined as the ratio between geometric and arithmetic mean of a certain number (100 in our case) of spectral magnitudes

$$SFM_{dB} = 10\log_{10}\frac{\sqrt[N]{\prod_{i=0}^{N-1} mag(i)}}{\sum_{i=0}^{N-1} mag(i)}. \tag{1}$$

2.2. Spatial features

Several spatial features based on cross-correlation were introduced to improve spectral overlap detection on distant channel data. The first spatial feature is the value of the principal cross-correlation peak, which is a measure of *coherence* between signals. For a pair of microphones *i* and *j* it is defined as

$$C_{ij} = \max(R_{ij}(\tau)), \tag{2}$$

where $R_{ij}(\tau)$ is the Generalized Cross Correlation with Phase Transform weighting (GCC-PHAT) [11] that is often used in order to improve robustness in reverberant environments. Ideally, the coherence value should be high for single-source situations and low if noises, reverberation or concurrent acoustic sources are present. In the general case, the main peak is attenuated when multiple sources introduce random peaks.

In situations dealing with multiple, possibly moving, concurrent speakers it was observed that time delay estimates (TDEs) produced by the GCC-PHAT jump from one speaker to another at a very high rate as one source dominates due to the non-stationarity of the voice. Thus, the first order derivative of the time delay estimate *delta TDE* is expected to carry certain degree of information on overlaps. TDE is defined as

$$\hat{\tau}_{ij} = \operatorname*{argmax}_{\tau} R_{ij}(\tau). \tag{3}$$

Derived from the coherence value, we are also proposing to extract the coherence *dispersion ratio*, as shown in eq. 4. This value is computed as the relation of the square of main peak value and the sum of secondary peaks square values corresponding to other acoustic sources that may be present in the scenario,

$$D_{ij} = \frac{C_{ij}^2}{\sum_{t=-w_{ij}}^{w_{ij}} R_{ij}^2(\hat{\tau}_{ij}+t)}, \tag{4}$$

where the size of the window w_{ij} is adjusted in accordance with the TDE standard deviation of microphone pair *(i,j)*.

The dimensionality of a spatial-feature vector can be very high since we extract three features for every microphone pair. Furthermore, the number of microphones differs from site to site, making it difficult to train a general model. In order to deal with these problems, we are proposing to unify and reduce the number of microphone pairs with a PCA transformation. For each discussed spatial feature and for every site we estimated a transformation matrix and then applied just the first component. Consequently, we obtained three transformed features (coherence, dispersion, delta TDE) for each frame.

2.3. System architecture

A schematic block diagram of the overlap detection system with link to speaker diarization is given in Fig. 1. The system considers three acoustic classes representing non-speech, single-speaker speech and overlapped speech. For each class a three-state HMM is defined where every state is modeled with a GMM with diagonal covariance. Since the amount of training data is not balanced for all classes, we are using 256 Gaussian components for single-speaker speech and 32 or 64 components for overlapped speech and non-speech. GMMs are created by iterative Gaussian splitting and subsequent re-estimation. Spectral and spatial features are modeled with separate GMMs, with output probability calculated with feature stream weights of 0.75 and 0.25, respectively. In the case of spatial GMMs, the means and variances are shared across the three states.

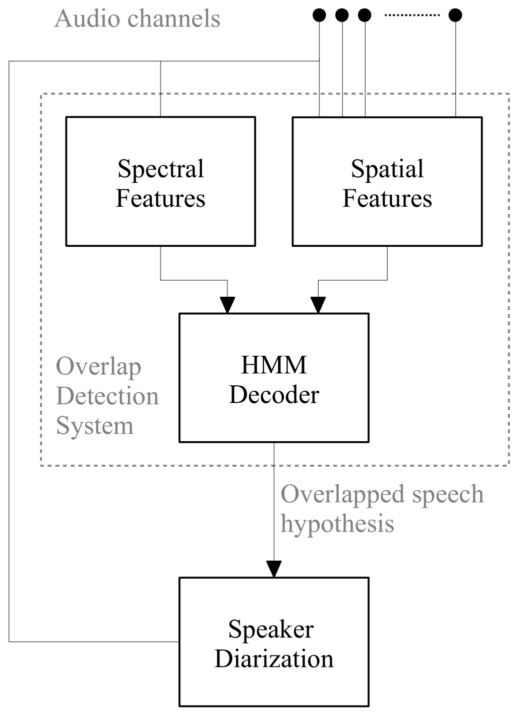


Figure 1: *Overlap detection system block diagram*

The detection hypothesis is obtained by Viterbi decoding and applying a word network. For precision purposes the transition from single-speaker speech to overlapped speech can be penalized with an overlap insertion penalty (OIP) and direct transitions between non-speech and overlapped speech are completely forbidden.

3. Speaker diarization system

Our speaker diarization system, detailed in [12], follows the commonly used agglomerative clustering approach. In the beginning, speech is broken into rather short uniform segments and the successive clustering stage groups acoustically similar segments and assigns them to speaker clusters. The number of initial clusters is determined automatically from audio length with minimal and maximal value constraints. Clusters are modeled with GMMs and cluster pair merging in each iteration is driven by Bayesian information criterion (BIC). The system operates with 20 MFCCs extracted from 30 ms frames. The performance of diarization is evaluated by means of diarization error rate (DER), which is the sum of missed seech rate, false alarm rate and speaker error rate.

Overlap handling in diarization comprises the exclusion and/or labeling of simultaneous speech. The first technique blocks overlap frames from being included into cluster initialization and GMM training, but does not prevent decoding them. The aim of this technique is to get lower speaker detection error rates with more precise clusters. Overlap labeling technique seeks to select the two most likely clusters in Viterbi decoding instead of only one. In this way the missed speaker time should be decreased.

In order to evaluate just the impact of overlapped speech on speaker segmentation, detected overlaps are masked with reference speech/non-speech segments before given to diarization system. The diarization system is using reference speech segments as well.

4. Experiments

4.1. Database and experimental setup

The database used for our experiments was the AMI Meeting corpus, which comprises 100 hours of meeting recordings. We were working with far-field microphone array channels sampled at 16 kHz. We defined single- and multi-site scenarios. The first included recordings only from Idiap site and the latter also from Edinburgh and TNO sites. The recordings were then divided into training set (22 for both single- and multi-site scenario), development set (3 and 9) and evaluation set (11 and 10). The average amount of overlapped speech in these scenarios was 14.40% and 15.10%, respectively. Training and evaluation of the overlap detection system are performed with forced-alignment annotations obtained by the SRI's DECIPHER recognizer.

In the presented experiments, we are comparing the results obtained with two feature setups for the detection of speaker overlap. The first one is a baseline spectral system (*Spct*) and the second is a system based on the combination of spectral and PCA-transformed spatial features (*Spct+Spat*). Overlap detection performance is measured with Recall—ratio between true detected and reference overlap time, Precision—ratio between true and all detected overlaps, and with Error—the sum of missed and false overlaps divided by reference overlap time. Results depend very much on the value of the overlap insertion penalty, which controls the amount of overlaps the system will posit. It can be perceived as a compensation for an undertrained model. Initially, four values of OIP were defined based on different detection characteristics on development data, accounting for hypotheses with the highest recall, the highest F-ratio, the lowest error rate and an acceptably high precision.

It is assumed that hypotheses exhibiting high recall will be suitable for overlap exclusion, because as many overlaps as possible will be discarded from model building. On the contrary, high precision and low error will be needed for successful overlap labeling, since all of the false overlaps will be propagated to DER, but only a perfect labeling would transform all true overlaps into reduction of missed speaker time.

Obviously, the nature of the two techniques is very different. Therefore, it is not useful to use necessarily the same overlap hypohesis for both, but rather two independent overlap hypotheses, one for each technique.

4.2. Single-site experimental results

The DER relative improvements of handling overlapped speech over the diarization baseline for single-site recordings are given in the right column of Table 1. We can see that the overall better *Spct+Spat* overlap detection hypothesis used for exclusion resulted in higher DER improvement than the *Spct* hypothesis. In the case of labeling, the hypothesis with lower error (*Spct+Spat*), though lower precision as well, gained more improvement as the other (*Spct*). The highest improvement of 11.6% was achieved by applying overlaps detected by the combined spectral and spatial system.

Table 1: *Speaker diarization with excluding and/or labeling overlapped segments on single-site evaluation data. Overlap detection recall, precision, error and corresponding DER rel. improvement over the baseline (all values in %)*

Baseline DER					38.3
Overlap det.:	**Rcl.**	**Prc.**	**Err.**	**DER rel. imp. [%]**	
Spct	45.7	52.2	96.1	+Excl.	+3.9
	27.0	83.9	78.2	+Labl.	+4.9
	”	”	”	+Both	**+6.9**
Spct+Spat	49.2	59.0	85.0	+Excl.	+5.2
	35.4	80.5	73.2	+Labl.	+5.5
	”	”	”	+Both	**+11.6**

Table 2: *Speaker diarization with excluding and/or labeling overlapped segments on multi-site evaluation data. Overlap detection recall, precision, error and corresponding DER rel. improvement over the baseline (all values in %)*

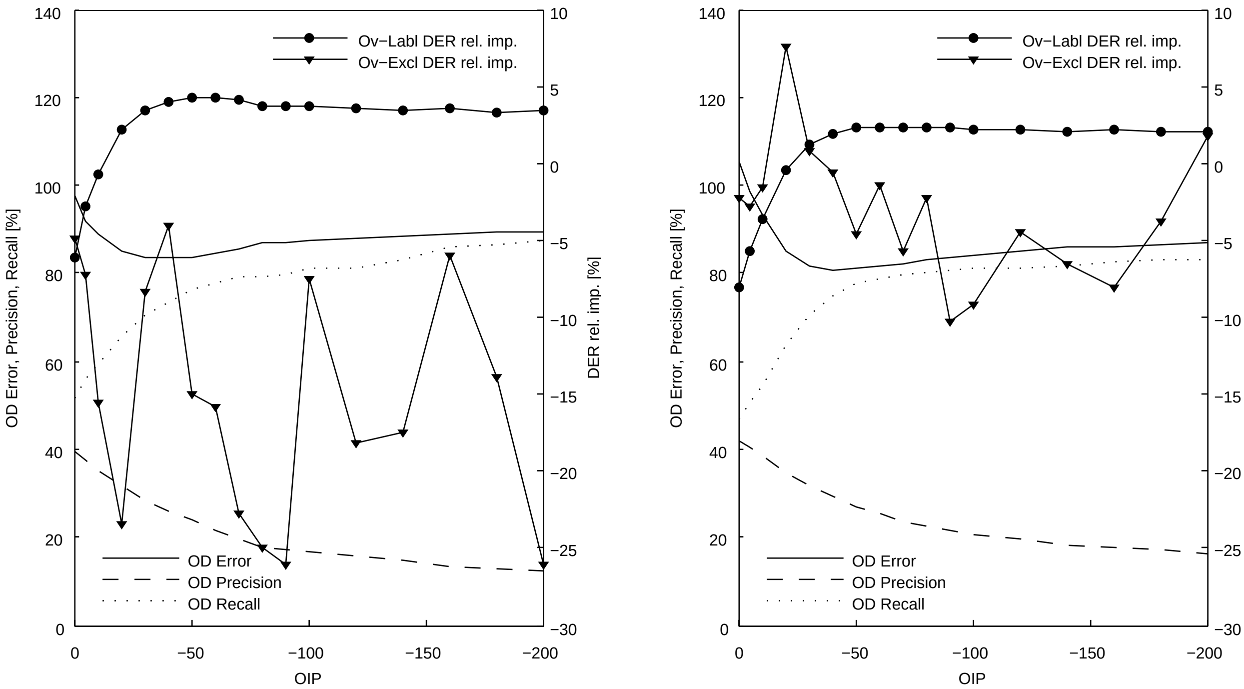
Baseline DER					37.3
Overlap det.:	**Rcl.**	**Prc.**	**Err.**	**DER rel. imp. [%]**	
Spct	49.5	41.8	119.5	+Excl.	+7.5
	25.4	74.9	83.1	+Labl.	+2.1
	”	”	”	+Both	**+10.2**
Spct+Spat	59.3	42.3	121.5	+Excl.	+6.9
	30.4	70.5	82.3	+Labl.	+1.7
	”	”	”	+Both	**+9.5**

4.3. Multi-site experimental results

The overlap detection results in multi-site scenario, given in Table 2, are considerably worse than in the case of single-site data. Despite this worse overlap detection performance, excluding overlapped segments led to surprisingly higher relative DER improvements. More expected are the modest improvements by labeling. The best improvement of up to 10.2% is obtained with (*Spct*) hypotheses. The lower performance of spatial setup could be eventually explained by the fact that the spatial features are, in general, not commensurable accross different microphone pairs, since they are tied to physical characteristics of a particular pair. The PCA-based transformation of features from multiple sites probably lacked some robustness in this case.

4.4. Overlap detection and diarization improvement

In order to investigate more on the relationship between overlap detection performance and the obtained improvements in diarization, we performed a further set of experiments on single- and multi-site development data. A large number of overlap hypotheses produced with several overlap insertion penalties was employed for exclusion and labeling.



(a) Single-Site (b) Multi-Site

Figure 2: *Overlap detection performance of spectral system and corresponding speaker diarization relative DER improvement over the baseline by excluding and labeling detected overlap segments for (a) single- and (b) multi-site development data.*

The detection metrics, i.e., recall, precision and error, and the corresponding DER improvements are given in Fig. 2 (a) and (b). Note that the peak of labeling performance lies within the region of lowest detection error and is also somewhat shifted to the right towards higher precisions. This observation is in compliance with our former assumption and the fact that the complement of the error tells us how much we can theoretically gain by assigning second labels. Overlap exclusion exhibits a less predictable behaviour in terms of DER improvements, making it difficult to derive any kind of conclusion at this point. Still, the results from Tables 1 and 2 show that the improvements can be significant.

5. Conclusions

We have presented an overlap detection system based on spectral and spatial features. Detected overlaps were used in speaker diarization for increasing the purity of speaker models and to recover missed speech by assigning multiple speaker labels. Experiments on evaluation single- and multi-site data showed improvements over baseline diarization system. Further investigation using held-out data revealed the changeable nature of improvements by overlap exclusion, but also confirmed the labeling performance's dependence on overlap precision and error.

6. References

[1] Shriberg, E., "Spontaneous Speech: How People Really Talk and Why Engineers Should Care," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005, pp. 1781–1784.

[2] Wrigley, S.N. et al., "Speech and Crosstalk Detection in Multichannel Audio," *IEEE Transactions on Speech and Audio Processing*, vol 13. pp. 84–91, 2005.

[3] Pfau, E., Ellis, D.P.W and Stolcke, A., "Multispeaker Speech Activity Detector for the ICSI Meeting Recorder," in *Proc. ASRU '01*, Madonna di Campiglio, Italy, 2001, pp. 107–110.

[4] Laskowski, K., Jin, Q. and Schultz, T. "Crosscorrelation-based Multispeaker Speech Activity Detection," in *Interspeech '04*, Jeju Island, Korea, 2004, pp. 973–976.

[5] Lathoud, G. and McCowan, L., "Location Based Speaker Segmentation," in *Proc. ICME '03*, Baltimore, USA, 2003, pp. III-621–4 vol.3.

[6] van Leeuwen, D.A. and Huijbregts, M., "The AMI Speaker Diarization System for NIST RT06s Meeting Data," in *Machine Learning for Multimodal Interaction*, LNCS, vol. 4299/2006, Springer Berlin/Heidelberg, 2006, pp. 371–384.

[7] Boakye, K., Trueba-Hornero, B., Vinyals, O., and Friedland, G., "Overlapped Speech Detection for Improved Speaker Diarization in Multiparty Meetings," in *Proc. ICASSP '08*, Las Vegas, USA, 2008, pp. 4353–4356.

[8] Otterson, S., "Improved Location Features for Meeting Speaker Diarization," in *Proc. Interspeech '07*, Antwerp, Belgium, 2007, pp. 1849–1852.

[9] Sundaram, N. et al., "Usable speech detection using linear predictive analysis - a model based approach," in *Proc. of ISPACS*, Awaji Island, Japan, 2003, pp. 231–235.

[10] Yantorno, R., "The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Emplyed as a Co-Channel Detection System," in *Proc. of IEEE Workshop on Intelligent Signal Processing*, 2001.

[11] Brandstein, M. S. and Silverman, H. F., "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. ICASSP '97*, Munich, Germany, 1997, pp. 375–378.

[12] Luque, J. et al., "Speaker Diarization for Conference Room: The UPC RT07s Evaluation System," in *Multimodal Technologies for Perception of Humans*, LNCS, vol. 4625/2008, Springer Berlin/Heidelberg, 2008, pp. 543–553.

Low-Latency Speaker Tracking and SOA-Compliant Services for Ambient Intelligence Environments

Maider Zamalloa [1,2], *Luis Javier Rodríguez-Fuentes* [1], *Germán Bordel* [1], *Mikel Penagarikano* [1], *Jorge Parra* [2], *Aitor Uribarren* [2], *Juan Pedro Uribe* [2]

[1] Department of Electricity and Electronics, University of the Basque Country, Spain
[2] Ikerlan - Technological Research Centre, Spain
{maider.zamalloa, luisjavier.rodriguez, german.bordel, mikel.penagarikano}@ehu.es, {jparra, auribarren, JPUribe}@ikerlan.es

Abstract

As the most natural interface for human interaction, speech can be exploited to track users and then customize services as they get available. Low latency is required, since adaptation to user profiles must be done in a continuous fashion. However, most speaker tracking approaches found in the literature work offline, fully processing pre-recorded audio files by means of a two-stage procedure involving acoustic segmentation and speaker detection. In this work, a low-latency online speaker tracking approach is applied, which deals with continuous audio streams and outputs a decision at fixed intervals, by scoring fixed-length audio segments with regard to a set of target speaker models. Experimental results are reported on the AMI Corpus of meeting conversations, revealing the effectiveness of the proposed approach with regard to a traditional approach working offline. A speaker tracking service and a lower-level auxiliary speaker detection service have been also designed, based on the online low-latency speaker tracking approach mentioned above. These services are SOA-compliant and provide an interoperable, reusable and easily evolvable means to develop SOA-based speaker tracking applications for Ambient Intelligence (AmI) environments.

Index Terms: low-latency, speaker tracking, Service Oriented Architecture, Ambient Intelligence.

1. Introduction

In Ambient Intelligence (AmI) environments, human-computer interaction must be driven by intelligent and natural interfaces. Speech is a natural interface for human interaction and can be exploited to extract user related information such as location, identity, emotional state, etc. Speech is also a suitable means to support user adaptation. User adaptation must be done in a continuous fashion, which requires users to be continuously tracked (identified and located) in the AmI environment, so that customized services can be provided to all of them.

Speaker diarization and speaker tracking are well known speech processing tasks which aim to answer the question *Who spokes when?*, that is, to detect speaker turns in a continuous audio stream. Speaker tracking aims to detect segments corresponding to a known set of target speakers [1], whereas speaker diarization aims to detect speakers without any prior knowledge about them [2][3][4]. There are three primary application domains for speaker tracking and diarization: broadcast news audio, recorded meetings and telephone conversations. The methodologies applied in such domains assume that audio recordings are fully available before processing. So, common approaches to speaker tracking and diarization consist of two steps applied offline: (1) audio segmentation and (2) speaker detection. In speaker diarization, segments hypothetically uttered by the same speaker are clustered together and assigned the same label. In speaker tracking, once the audio stream is segmented, speaker detection is carried out through classical speaker recognition techniques [5][6][7]. In any case, these methodologies are not suitable for low-latency online speaker detection.

This paper presents results from our basic research on low-latency online speaker tracking, and describes tools which help shortening the deployment time of speaker tracking applications. Both elements were designed for an AmI scenario, for example an intelligent home environment, where the system continuously tracks known speakers (users), and the expected number of target speakers is low (i.e. the members of a family). As noted above, this scenario requires taking almost instantaneous (low latency) speaker tracking decisions.

The speaker tracking approach applied in this work jointly performs audio segmentation and speaker detection, by defining and processing fixed-length audio segments and scoring each of them to decide whether it belongs to a target speaker or to an impostor.

The performance of the proposed approach is compared to that of an offline system developed for reference, which follows a two-stage uncoupled approach. Speaker tracking experiments applying both systems were carried out on the AMI Corpus (Augmented Multi-party Interaction) [8], which contains human conversations in the context of smart meeting rooms, close to the AmI scenario described above.

From a practical point of view, the main contribution of this work regards the design of services helping the deployment of speech-based speaker tracking applications in a Service Oriented Architecture (SOA) framework [9][10]. SOA-based systems provide services to either end-user applications or to other services distributed in a network, via published and discoverable standard interfaces. SOA promotes the loose coupling between software components published as services, so they can be combined by service composition and reused in many applications. In addition, interoperability is also achieved, since services are neither dependent on the platform nor the programming language. In this work, two SOA-compliant UPnP services have been defined (a speaker tracking service and a lower-level auxiliary speaker detection service) based on the low latency real time speaker tracking approach described above.

The rest of the paper is organized as follows. In section 2, the main features of the online and offline speaker tracking systems are described, including speaker detection, score calibration and score smoothing. Section 3 gives details about the experimental setup. Section 4 presents and briefly discusses results attained in speaker tracking experiments. Section 5 describes the SOA-compliant speaker tracking services. Finally, conclusions are summarized in section 6.

2. Speaker tracking systems

2.1. Speaker detection

The online speaker tracking system applied in this work computes a detection score per target speaker and outputs a speaker identification decision at fixed-length intervals. That length has been empirically set to one second, which provides relatively good time resolution and spectral richness, and a reasonably small latency for most online speaker tracking scenarios. The offline system developed for reference does the same computation, but using the segments produced by an audio segmentation algorithm [11]. Regardless the way audio segments are obtained, scores are computed by means of acoustic models (corresponding to target speakers) estimated via Maximum A Posteriori (MAP) adaptation of a Universal Backround Model (UBM) [12]. Besides yielding good speaker recognition performance, the MAP-UBM methodology allows for a fast scoring technique which speeds up the score computation.

Acoustic vectors consist of 12 Mel-Filter Cepstral Coefficients (MFCC) + 12 ΔMFCC. Given an acoustic observation X (consisting of a sequence of acoustic vectors), the acoustic model λ_s for the target speaker s and the UBM, λ_{UBM}, the detection score $\Delta_s(X)$ is computed as follows:

$$\Delta_s(X) = L(X|\lambda_s) - L(X|\lambda_{UBM}) \qquad (1)$$

where $L(X|\lambda)$ is the log-likelihood of X given λ. Once the detection scores are computed for all the target speakers, speaker detection can be accomplished according to two possible approaches:

1. In the *speaker identification (SI) approach*, X is marked as coming from the most likely target speaker $s^* = \arg\max_{s\in[1,S]}\{\Delta_s(X)\}$, if $\Delta_{s^*}(X) > \theta_I$. Otherwise X is marked as coming from an impostor. The decision threshold θ_I can be heuristically established to optimize the discrimination among target speakers. Note that, for any given segment X, there could actually be two or more speakers speaking at the same time. However, the detection approach described above cannot inform of speaker overlaps, because only the most likely speaker can be detected.

2. In the *speaker verification (SV) approach*, each target speaker s is accepted or rejected by comparing the detection score $\Delta_s(X)$ to a decision threshold θ_V. If $\Delta_s(X) > \theta_V$ the target speaker s is accepted; otherwise it is rejected. This approach allows to handle segments with overlapped speech, since all the target speakers for which $\Delta_s(X) > \theta_V$ are accepted.

2.2. Score calibration

Calibration maps detection scores $\{\Delta_s \mid s\in[1,S]\}$ to likelihood ratios $\{C(\Delta_s) \mid s\in[1,S]\}$ without any specific application in mind. The scaling parameters are computed over a development corpus by maximizing *Mutual Information*, which is equivalent to minimizing the so called C_{LLR} (a metric defined in [13]), which integrates the expected cost over a wide range of operation points. The final decision is taken by applying the minimum expected cost Bayes decision threshold to calibrated scores $C(\Delta)$. The target speaker is accepted only if the following inequality holds:

$$C(\Delta) \geq \ln\left(\frac{C_{fa}(1-P_{target})}{C_{miss}\,P_{target}}\right) \qquad (2)$$

where C_{miss} and C_{fa} are miss and false-acceptance error costs, and P_{target} the prior probability of target speakers. Scores are calibrated by means of the *FoCal toolkit*, applying a linear mapping strategy (see http://www.dsp.sun.ac.za/~nbrummer/focal/).

2.3. Score smoothing

Since speaker detection is done for very short (one-second length) segments, the performance of the low-latency online speaker tracking system may degrade due to local variability. To increase the robustness to such variability, information from previous segments can be taken into account, that is, the acoustic scores of target speakers may be computed on speech segments lasting more than one second. Assuming that no speaker change takes place in the previous segments, scores will be more accurate as more samples are used to compute them. On the other hand, this does not affect the online processing and low-latency decision-making constraints. In practice, a smoothed score is computed by linearly combining the scores of the last *w* (one-second length) segments, weighting them according to rectangular (uniform) or triangular (linearly decreasing as going back in time) functions.

3. Experimental setup

3.1. The AMI Corpus

Experiments were carried out on the AMI Corpus of meeting conversations (http://corpus.amiproject.org/). The AMI Corpus is a multimodal dataset concerned with real-time human interaction in the context of smart meeting rooms. Data, collected in three instrumented meeting rooms, include a range of synchronized audio and video recordings. Meetings contain speech in English, mostly from non native speakers.

In this work, data for train, development and evaluation of speaker tracking systems were taken from a subset of the AMI Corpus, the Edinburgh scenario meetings, including 15 sessions: ES2002-ES2016, with four meetings per session, each meeting being half an hour long on average. The audio stream is obtained by mixing the signals from the headset microphones of the speakers. Three of the four speakers participating in each session are taken as target speakers, the remaining one being assigned the role of impostor. Careful impostor selection –not random– is made to account for gender unbalanced sessions. In sessions containing just one female speaker, the impostor is forced to be male (and vice versa), in order to avoid that gender favors impostor discrimination.

In order to assess the speaker tracking performance in realistic conditions, two independent subsets are defined, consisting of different sessions (and therefore different speakers), for development and evaluation purposes, respectively. The development set, consisting of 8 sessions (32 meetings), is used to tune the configuration parameters of the speaker tracking systems. The evaluation set, including the remaining 7 sessions (28 meetings), is used only to evaluate the performance of the previously tuned speaker tracking systems. Both the development and evaluation subsets are further divided into train and test datasets. Two meetings per session are randomly selected to estimate the UBM and the speaker models, and the remaining two are left for testing purposes. Time references are based on manual annotations provided in the AMI Corpus.

3.2. Performance measures

The performance of speaker tracking systems is commonly analyzed by means of *Detection Error Tradeoff* (DET) plots [14]. Performance is measured in terms of time that is correctly or incorrectly classified as belonging to a target. Therefore, miss and false alarm rates are computed as a function of time [1] and not as a function of trial number, like in speaker detection experiments. DET performance can be summarized in a single figure by means of the *Equal Error*

Rate (EER), the point of the DET curve at which miss and false alarm rates are equal. Obviously, the lower the EER, the higher the accuracy of a speaker tracking system.
Another way to summarize in a single figure the performance of a speaker tracking system is the so called *F-measure*, defined as follows:

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (3)$$

where precision (P) and recall (R) are related to false alarm and miss rates respectively. Precision measures the correctly detected target time from the total target time detected. Recall computes the correctly detected target time from the actual target time. The F-measure ranges from 0 to 1, with higher values indicating better performance. Collar periods of 250 milliseconds at the end of speaker turns are ignored for scoring purposes. Thus, speaker turns of less than 0.5 seconds are not scored.

4. Speaker tracking experiments

4.1. Online vs. offline systems under the speaker identification approach

Under the speaker identification approach, speaker overlaps cannot be detected, so all the segments containing speech from two or more speakers are removed when scoring test meetings. As expected, the classical offline system outperformed the proposed low-latency online system, but the performance of the latter was quite good, yielding only a 10.85% relative degradation (from 19.07 to 21.14% EER).

4.2. The effect of the speaker detection approach and score calibration

Attending to DET curves (not shown here for a lack of space) and EER, the system applying the speaker verification (SV) approach outperformed that applying the speaker identification (SI) approach. Using uncalibrated scores, the EER was 21.14% for the SI system and 12.03% for the SV system. But attending to the F-measure, the SI system outperformed the SV system (see Table 1). How may this be possible?
The DET curve is a very valuable means to compare the global discrimination capability of several speaker detection systems by presenting them the same set of trials. Note, however, that the sets of trials considered in DET curves for the SI and SV systems were different. The SV system considered as many trials as target speakers per test utterance (meaning that the same test utterance was evaluated many times), whereas the SI system considered a single trial per test utterance. Therefore, DET curves of SV systems were computed on much more trials than those of SI systems, and comparing them makes no sense. SV systems featured a high number of impostor trials. Since most of them were rejected, false alarm rates resulted remarkably lower than those of SI systems, thus yielding a better performance.
On the other hand, the F-measure is not defined in terms of trials but in terms of the time that was correctly detected. Therefore the F scores of SI and SV systems can be directly compared. Since the F scores of SI systems were better than those of SV systems, we conclude that SI systems outperform SV systems on the speaker tracking task defined on the AMI Corpus.
The F scores presented in Table 1 correspond to the operation points (thresholds) considered optimal in the DET curve. The threshold used for calibrated scores is based on application-dependent costs and target priors, which are adjusted using the development corpus. For uncalibrated scores, the threshold is fixed to zero, i.e. a target speaker is detected if the likelihood of the null hypothesis is higher than that of the alternative hypothesis. Results in Table 1 demonstrate the usefulness of the calibration stage, which leads to better performance in all the cases. The relative improvement is higher for SV systems, because calibration can compensate for the high number of false alarms at the cost of some misses.

Table 1. Precision, Recall and F-measure of SI and SV online speaker tracking systems in experiments on the evaluation set of the AMI Corpus.

	Uncalibrated			***Calibrated***		
	precision	***recall***	***F***	***precision***	***recall***	***F***
SI	0.71	0.91	0.80	0.81	0.85	0.83
SV-ExcOvlp	0.49	0.96	0.65	0.76	0.83	0.80
SV-IncOvlp	0.44	0.96	0.60	0.72	0.81	0.76

In the case of SV systems, which could theoretically detect various speakers at the same time, scores were computed either excluding or including overlapped segments. Both results (SV-ExclOvlp and SV-InclOvlp) are presented in Table 1. As expected, the performance of the SV-InclOvlp system was worse than that of the SV-ExclOvlp system: 7.69% worse when using uncalibrated scores, and 5% worse when using calibrated scores.

4.3. The effect of smoothing scores

The optimal *w* for the smoothing functions (which somehow depends on the average length of speaker turns) was heuristically determined on the development set. For the rectangular function, the optimal value was *w=2*. For the triangular function, it was *w=3*. Smoothing the scores consistently improved the speaker tracking performance on the test set of the AMI Corpus, the EER decreasing from 21.14% (no smoothing, *w=1*) to 19.37% (rectangular, *w=2*) and 18.64% (triangular, *w=3*), respectively. In terms of F-measure, a relative improvement of 3.61% was observed, from F=0.83 (no smothing, *w=1*) to F=0.86 (triangular, *w=3*).

5. SOA-compliant services

Two SOA-compliant services have been designed and implemented as the core elements for the deployment of SOA-based speaker tracking applications: a speaker tracking service (STservice) and a lower-level auxiliary speaker detection service (SDservice).
The SDservice captures the audio stream from an audio source (a field microphone integrated in the AmI environment) and outputs the likelihood scores of the target speakers at fixed-length (typically, one second) intervals, based on the analysis of the most recent window of speech. The SDservice performs feature extraction, speaker detection (based on target speaker models) and score calibration (based on a linear transform, optimized on a development corpus).
Taking advantage of service composition, a speaker tracking service (STservice) has been also designed which outputs actual speaker tracking decisions based on the outputs (detection scores) received from the SDservice. Decisions may be taken following either the speaker identification or the speaker verification approaches described in section 2. When performing speaker identification, the STservice outputs the identity of the most likely speaker. When performing speaker verification, several speakers can be detected simultaneously; on the other hand, if none of the scores is higher than the verification threshold, the STservice will output an impostor identifier.

The detection criterion (identification or verification) applied in making decisions, as well as the use of score calibration and the use of score smoothing, are optional features and depend on configuration parameters of the SDservice and the STservice, which can be updated from the speaker tracking application.

In practice, a speaker tracking application will invoke one STservice instance for each audio source (microphone) detected in the environment. As noted above, configuration parameters are determined by the application and depend on the scenario: decisions may be based on speaker verification if overlapped speech is allowed; smoothing based on past scores is activated if tracking robustness has to be increased, etc. Each STservice instance invokes a SDservice which continuously captures an audio stream and outputs a detection score per target speaker. The interaction between services and applications is based on subscriptions. Event subscription allows to get SDservice detection scores in a STservice instance, and to get STservice decisions in a speaker tracking application. Currently, the SOA based speaker tracking application is implemented following the UPnP standard.

The SDservice could be further composed by an audio capturing service which could be reused from many speech-based services and applications, such as speech recognition, language identification, etc. The SDservice could be also invoked for speaker adaptation in speech recognition tasks.

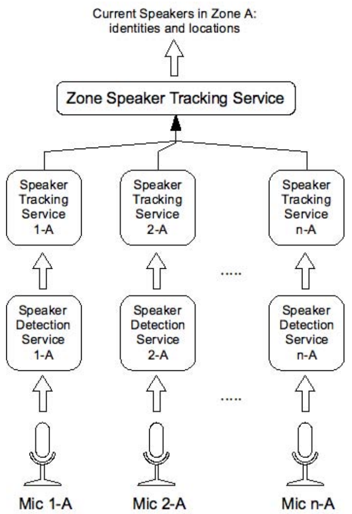


Figure 1. Definition of a Zone Speaker Tracking service.

Finally, a Zone Speaker Tracking service (ZSTservice) has been also defined, which assumes that AmI environments may be divided in separate spaces (e.g. rooms), and various microphones installed in each space. As shown in Figure 3, STservices can be organized hierarchically and their information gathered and interpreted by one ZSTservice. This way, by defining various ZSTservices (one per room), a high level management service could easily locate users in the AmI environment and follow their movements to customize services as users change their location.

6. Conclusions

A low-latency online speaker tracking approach and two SOA-compliant services, based on such approach, have been presented in this paper. Both elements have been designed for an Ambient Intelligence scenario with few users. The online speaker tracking system processes continuous audio streams and outputs a speaker identification decision for fixed-length (one second) segments. Speaker detection is done by means of a MAP-UBM speaker verification backend.

The proposed system was compared to a traditional system working offline, in experiments on a subset of the AMI Corpus of meeting conversations. Though offline segmentation of audio streams led to better results than using fixed-length segments, depending on the scenario and the required latency, offline audio segmentation may be unfeasible. The proposed approach provides low-latency online speaker tracking with little performance degradation.

To increase the robustness to local variability, a simple smoothing scheme was applied, consisting on a linear combination of the current score and a number of past scores. Promising results have been obtained in preliminary experiments.

Finally, two SOA-compliant services have been defined and implemented using UPnP, for speaker detection and tracking on a single audio source. A Zone Speaker Tracking Service has been also defined, which illustrates how those services could be connected and integrated in a home environment.

7. Acknowledgements

This work has been partially funded by the Government of the Basque Country, under program SAIOTEK, project S-PE09UN47; and the Spanish MICINN, under PN-I+D+i, project TIN2009-07446.

8. References

[1] A.F. Martin and M.A. Przybocki, "Speaker Recognition in a Multi-Speaker Environment", in Proceedings of European Conference on Speech Communications, Denmark, 2001.

[2] S.E. Tranter and D.A. Reynolds, "An Overview of Automatic Speaker Diarization Systems", IEEE Transactions on Audio, Speech and Language Processing, 14(5): 1557-1565, 2006.

[3] S. Meignier, D. Moraru, C. Fredouille, J.F. Bonastre and L. Besacier, "Step-by-step and Integrated Approaches in Broadcast News Speaker Diarization", Computer Speech and Language, 20:303-330, 2006.

[4] M. Kotti, V. Moschou and C. Kotropoulos, "Speaker Segmentation and Clustering", Signal Processing, 88: 1091-1124, 2008.

[5] D. Moraru, M. Ben and G. Gravier, "Experiments on Speaker Tracking and Segmentation in Radio Broadcast News", in Proceedings of the ICSLP, Lisboa, 2005.

[6] D. Istrate, N. Scheffer, C. Fredouille and J.F. Bonastre, "Broadcast News Speaker Tracking for ESTER 2005 Campaign", in Proceedings of the ICSLP, Lisboa, 2005.

[7] J.F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin and C. Wellekens, "A Speaker Tracking System based on Speaker Turn Detection for NIST Evaluation", in Proceeding of the IEEE ICASSP, Istanbul, Turkey, 2000.

[8] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus". Language Resources and Evaluation Journal , 41(2): 181-190, 2007.

[9] Reference Model for Service Oriented Architecture. http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.pdf

[10] Reference Architecture for Service Oriented Architecture. http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/soa-ra-pr-01.pdf

[11] L.J. Rodríguez, M. Peñagarikano and G. Bordel, "A Simple But Effective Approach to Speaker Tracking in Broadcast News", Pattern Recognition and Image Analysis, LNCS 4478: 48-55, Springer-Verlag, 2007.

[12] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 10:19-41, 2000.

[13] N. Brummer and J. du Preez, "Application Independent Evaluation of Speaker Detection", Computer Speech and Language, 20:230-275, 2006.

[14] A.F. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET curve in assessment of detection task performance", in Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech), Vol. 4, pp. 1895-1898, 1997.

Intra-session Variability Compensation for Speaker Segmentation

Carlos Vaquero, Alfonso Ortega, Eduardo Lleida

Communications Technology Group (GTC)
Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{cvaquero,ortega,lleida}@unizar.es

Abstract

This paper addresses the problem of speaker segmentation in two speaker telephone conversations, proposing a segmentation approach based on factor analysis and a novel method for intra-session variability compensation to improve segmentation performance. The segmentation system is evaluated on the NIST Speaker Recognition Evaluation 2008 summed channel test condition, showing that intra-session variability compensation allows to obtain around a 20% relative improvement in terms of speaker segmentation error.

Index Terms: Speaker segmentation, speaker and session variability, intra-session variability

1. Introduction

Recently, there has been a great advance in the field of speaker identification, in part motivated by the NIST Speaker Recognition Evaluations (SRE). One of the main breakthroughs of the last years has been the formulation of the Joint Factor Analysis (JFA) for speaker verification [1]. Nowadays most state of the art speaker verification systems are based on this approach. Since then, researchers have explored its application to different areas, specially to study new speaker diarization methods. One of the most interesting of these methods is the one presented in [2], a novel approach for streaming speaker diarization, which shows several differences with traditional diarization systems. This method makes use of a simple Factor Analysis (FA) model composed of only eigenvoices [3] to obtain high accuracy in a two speaker segmentation task on telephone conversations. However, performance decreases significantly when the number of speakers is unknown.

Consequently, the speaker identification community has focused on improving the performance in the two speaker segmentation task on telephone conversations, a task quite related to speaker verification. In [4] several approaches using JFA and Variational Bayes are proposed, obtaining better performance than the traditional Bayesian Information Criterion (BIC) based Agglomerative Hierarchical Clustering (AHC) [5]. However all approaches presented in [4] only model inter-speaker variability to perform speaker segmentation. In [10] the same approaches are analyzed and inter-session variability compensation is added, showing that it decreases performance, since inter-session variability may contain useful information to separate different speakers in a single session, specially if they are talking over different channels.

In this work we address the problem of speaker segmentation in two speaker conversations. We propose two methods to compensate the variability presented by a single speaker during a session (intra-session variability) and an eigenvoice based approach for two speaker segmentation similar to the one presented in [2], obtaining competitive performance compared to state-of-the-art 2-speaker segmentation systems [4], and showing further improvement when the mentioned variability is compensated.

This paper is organized as follows: In Section 2 we present the proposed segmentation system, and we describe different types of variability that may affect a diarization system in Section 3. In Section 4, we introduce two approaches to compensate intra-session variability and in Section 5 we evaluate the speaker segmentation system and the proposed intra-session variability compensation approaches. Finally, in Section 6 we summarize the conclusions of this study.

2. Segmentation System

In the proposed speaker segmentation system, described in [6], we use a factor analysis approach to model the desired sources of variability. As a starting point we try to capture the variability present among different speakers. For this purpose, we model every speaker by a Gaussian Mixture Model (GMM) adapted from a Universal Background Model (UBM) using an eigenvoice approach [3], according to:

$$M_s = M_{UBM} + Vy. \tag{1}$$

Where M_s is the speaker GMM supervector, obtained concatenating all Gaussian means, M_{UBM} is the UBM supervector, V is the low rank eigenvoice matrix, and y is the set of speaker factors, which follows a standard normal distribution $N(y|0, I)$ a priori. This way every speaker is represented by a GMM supervector in a high dimension space, and in such space we allow the speakers to lay in the low dimension subspace generated by the column vectors of V, which point to the directions of maximum variability among speakers. We refer to this variability as inter-speaker variability and to the low rank subspace as the speaker subspace.

In our approach we use a 256 Gaussian UBM, and as feature vectors we use 12 Mel Frequency Cepstral Coefficients (MFCC) including C0, computed every 10 ms over a 25 ms window. The dimension of the speaker subspace is 20, compared to the dimension of the supervector space that is $256 \times 12 = 3072$. This way every point estimate for a given speaker is defined by a set of 20 speaker factors.

To perform speaker segmentation given a sequence of feature vectors, as in [2], we estimate the speaker factors for every frame over a 100 frame window, with an overlap of 990 ms, an we estimate a 2-Gaussian GMM to model the stream of speaker factors obtained, after removing silence frames according to a Voice Activity Detector (VAD). Each one of these Gaussians

This work has been partially funded by the national project TIN2008-06856-C05-04.

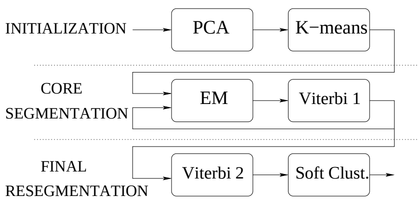


Figure 1: *Block diagram of the proposed segmentation system*

will be assigned to a single speaker. In contrast to [2], we estimate the GMM using all available data in the recording, rather than processing 1 minute slices and applying a clustering technique. The later allows stream processing with 1 minute latency but the former yields better results. A block diagram of the proposed segmentation system is shown in Fig. 1.

2.1. Initialization

We have detected that a good initialization is quite important to ensure that every Gaussian in the GMM corresponds to a single speaker. In our approach, we use prior knowledge about speaker factors proposed in [1]: A priori, speaker factors are assumed to be distributed according to the standard normal distribution $N(y|0, I)$. Since we obtain speaker factors from a small data sample (100 frames, which is small compared to the number of frames that speaker recognition systems usually manage, around 10000), using MAP estimation, we can expect the posterior distribution of speaker factors for a single speaker to keep some properties of the prior. Assuming that the posterior variance is close to I, we can perform PCA to obtain the direction of maximum variability in the speaker factor space. Such direction should be the best one to separate speakers, since both are supposed to have a variance close to I and a different mean.

This strategy gives two clusters that can be seen as a first speaker segmentation, and then K-means clustering is performed to reassign frames to the two clusters and a single Gaussian is trained on each of them. Using this frame assignment directly as segmentation output gives reasonably good results, as we will see later, in Section 5.

2.2. Core Segmentation

The 2 Gaussians previously trained serve as initial GMM of the whole recording. Then a two stage iterative process is applied until convergence: first several Expectation-Maximization (EM) iterations are used and then, every Gaussian is assigned to a single speaker and a Viterbi segmentation is performed (Viterbi 1 in Fig. 1). According to this new frame assignment, 2 Gaussian models are trained and the iterative process restarts again. Convergence is reached when the segmentation of the current iteration is identical to that obtained in the previous one.

To avoid fast speaker changes, in the Viterbi segmentation, we modify the speaker turn duration distribution using a sequence of tied-states [7] for every speaker model. This way, we avoid the state duration to follow a geometric distribution that cannot accurately model real speaker turn durations. Each speaker model is composed of 10 states that share the same observation distribution, a single Gaussian in this case. Tied-states are not considered for the silence, but a single state without an observation distribution is used, since the algorithm is forced to go through the silence state according to the VAD labels. We have observed that this way of modeling speaker turn duration yields better results than modifying the transition probability.

2.3. Viterbi Resegmentation and Soft Clustering

The output of the core segmentation system gives accurate speaker labels in most cases, but these labels can be refined by means of Viterbi resegmentations (Viterbi 2 in Fig. 1). In this case we model every speaker with a 32 component GMM according to the output of the core segmentation system using as features 12 MFCC including C0. Again we use 10 tied-states for speaker models and a single state for all silence frames.

After this resegmentation we retrain the GMM models and run a forward backward decoding to perform a soft reassignment of the frames to the two speakers. GMM models are retrained according to the soft reassignment and a final Viterbi resegmentation is performed. This approach was first presented in [4] as soft-clustering.

3. Speaker, Session and Intra-session Variability in Speaker Diarization

In the proposed approach for speaker segmentation we only take into account inter-speaker variability. However there are other sources of variability that may affect a segmentation or diarization system. In speaker recognition systems, one of the hardest problems is to deal with the variability present in a speaker recorded over different sessions. This is known as inter-session variability and includes variability due to the speaker, since his speech may vary along different recording sessions, as well as variability due to the recording environment. There are several techniques to model this variability. Some of the more recent and successful approaches have been Nuissance Attribute Projection (NAP) for SVM-GMM speaker recognition systems [8], Eigenchannel modeling, or JFA [9]. All this techniques assume that the speaker is modeled by a supervector (usually a GMM-sv) in a high dimension space and different sessions for a given space produce different estimations of the speaker supervector. The variability in these estimations or inter-session variability is assumed to lay in a low dimension sub-space, so all inter-session variability compensation techniques try to estimate the component of the speaker session in such space and remove it to obtain a session independent speaker supervector.

The question is if inter-session variability compensation is useful for speaker diarization. Speaker diarization systems aim at answering the question “Who spoke when?” in a unsupervised fashion. We can think that inter-session variability compensation do not help for speaker diarization, for two main reasons: First, diarization is performed over one session without prior knowledge of the speakers involved, so we will never get the same speaker over different sessions. Secondly, in many scenarios session variability models may enhance diarization performance since different speakers may use different communication channels. This is the case of telephone conversations or meetings in a room where the speakers remain static.

Finally, a single speaker can present variability during a single session when we process such session in small segments. We will refer to this variability as intra-session variability. Some examples of this variability includes emotions or excitement of the speaker as the conversation evolves, or the unbalanced phonetic load present in small segments as in the proposed system (1 second segments). Intra-session variability is not usually taken into account for speaker recognition, since state of the art systems usually integrate over all observations of a given speaker obtaining an average model, which may differ from session to session. In such case intra-session variability modeling and compensation will only be useful as far as it is related to inter-session variability. Actually, both intra and inter-session variability share many sources of variability, but some of them are more critical than others. For example, channel is a source of inter-session variability that in general will not introduce intra-session variability (but it could, e.g., if a speaker is recorded in a room with a far field microphone and he moves as he talks). On the other hand, unbalanced phonetic load will be more critical for intra-session variability modeling, specially as the segments to analyze in a given session become smaller.

However, intra-session variability is very important and should be taken into account in the task of speaker diarization, since in such task we analyze small and pure segments and try to agglomerate them to obtain pure clusters that should belong to a single speaker. In the following section we describe an approach for supervised intra-session variability compensation.

4. Intra-session variability compensation

Given a recording, the segmentation system proposed in section 2, produces a sequence of speaker factor vectors estimated every 10 ms over 1 sec. window. Assuming that a set of S recordings is available and each recording contains a single speaker, we can obtain a sequence $y^s = y_1^s, ..., y_{N^s}^s$ of N^s speaker factor vectors for every recording session s. The speaker factors obtained from a session belongs to the same class (same speaker), so we can study the inter-session and intra-session as between-class and within-class variability respectively. This approach is similar to the one presented in [11], but in that case it was used for speaker recognition and inter-session variability compensation.

4.1. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a well known technique for dimensionality reduction in pattern recognition that, given a set of features belonging to different classes and laying in the feature space, seeks the orthogonal basis for such space that enables better discrimination between different classes by maximizing between-class variance and minimizing within class variance. Linear discriminant analysis assumes that the observations belonging to each class are normally distributed and that within class covariance is kept across different classes. The speaker factor vectors satisfy the first assumption, while the second is expected to be satisfied since we do not expect the posterior covariance of y^s to be very different of the prior as we explained in section 2.

In our problem we estimate between-class covariance (S_b) and within class (S_w) covariance as:

$$S_b = \frac{1}{S-1}\sum_{s=1}^{S}(\mu^s - \mu)(\mu^s - \mu)^T \tag{2}$$

$$S_w = \frac{1}{S-1}\sum_{s=1}^{S}\frac{1}{N^s-1}\sum_{n=1}^{N^s}(y_n^s - \mu^s)(y_n^s - \mu^s)^T \tag{3}$$

$$\mu^s = \frac{1}{N^s}\sum_{s=1}^{N^s} y_n^s \tag{4}$$

$$\mu = \frac{1}{S}\sum_{n=1}^{N^s}\mu^s \tag{5}$$

The problem reduces to find the matrix v of eigenvectors that satisfies:

$$S_b v = \lambda S_w v, \tag{6}$$

and project the speaker factors onto v or onto a low rank matrix A obtained selecting those eigenvectors having higher eigenvalues, for dimensionality reduction.

4.2. Within Class Covariance Normalization

Within class covariance normalization (WCCN) is a normalization method that allows to obtain a linear transformation for a given set of features belonging to different classes so that the within class covariance matrix S_w defined in Eq. 3 is equal to the identity matrix I. Again this technique assumes that all classes have the same covariance matrix.

To obtain the linear transformation we first obtain S_w as shown in Eq. 3 and then we apply Cholesky decomposition, so the transformed speaker factors y' will follow this expression:

$$y' = Ry \tag{7}$$

$$S_w^{-1} = R'R \tag{8}$$

where R is the upper triangular matrix obtained by Cholesky decomposition.

5. Performance Analysis

5.1. Experimental Setup

We study the performance of the proposed segmentation system and intra-speaker variability compensation in terms of segmentation error rate. As development data to train the UBM, V matrix, LDA and WCCN we use all telephone data available from 1conv and 8conv conditions from the NIST SRE evaluations 2004, 2005 and 2006. As evaluation data we use the summed channel test condition from the NIST SRE 2008. This condition comprises 2213 2-speaker telephone conversations of around five minutes length each. As ground truth for segmentation error rate computation we extract the segmentation labels from the ASR NIST transcriptions obtained separately on each telephone of the conversation.

5.2. Segmentation Performance: Baseline

As we explained in Section 2, the proposed segmentation system comprises several steps, including PCA initialization, K-means clustering, iterative EM and Viterbi segmentation in the speaker factor space, a Viterbi resegmentation using MFCC features and a last soft-clustering resegmentation. Table 1 shows the results obtained by the segmentation system after every step:

Segmentation system	Seg error (%)	σ(%)
PCA	20.2	14.3
+K-means	4.9	8.8
Core segmentation system	3.1	6.6
+Viterbi resegmentation	2.3	6.2
+Soft-clustering	2.2	6.1

Table 1: *Performance of the segmentation system and standard deviation step by step.*

Given these results we can extract several conclusions. First, speaker factors enable easy separability between speakers. Just with PCA and K-means clustering we get 4.9% segmentation error. Note that at that point, frames are assigned to one speaker or the other assuming statistical independence, no context or temporal information is used. Completing the core system gives great improvement and results are comparable to those obtained with the best systems presented in [4]. Moreover, after resegmentations results improve further.

5.3. Intra-speaker Variability Compensation

To study the performance of intra-speaker variability compensation we compare the segmentation error obtained before the resegmentation stages (after the core segmentation in Fig. 1) with and without using the intra-speaker variability compensation methods described in Section 4. For comparison when using LDA for dimensionality reduction we show results using 20 speaker factors (baseline system) and 50 speaker factors.

Segmentation system	Seg error (%)	$\sigma(\%)$
Baseline (20 spk factors)	3.1	6.6
WCCN (20 spk factors)	**2.5**	**5.5**
50 spk factors	2.9	6.9
LDA 50 to 20	2.7	5.7
LDA 50 to 20 + WCCN	2.5	5.7
50 spk factors + WCCN	**2.1**	**5.6**

Table 2: *Performance of the core segmentation system with and without intra-speaker variability compensation.*

As we can see in Table 2, both LDA and WCCN approaches for intra-speaker variability compensation outperforms our baseline. Using WCCN directly on 20 speaker factors reduces the segmentation error from 3.1% to 2.5%, obtaining a 20% of relative improvement. Using LDA to obtain 20 dimension vectors from 50 speaker factors improves also the performance of the system compared to the baseline using both 20 and 50 speaker factors. In addition, the performance can be further improved applying WCCN after LDA.

However using LDA+WCCN on 50 speaker factors is not significantly better than using WCCN directly on 20 speaker factors. Moreover, the most critical step in the proposed system regarding computational cost is the speaker factor computation ($O(d^2)$, with d the dimension of the speaker factors), and once speaker factors are computed, the classification algorithm is fast compared to speaker factor computation ($O(d)$). Therefore, the computational cost of the system using LDA for dimensionality reduction is comparable to the cost of the system using 50 speaker factors and is much higher than the cost of the system using 20 speaker factors. For this reason, we show the results obtained with 50 speaker factors and WCCN for intra-speaker compensation. We obtain a 28% relative improvement when using WCCN on 50 speaker factors. It seems that a higher dimensionality enables WCCN to improve further.

Taking into account the computational cost, we can affirm that, even though LDA based intra-session variability compensation shows improvements, it is not useful for our system since using WCCN on low dimension speaker factor space performs as good as using LDA+WCCN on a higher dimension speaker factor, but this second approach is much more costly, and if we use WCCN directly on the higher dimension speaker factor space we obtain further improvement keeping the computational cost comparable to LDA+WCCN.

5.4. Results with the Full Segmentation System

In the previous subsections we have shown results for the core segmentation system, but the proposed segmentation system can increase its performance using Viterbi resegmentation after obtaining the core segmentation output.

Results in Table 3 show that while increasing the number of speaker factors is not effective after Viterbi and soft-clustering resegmentations, intra-session variability compensation using WCCN is still effective, obtaining a relative performance improvement of 18% for 20 speaker factors and 23% for 50 speaker factors. In addition it is shown that increasing the number of speaker factors may not be helpful if intra-session variability is not compensated, probably because some directions of the speaker space are related to intra-session variability.

Segmentation system	Seg error (%)	$\sigma(\%)$
Baseline + reseg	2.2	6.1
WCCN + reseg	**1.8**	**5.0**
50 spk factors + reseg	2.2	6.2
50 spk factors + WCCN + reseg	**1.7**	**5.2**

Table 3: *Performance of the segmentation system with WCCN for intra-speaker variability compensation.*

6. Conclusions

In this study, we have introduced two methods for intra-session variability compensation in the task of speaker segmentation and diarization, based on LDA and WCCN. In addition, we have proposed a two speaker segmentation system based on the one presented in [2], introducing a set of improvements, including a novel PCA initialization and a modification of the speaker turn duration distribution, that enables us to obtain a 2.2% segmentation error on the summed dataset from the NIST SRE 2008. We have shown that intra-session variability compensation can improve performance of a segmentation system, reducing the segmentation error rate to 1.8%.

7. References

[1] P. Kenny et al, “A Study of Inter-Speaker Variability in Speaker Verification”, IEEE Trans. Audio, Speech Proc., 2008

[2] Castaldo, F. et al, “Stream Based Speaker Segmentation Using Speaker Factors and Eigenvoices”, in Proc ICASSP, 4133-4136, Las Vegas, NV, 2008.

[3] R. Kuhn et al. “Rapid Speaker Adaptation in Eigenvoice Space”, IEEE Trans. Speech Audio Proc. Vol 8, no. 6, 695–707, 2000.

[4] Reynolds, D. et al “A Study of New Approaches to Speaker Diarization”, in Proc Interspeech, 1047–1050, Brighton, UK, 2009

[5] Reynolds, D. A. and Torres-Carrasquillo, P., “Approaches and applications of audio diarization”, In Proc ICASSP, V:953–956, Philadelphia, PA, 2005.

[6] Vaquero, C. et al “Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification”, to appear in Interspeech, Makuhari, Japan, 2010.

[7] Levinson, S.E., “Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition”, Computer Speech and Language, I:29–45, 1986.

[8] Campbell, W. et al “SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation”, In Proc ICASSP, Toulouse, France, 2006.

[9] Kenny, P. et al, “Joint factor analysis versus eigenchannels in speaker recognition” IEEE Trans. Audio, Speech Proc. 15 (4), pp. 1435-1447, 2007

[10] Kenny, P. et al, “Diarization of Telephone Conversations using Factor Analysis” IEEE Journal of Selected Topics in Signal Processing, 2010.

[11] Dehak, N. et al, “Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysi”, in Proc ICASSP, Taipei, Taiwan, 2009.

Detection of Overlapped Acoustic Events using Fusion of Audio and Video Modalities

Taras Butko and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Spain
taras.butko@upc.edu, climent.nadeu@upc.edu

Abstract

Acoustic event detection (AED) may help to describe acoustic scenes, and also contribute to improve the robustness of speech technologies. Even if the number of considered events is not large, that detection becomes a difficult task in scenarios where the AEs are produced rather spontaneously and often overlap in time with speech. In this work, fusion of audio and video information at either feature or decision level is performed, and the results are compared for different levels of signal overlaps. The best improvement with respect to an audio-only baseline system was obtained using the feature-level fusion technique. Furthermore, a significant recognition rate improvement is observed where the AEs are overlapped with loud speech, mainly due to the fact that the video modality remains unaffected by the interfering sound.

Index Terms: Acoustic Event detection, Multimodal Fusion, Fuzzy Integral, Acoustic Localization

1. Introduction

Acoustic event detection (AED) aims at determining the identity of sounds and their temporal position in the signals that are captured by one or several microphones. It can provide a support for a high-level analysis of the underlying acoustic scene. This analysis includes the description of human activity which is reflected in a rich variety of AEs, either produced by the human body or by objects handled by them. Moreover, AED can contribute to improve the performance and robustness of speech technologies such as speech and speaker recognition, speech enhancement.

AED is usually addressed from an audio perspective and many reported works are intended for indexing and retrieval of multimedia documents [1], or to improve robustness of speech recognition [2]. AED has been adopted as a relevant technology in several international projects, like CHIL [3], and evaluation campaigns [4]. The last international evaluations in seminar conditions have shown that AED is still a challenging problem. According to those results, the detection of AEs from only audio information shows a large amount of errors, which are mostly due to temporal overlaps of sounds.

The overlap problem may be faced by developing more efficient algorithms either at: the signal level, using source separation techniques like independent component analysis [5]; the feature level, by means of specific features [6]; or the model level [7]. An alternative approach consists of using an additional modality that is less sensitive to the overlap phenomena present in the audio signal.

Most of human produced AEs have a visual correlate that can be exploited to enhance detection rate. This idea was first presented in [8], where the detection of footsteps was improved by exploiting the velocity information obtained from a video-based person-tracking system. Further improvement has been achieved by the authors in [9] [10] where the concept of multimodal AED is extended to detect and recognize a set of 11 AEs. In that work, not only video information but also acoustic source localization information was considered. Either a decision-level fuzzy integral fusion [9] or a feature-level fusion [10] was used to increase the accuracy of detection of isolated AEs. But for most of the AEs a statistically significant improvement was not observed due to the fact that in clean conditions the baseline recognition results are relatively high, so the additional modalities can not contribute significantly.

In this work we compare feature-level and decision-level fusion techniques for AED in more realistic conditions where the AEs are overlapped with speech. Feature-level fusion is performed by means concatenation of features from different modalities into one super-vector. Decision-level fusion is carried out with the Weighted Arithmetical Mean (WAM) approach and the Fuzzy Integral (FI) statistical approach [11].

2. Database and metric

There are several publicly available multimodal databases designed to recognize events, activities, and their relationships in interaction scenarios [3]. However, these data are not well suited to audiovisual AED since the employed cameras do not provide a close view of the subjects under study. In order to assess the performance of the proposed multimodal fusion approaches, the subset of isolated AEs from a recently recorded multimodal database [9] [10] was used. The video signals were recorded with 5 calibrated cameras at pixel resolution 768x576 and 25 fps. Audio signals were collected from 6 T-shaped 4-microphone clusters, and sampled at 44.1 kHz (in total, 24 microphones are used). All sensors were synchronized. In the recorded scenes, 5 different subjects performed several times the AEs employed in this work, adding up to around 100 instances for every AE, and 2 hours. This multimodal database is publicly available from the authors. We consider 12 classes of AEs which naturally occur in meeting-room environments, like in [7], [8], [9] and [10]: "Door knock", "Door open/slam", "Steps", "Chair moving", "Spoon/cup jingle", "Paper work", "Key jingle", "Keyboard typing", "Phone ring", "Applause", "Cough", and "Speech".

The meeting scenario adopted for this work assumes that there are two simultaneous acoustic sources in the room: one is always speech and the other is a specific AE. Taking into account this assumption, our UPC's smart-room has been considered ideally subdivided in the two areas: left and right (Figure 1 (a)). In the left part the speaker produces speech, and in the right part the listener produces different types of AEs. This assumption allows us to analyze the left and right parts of the room independently for the extraction of acoustic source localization features.

The speech of the speaker was recorded separately and it was artificially overlapped with the database of isolated AEs. To do that, for each AE instance, a segment with the same length was extracted from a random position inside the speech signal. The overlapping was performed with 5 different

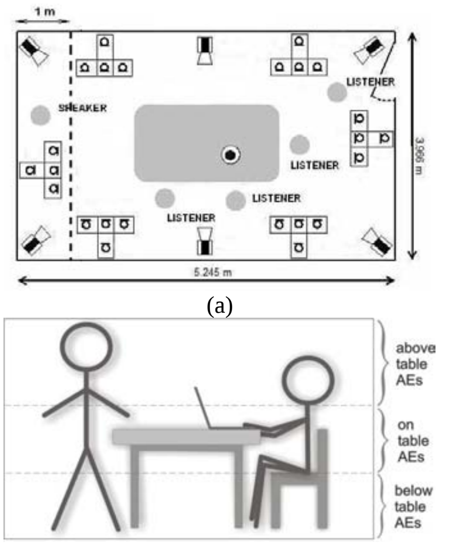
Signal-to-Noise Ratios (SNRs): 20 dB, 10dB, 0dB, -10dB. -20 dB, where speech is considered as "noise".

Although the database with overlapped AEs is generated in an artificial way, it has some advantages:

a) The behavior of the system can be analyzed for different levels of overlap.

b) The existing databases of isolated AEs with high number of instances can be used for evaluations.

The metric referred to as AED-ACC [7], which is the F-score (harmonic mean between precision and recall), is employed to assess the final accuracy of the presented algorithms. .



(a)

(b)

Figure 1: *(a) Top view of the room. (b) The three categories along the vertical axis.*

3. Feature extraction

A first stage of the proposed multimodal AED system is to determine the most informative features related to the AEs of interest for every input modality. Although audio and localization are originated from the same physical acoustic source, they are regarded as two different modalities.

3.1. Spectro-temporal audio features

A set of audio spectro-temporal features, like those used in automatic speech recognition, is extracted to describe every audio frame. It consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives [10], which represent the spectral envelope of the audio waveform within a frame, as well as its temporal evolution. In total, a 32-dimensional feature vector is used. The FF feature extraction scheme consists in calculating a log filter-bank energy vector for each signal frame (in our experiments the frame length is 30 ms with 20 ms shift, Hamming window is applied) and then applying a FIR filter $h(k)$ on this vector along the frequency axis. We use the $h(k)=\{1, 0, -1\}$ filter in our approach. The end-points are taken into account. Notice that FF requires less computation than the classical MFCC.

3.2. Room model and localization features

To enhance the recognition results of the baseline system additional features are proposed. In our case, as the characteristics of the room are known beforehand (Figure 1 (a)), the position (x, y, z) of the acoustic source may carry useful information. In fact, events as door slam and door knock can only appear near the door, so a feature which describes the distance from the door is employed in this paper. On the other hand, usually each AE has an associated height, so the z position of the acoustic source may help to distinguish among AEs. The following categories are defined as indicated in Figure 1 (b): *below table, on table, and above table.*

The acoustic localization system used in this work is based on the SRP-PHAT [12] localization method, which is known to perform robustly in most scenarios. In short, this algorithm consists of exploring the 3D space, searching for the maximum of the global contribution of the PHAT-frequency-weighted cross-correlations from all the microphone pairs.

3.3. Video features

Tracking of multiple people present in the analysis area basically produces two figures associated with each target: position and velocity. The human velocity is readily associated to the footsteps AE. Multiple cameras are employed to perform tracking of several people interacting in the scene, by applying the real-time performance algorithm presented in [13].

The motion visual analysis is also used to detect two other acoustic events: paper wrapping and door slam. A motion of a white object near a human in the scene can be associated to paper wrapping (under the assumption that a paper sheet is distinguishable from the background color). The movement of the door can be well detected by the camera oriented towards the door. In order to visually detect a door slam AE, we exploited the a-priori knowledge about the physical location of the door. Analyzing the zenithal camera view, activity near the door can be addressed by means of a foreground/background pixel classification [14]. A high enough amount of foreground pixels in the door area will indicate that a person has entered or exited, hence allowing the visual detection of a door slam AE.

Detection of certain objects in the scene can be beneficial to detect AEs such as phone ringing, cup clinking or keyboard typing. Unfortunately, phones and cups are too small to be efficiently detected in the scene but, the case of a laptop can be addressed. In our case, the detection of laptops is performed from a zenithal camera located at the ceiling of the scenario. The algorithm initially detects the laptop's screen and keyboard separately and, in a second stage, assesses their relative position and size [15]. Once the position of the laptop is detected, the amount of "skin" pixels over this position will allow to decide about a keyboard typing AE.

4. Multimodal Acoustic Event Detection

Typically, low energy AEs such as paper wrapping, keyboard typing or footsteps are hard to detect using audio features, so both the visual correlate and the acoustic localization measures of these AEs may help to increase the detection performance.

In this paper, three data sources are combined for multimodal AED. First, two information sources are derived from acoustic data processing: single channel audio provides audio spectro-temporal (AST) features, while microphone array processing estimates the 3D location of the audio source. Second, information from multiple cameras covering the scenario allows extracting cues related to some AEs (described in Section 3.3). The features obtained from all modalities are combined together at feature and decision levels (Figure 2).

We employ a one-against-all detection strategy, so only two models are used for each AE, which will herewith be called "Class" and "non-Class". The first model is trained using the signals coming from the given class of interest, while the second model is trained using the rest of signals. In total, 12 HMM-based binary detectors working in parallel are needed to perform detection of all AEs [10].

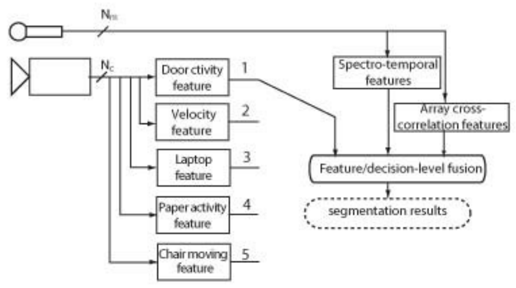


Figure 2: *System flowchart.*

4.1. Fusion of different modalities

The information fusion can be done on data, feature, and decision levels. Data fusion is rarely found in multi-modal systems because raw data is usually not compatible among modalities. For instance, audio is represented by one-dimensional vector of samples, whereas video is organized in two-dimensional frames. Concatenating feature vectors from different modalities into one super vector is an easy and simple way for combining audio and visual information. This approach has been reported, for instance, in [16] for multimodal speech recognition. An alternative to feature-level fusion is to model each different feature set separately, design a specialized classifier for this feature set, and combine the classifier output scores. Each such classifier acts as an independent "expert", giving its opinion about the unknown AE. The fusion rule then combines the individual experts' match scores. This approach is referred here as decision-level fusion. In the presented work, fusion is carried out on the decision level using weighted arithmetical mean (WAM) and fuzzy integral (FI) [11] fusion approaches. Unlike non-trainable fusion operators (mean, product), the statistical approaches WAM and FI avoid the assumption of equal importance of information sources. Moreover the FI fusion operator also takes into account the interdependences among modalities.

4.1.1. *Feature-level fusion approach*

In this work we use a HMM-GMM approach with feature-level fusion, which is implemented by concatenating the feature sets *Xs* from *S* different modalities in one super-vector:

$$\boldsymbol{Z} = \boldsymbol{X}_1 \cup \boldsymbol{X}_2 \cup \ldots \cup \boldsymbol{X}_S \qquad (1)$$

Then, the likelihood of that observation super-vector at state *j* and time *t* is calculated as:

$$b_Z(t) = \sum_m p_m N(\boldsymbol{Z}_t; \boldsymbol{\mu}_m; \boldsymbol{\Sigma}_m). \qquad (2)$$

where $N(.;\mu;\Sigma)$ is a multi-variate Gaussian pdf with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and p_m are the mixture weights. Assuming uncorrelated feature streams, diagonal covariance matrices are considered.

Feature-level fusion becomes a difficult task when some features are missing. Although the AST features can be extracted at every time instance, the feature that corresponds to the localization of acoustic source has an undefined value in the absence of any acoustic activity. In our experiments we substitute the missing features (*x*, *y*, *z* coordinates) with a predefined "synthetic" value (we use -1 value in our experiments). In this case we explicitly assign the 3D "position" of the silence event to have the value (*-1*, *-1*, *-1*).

4.1.2. *Decision-level fusion approach*

The decision-level fusion process is schematically depicted in Figure 3. First, a HMM segmentation based on the spectro-temporal features is performed to find all non-silence segments in the input audio. Given the "Class" and "non-Class" HMM models the log-likelihood ratio (LLR) is obtained for each non-silence segment S_i and each modality separately. A high positive LLR score would mean a high confidence that the non-Silence segment belongs to the "Class", while a low negative score would mean that the segment more likely belongs to "non-Class". A value close to zero indicates low confidence of decision. Furthermore, the obtained scores are normalized to be in the range [0…1] and their sum equal to 1. Then the normalized values are fused together using either Weighted Arithmetical Mean (WAM) or Fuzzy Integral (FI) fusion operators. To estimate the weights in WAM operator we use constrained regression approach to minimize the variance of error on development data [11]. The individual weights for the fuzzy integral fusion are also trained on development data using the gradient descent training algorithm.

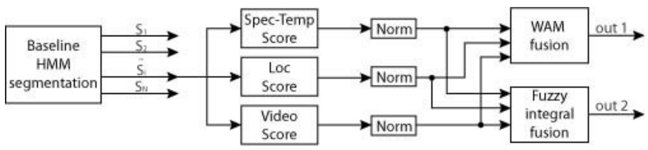


Figure 3: *Flowchart of decision-level fusion.*

5. Experiments and results

The detection results corresponding to two mono-modal AED systems based on AST and video features, respectively, are presented in Figure 4. The results for the video-based system are presented as an average accuracy score for those AEs for which the video counterpart is taken into consideration.

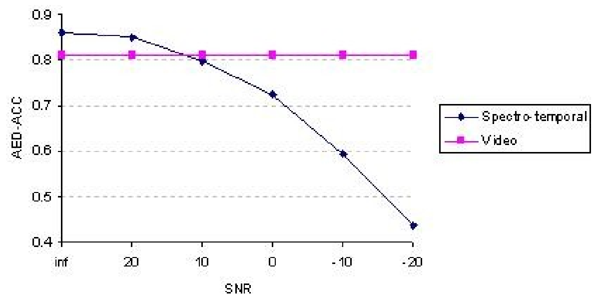


Figure 4: *Mono-modal AED results.*

Note the recognition results do not change for different SNR conditions since the video signals are not affected by overlapped speech. We do not present results for the AED system based on localization features since the information about the position of the acoustic source enables to detect just the category but not the AE within it. As we see from Figure 4, the recognition results of the baseline system decrease significantly for low SNRs.

The average relative improvement obtained by the multimodal system with respect to the baseline system (that uses the AST features only) for different fusion techniques is displayed in Figure 5. The feature-level fusion performs better for all AEs than both WAM and FI decision-level fusion approaches, and the both decision-level fusion techniques showed similar results in our experiments.

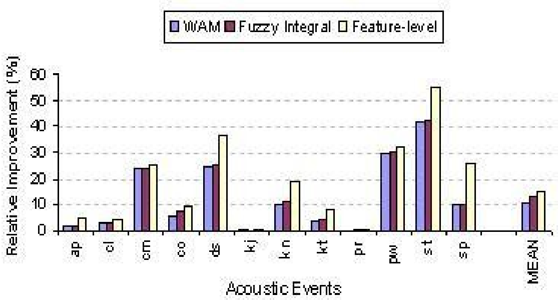


Figure 5: *Average relative improvement obtained by the multimodal system*.

The next Figure 6 summarizes (averaged over all Aes) the relative improvement obtained with the feature-level and decision-level (using fuzzy integral) fusion techniques for different levels of SNRs. According to these results, video signals as well as signals from arrays of microphones showed to be a useful additional source of information to cope with the problem of AED in overlapping conditions.

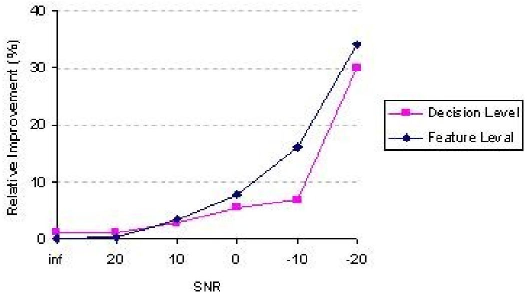


Figure 6: *The relative improvement obtained from multimodal features for different SNRs.*

6. Conclusions

In this paper, a comparison between multimodal systems based on a feature-level and decision-level fusion approaches have been presented. The acoustic data is processed to obtain a set of spectro-temporal features and the three localization coordinates of the sound source. Additionally, a number of features are extracted from the video signals by means of object detection, motion analysis, and multi-camera person tracking to represent the visual counterpart of several AEs.

The obtained results showed that although in clean conditions the video and localization information does not contribute significantly, more improvement can be achieved in the conditions where the audio signals are overlapped with speech.

Future work will be devoted to extend the multimodal AED system to other classes as well as the elaboration of new multimodal features.

7. Acknowledgements

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is partially supported by a grant from the Catalan autonomous government. Particular thanks are given to Carlos Segura for providing the acoustic localization features in the proposed scenario. The authors would like to thank Cristian Canton-Ferrer and Xavier Giró for providing the features from the video signals.

8. References

[1] L. Lu, H. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation", IEEE Trans. on Speech and Audio Processing, vol. 10, pp. 504–516, 2002.

[2] T. Nishiura, S. Nakamura, K. Miki, and K. Shikano, "Environmental sound source identification based on hidden Markov models for robust speech recognition", in Proc. Eurospeech, pp. 2157–2160, 2003.

[3] CHIL: Computers in the Human Interaction Loop, <http://chil.server.de/>.

[4] CLEAR, 2007. Classification of Events, Activities and Relationships. Evaluation and Workshop. <http://www.clear-evaluation.org/>.

[5] A. Hyvärinen, J. Karhunen, E. Oja, "Independent Component Analysis", John Wiley & Sons, 2001.

[6] S. Wrigley, G. Brown, V. Wan, S. Renals, "Speech and crosstalk detection in multi-channel audio", IEEE Trans. Speech Audio Process, v. 13, pp. 84–91, 2005.

[7] A. Temko, C. Nadeu, "Acoustic event detection in meeting-room environments", in Pattern Recognition Letters, v. 30, pp. 1281-1288, 2009.

[8] T. Butko, A. Temko, C. Nadeu and C. Canton, "Inclusion of Video Information for Detection of Acoustic Events using the Fuzzy Integral", in Machine Learning for Multimodal Interaction, LNCS, vol. 5237/2008, pp. 74-85, Springer, 2008

[9] C. Canton-Ferrer, T. Butko, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, "Audiovisual Event Detection Towards Scene Understanding", in Proc. IEEE Int. Conference on Computer Vision and Pattern Recognition, 2009.

[10] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, J.R. Casas, "Improving Detection of Acoustic Events Using Audiovisual Data and Feature Level Fusion", in Proc. Interspeech, 2009.

[11] L. Kuncheva, Combining Pattern Classifiers, John Wiley & Sons, 2004.

[12] J. Dibiase, H. Silverman, M. Brandstein, "Microphone Arrays. Robust Localization in Reverberant Rooms", Springer, 2001

[13] C. Canton-Ferrer, R. Sblendido, J.R. Casas, M. Pardàs, "Particle filtering and sparse sampling for multi-person 3D tracking", in Proc. IEEE Int. Conf. on Image Processing, pp. 2644–2647, 2008.

[14] C. Stauffer, W. Grimson, "Adaptive background mixture models for real-time tracking", in Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 252–259, 1999.

[15] X. Giró and F. Marqués, "Composite object detection in video sequences: Applications to controlled environments", in Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services, pp. 1–4, 200.

[16] M. Chan, Y. Zhang, T. Huang, "Real-time lip tracking and bi-modal continuous speech recognition", in Proc. IEEE Workshop on Multimedia Signal Processing, 1998.

Oral Session 4: Machine Translation and Technology Development

Translation Dictionaries Triangulation

Alberto Simões[1], *Xavier Gómez Guinovart*[2]

[1]Universidade do Minho
[2]Grupo TALG, Universidade de Vigo
ambs@di.uminho.pt, xgg@uvigo.es

Abstract

Probabilistic Translation Dictionaries (PTD) are translation resources that can be obtained automatically from parallel corpora. Although this process is simple, it requires the existence of a parallel corpora for the involved languages.

Minoritized languages have a limited amount of available resources. For example, while they can have a few parallel corpora, the number of parallel language-pairs uses to be restricted.

We defend that if a minoritized language A has a parallel corpus with a language B, and language B has a parallel corpus with another language C, then we can obtain a helpful probabilistic translation dictionary between A and C.

In this document we will formalize the probabilistic translation dictionaries triangulation, perform some experiments making the triangulation between Galician, English and Italian, and conclude with an evaluation of the proposed approach.

Index Terms: probabilistic translation dictionaries, parallel corpora

1. Introduction

Translation between languages require as a basic resource a set of translation dictionaries, that is, a mapping from words or terms from one language to the other language. The main problem is that these resources are hard to create manually (time consuming, error prone, and hand-work intensive). Fortunately there are some methods [1, 2, 3, 4, 5] to analyze words used in a parallel corpus and extract automatically translation dictionaries.

These dictionaries are commonly named as Probabilistic Translation Dictionaries (PTD) as they are created in a statistic base. Nevertheless, they proven to be useful and have been used for different tasks:

- [6] describes a method to bootstrap a conventional translation dictionary from a PTD. A PTD was created and a try-and-error approach was used to define a translation probability threshold for filtering purposes. The filtered dictionary was used for manual validation.
- [7] uses PTDs for a similar task, bootstraping a machine translation dictionary. In this case the manual validation was not required.
- [8] uses PTDs as a mechanism to present parallel concordances and guessing translations of the searched terms, highlighting them when presenting the search result.
- [9, 10] presents methods to extract bilingual terminology using traslation patterns and PTDs for translations alignment.
- [11] also uses PTDs as a mechanism to align chunks of text when creating translation examples.
- [12] uses PTDs for cross-language information retrieval.

While useful, PTDs can only be obtained automatically if we have access to a parallel corpora in the required languages. That is, when computing a probabilistic translation dictionary between languages A and B, we need access to a parallel corpora between languages A and B. Unfortunately not all the language pairs in the world have available parallel corpora.

Consider, for instance, the Galician language. While we can find fairly easily parallel corpora with English, French, Portuguese or Spanish, it is not as easy to find parallel corpora with other languages like Arabic, Italian, Dutch or Danish.

Nevertheless, there are some parallel corpora from English, Portuguese or Spanish to these other more "exotic" languages. The question we want to answer is if it is possible to use two pair of transitive corpora, say Galician–English and English–Italian, to compute a Galician–Italian PTD. The choice of English as pivot language in our experiment is motivated by the high number of parallel corpora avalaible for English with very different languages. On the other hand, we feel that if it is possible to go from a Romance to a Germanic language, and then go back to a Romance language, then the same approach would hopefully retrieve even better results if we could find a pivot language in the same language family of source and target language.

For the better understanding of the concept of Probabilistic Translation Dictionaries, section 2 shows a simple example and discusses their structure details. Section 3 will formalize the triangulation (or composition) approach, and includes a full composition example. In section 4 we present some experiments on composing a Galician–English dictionary with an English–Italian dictionary, and evaluate the obtained results. Finally, we conclude on section 5 with some comments and future work.

2. Probabilistic Translation Dictionaries

Probabilistic Translation Dictionaries (PTD) are extracted automatically from (sentence-aligned) parallel corpora. The process is completely automatic and already proven to scale for big corpora. It results in a pair of dictionaries: one mapping words from the corpus source-language to its target-language, and another mapping words from the corpus target-language to its source-language[1].

Each dictionary maps words from a source-language S to a set of possible translations on a target-language T. Each possible translation have an associated probability measure:

$$\mathcal{T}\,(\text{codificada}) = \begin{cases} \text{codified} & 62.83\% \\ \text{uncoded} & 13.16\% \\ \text{coded} & 6.47\% \\ \ldots \end{cases}$$

[1] Refer to [4] for further discussion about why these methods compute a pair of dictionaries instead of just one.

Together with this information we keep track, for each word on the source-language, of the number of their occurrence in the corpus. Given that a parallel corpus alignment produces a pair of PTD it is possible to query the number of occurrences for any word (being it from the source or target language), and to compute the total number of tokens for each corpus.

We will define formally a PTD (one of the two extracted dictionaries) from a language A to a language B as:

$$\begin{array}{lcl} PTD & = & langA \hookrightarrow info \\ info & = & \begin{array}{ccc} count & : & int \\ trans & : & trans \end{array} \quad \times \\ trans & = & langB \hookrightarrow prob \\ langA & = & term \\ langB & = & term \end{array}$$

The extraction tools usually discard some translations during PTD extraction, given computer memory limitations, keeping track of the best k translations and discarding translations with probabilities bellow a specified threshold. In the context of this article, we used NATools [5] and its default configuration values: the algorithm computes a maximum of 8 translations per word (the more probable), and rejects translations below a probability threshold.

3. PTD Triangulation

As described earlier, a PTD maps words into some information. Therefore, we can consider that a PTD behaves just like a function, receiving a word, and returning the probable translations structure for this word.

Given two dictionaries, D_1 and D_2, which map respectively words from language A to language B and words form language B to language C, it is possible to apply some kind of function composition between the two dictionaries, creating a dictionary $D = D_1 \circ D_2$, which maps words from language A to language C. Check [13] for an alternative approach of dictionaries composition.

As PTD do not map words into words, but words into some information structure, some decisions should be made so that this composition can be performed:

- **What occurrence should have each word on the target dictionary?**
  At the moment, our decision is to use the original occurrence count for that word, from the first dictionary. Another option could be to discard the value, or to multiply it by some factor related to the occurrence of the pivot word (the word being used for the composition).

  In the future we plan to perform experiments on using the a factor based on the multiplication of occurrences from both source languages (on both dictionaries).

- **What probability should be associated to each possible translation?**
  Although we can argue that the events of translating from A to B, and from B to C are not independent, we decided to just multiply the translation probabilities from both dictionaries, as defined below in the composition formalization.

  One problem of this approach is that the obtained probabilities are smaller than the ones we would usually obtain with a direct extraction, given the probability multiplication. With this in mind, after the composition we perform the dictionary totalization: sum up all the translation probabilities, consider this total to be 100% and recompute each word translation probability.

For a better understanding consider the following diagram. It presents some of the possible translations for the Galician word "afluencia" in English, together with their translation probability in the Galician-English PTD. For each English translation, we queried the English–Italian PTD, and added the more probable Italian translations (and their translation probability). The last column, in bold, presents the probability for the Italian word to be a correct translation of the original Galician word.

$$\text{afluencia}\left\{\begin{array}{lrl} \text{influx} & 18.6\% & \left\{\begin{array}{lrr} \text{afflusso} & 48.9\% & \mathbf{= 9.1\%} \\ \text{flusso} & 12.7\% & \mathbf{= 2.4\%} \\ \text{flussi} & 4.7\% & \mathbf{= 0.9\%} \end{array}\right. \\ \text{flow} & 12.9\% & \left\{\begin{array}{lrr} \text{flusso} & 46.9\% & \mathbf{= 6.0\%} \\ \text{flussi} & 9.9\% & \mathbf{= 1.3\%} \\ \text{gravi} & 1.7\% & \mathbf{= 0.2\%} \end{array}\right. \\ \text{inflow} & 6.1\% & \left\{\begin{array}{lrr} \text{sfogo} & 24.2\% & \mathbf{= 1.5\%} \\ \text{afflusso} & 16.8\% & \mathbf{= 1.0\%} \\ \text{ascritto} & 14.7\% & \mathbf{= 0.9\%} \end{array}\right. \\ \text{flood} & 5.9\% & \left\{\begin{array}{lrr} \text{inondazioni} & 5.6\% & \mathbf{= 0.3\%} \\ \text{flusso} & 4.4\% & \mathbf{= 0.3\%} \\ \text{alluvione} & 2.8\% & \mathbf{= 0.2\%} \end{array}\right. \\ \text{flows} & 4.7\% & \left\{\begin{array}{lrr} \text{flussi} & 72.3\% & \mathbf{= 3.4\%} \\ \text{flusso} & 1.6\% & \mathbf{= 0.1\%} \\ \text{ondate} & 1.5\% & \mathbf{= 0.1\%} \end{array}\right. \end{array}\right.$$

As the Italian translations for each of the English word might repeat, these values should be summed up. So, the final version of the triangulation task would result in:

$$\text{afluencia}\left\{\begin{array}{lr} \text{afflusso} & 10.08\% \\ \text{flusso} & 08.73\% \\ \text{flussi} & 05.51\% \\ \text{sfogo} & 1.46\% \\ \text{ascritto} & 0.89\% \\ \text{inondazioni} & 0.33\% \\ \text{gravi} & 0.22\% \\ \text{alluvione} & 0.16\% \\ \text{ondate} & 0.07\% \end{array}\right.$$

For the sake of completeness, we present formalization of the composition operator in mathematical notation.

$$\begin{array}{l} \text{compose: } PTD \times PTD \longrightarrow PTD \\ \text{compose}(A, B) \stackrel{\text{def}}{=} \\ \quad \begin{pmatrix} w \\ \text{composeI}\left(A\left(w\right), B\right) \end{pmatrix}_{w \in \text{dom}(A)} \end{array}$$

$$\begin{array}{l} \text{composeI: } info \times PTD \longrightarrow info \\ \text{composeI}(A, B) \stackrel{\text{def}}{=} \\ \quad info\left(count(A), \text{composeT}\left(trans\left(A\right), B\right)\right) \end{array}$$

$$\begin{array}{l} \text{composeT: } trans \times PTD \longrightarrow trans \\ \text{composeT}(A, B) \stackrel{\text{def}}{=} \\ \quad \textbf{let } \; D = \text{dom}\left(A\right) \\ \quad \textbf{in } \; \begin{pmatrix} w \\ \text{p}\left(A, B, t, w\right) \end{pmatrix}_{t \in D, w \in \text{dom}(trans(B(t)))} \end{array}$$

$$\begin{array}{l} \text{p: } trans \times PTD \times str \times str \longrightarrow double \\ \text{p}(A, B, t, w) \stackrel{\text{def}}{=} \\ \quad \textbf{let } \; tB = trans\left(B\left(t\right)\right) \\ \quad \textbf{in } \; A\left(t\right) \times tB(w) \end{array}$$

4. Triangulation Evaluation

For our experiments we used two Galician–English parallel corpora, TECTRA and UNESCO, [14] which are part of the CLUVI Parallel Corpus (`http://sli.uvigo.es/CLUVI/`), and the English–Italian pair from EuroParl v5 [15] parallel corpora. Table 1 summarizes extracted dictionary sizes.

Corpus	CLUVI		EuroParl v5	
Lang. pairs[2]	GL–EN	EN–GL	EN–IT	IT–EN
Types	100 740	69 861	118 001	170 159
T. per word[3]	4.61	6.12	5.94	5.15
Average prob.[4]	56%	49%	52%	58%
Average occs.[5]	18	28	403	288

Table 1: Statistics for the DPTs used in the experiment.

Table 2 summarizes the composed dictionaries (after probabilities totalization), for both language directions, and the composed dictionaries after a drastic filtering process.

The filtering used the following heuristics:

- the dictionary were totalized (table 2 values for the simple composition were taken after the totalization process);
- Then, were removed:
  - non-word entries (symbols, numbers, etc.);
  - entries with words occurring less than 5 times;
  - translations with probabilities bellow 20%;
  - translations $t \in \mathcal{T}_1(w)$ unless $\exists w \in \mathcal{T}_2(t)$;

Dictionary	**Simple Comp.**		**Filtered Comp.**	
Languages	GL–IT	IT–GL	GL–IT	IT–GL
Types	88 211	149 521	4 511	4 559
T. per word	31.2	47.82	1.1	1.1
Average prob.	34%	32%	50%	50%
Average occs.	21	323	168	4 746

Table 2: Statistics for the triangulated dictionary, before and after the drastic filtering process.

The resulting dictionaries are quite small, but the process of enlarging them is simple: just loosen the limits. Nevertheless, these limits should be defined accordingly with the final dictionary application. For simple automated tasks, like cross language information retrieval, there are no big losses on precision using weaker dictionaries. In the other hand, if preparing a dictionary for automatic or human translation, we might prefer fewer words and higher translation quality. Table 3 summarizes a looser approach, using the previous heuristics, but removing only entries with less than 2 occurrences (so, ignoring words occurring just once), and removing translations with probabilities bellow 10%.

[2] Language pairs with Galician (GL), English (EN) and Italian (IT).

[3] Average number of translations per dictionary entry. NATools limit this value to 8, justifying the number of translations on the original dictionaries. During composition that limit does not exist.

[4] Average probability for best entry translations. Higher values mean better translation confidence (and lower number of possible translations per entry).

[5] Average number of word occurrences in the source corpus.

Dictionary	**Simple Comp.**		**Filtered Comp.**	
Languages	GL–IT	IT–GL	GL–IT	IT–GL
Types	88 211	149 521	10 559	10 781
T. per word	31.2	47.82	1.4	1.4
Average prob.	34%	32%	39%	38%
Average occs.	21	323	97	2 628

Table 3: Statistics for the triangulated dictionary, before and after the looser filtering process.

These two changes make the average probability for the first translation to drop (as we have much more entries with fewer translation probabilities), and the average number of occurrences to drop as well (as we are including a lot of new words with occurrences ranging from 2 to 4).

For the triangulation evaluation 100 entries were randomly selected from both filtered dictionaries. These entries were evaluated manually, with three distinct classes: good translations (ignoring inflection), bad translations, and doubtful translations (where the translation is almost good, but misses something, for example, incomplete translation of one word to two word translation). Table 4 show obtained results. The number of evaluated translation pairs is not 100, as some entries have more than one possible translation.

Filtering	**Drastic Filtering**		**Looser Filtering**	
Languages	GL–IT	IT–GL	GL–IT	IT–GL
Good	104	101	100	106
Doubtful	5	4	7	3
Bad	1	4	26	29
Precision	95%	93%	75%	73%

Table 4: Evaluation of composed dictionaries after drastic and looser filtering approaches.

Note that, while the looser filtering approach increased the number of bad and doubtful entries, the number of good entries is almost the same.

Follows some examples of entries obtained with this method. The bad translations were tagged with a star. While those translations are bad, their presence is easy to understand.

$$
\begin{array}{rll}
\text{abandonados} & \left\{\begin{array}{l}\text{abbandonato}\\ \text{abbandonate}\\ \text{abbandonata}\end{array}\right. & \begin{array}{l}25.3\%\\ 11.5\%\\ 10.5\%\end{array}\\
\text{advertencias} & \left\{\begin{array}{l}\text{avvertimenti}\\ \text{avvertenze}\end{array}\right. & \begin{array}{l}41.8\%\\ 15.5\%\end{array}\\
\text{impostos} & \left\{\begin{array}{l}\text{imposte}\\ \text{fiscale}^\star\\ \text{tasse}^\star\end{array}\right. & \begin{array}{l}23.3\%\\ 21.2\%\\ 12.8\%\end{array}\\
\text{xenital} & \left\{\begin{array}{l}\text{mutilazioni}^\star\\ \text{genitali}\end{array}\right. & \begin{array}{l}20.8\%\\ 15.8\%\end{array}
\end{array}
$$

5. Conclusions

Scarcety of linguistic resources is one of the problems of minoritized languages. In this paper, we have suggested a solution for that problem in the field of bilingual dictionaries construction, using probabilistic translation dictionaries (PTD) extracted from transitive parallel corpora.

With that aim, we have analyzed the issues concerning to the construction of a Galician–Italian probabilistic dictionary.

As we have seen, PTDs are bilingual dictionaries extracted automatically from parallel corpora on a statistic base. Nevertheless, while we can expect easily to find parallel corpora from Galician to Portuguese, English or Spanish, it is not that easy (or possible) to find available parallel corpora from Galician to Italian.

Our claim is that it is possible to use some kind of transitivity to create translation dictionaries. So we have shown how a Galician-Italian PTD can be constructed without a Galician–Italian parallel corpus, by the combination or triangulation of two other PTDs: a Galician–English PTD and an English–Italian PTD extracted, respectively, from a GL–EN parallel corpus and an EN–IT one.

Next we have evaluated the performance of the triangulation and, as expected, the final combined dictionaries are not as good as the ones extracted directly from parallel corpora. Even so, we can conclude that the process is great for bootstrapping dictionaries when better data is not available.

At the moment we are working on the treatment of the extracted combined dictionaries to populate an Italian–Galician bilingual dictionary similar to the on-line corpus-based CLUVI English–Galician Dictionary developed at the University of Vigo [16] (`http://sli.uvigo.es/dicionario`) This work will include the automatic extraction of usage examples candidates, and the conversion of the filtered version (looser approach) of the PTD to the XML format being used by CLUVI dictionaries. The final product of this lexicographic work will require a process of human revision, during which better measures on the precision of the triangulation procedure will be calculated.

As an option to make the dictionary larger and better, we could use more than one composition (for instance, $GL \rightarrow EN \rightarrow IT$, $GL \rightarrow ES \rightarrow IT$ and $GL \rightarrow PT \rightarrow IT$, and other paths that could be found) to retrieve a set of dictionaries that can be compared and merged in an enhanced Italian–Galician PTD.

6. Acknowledgements

This work was partially funded by the project *Português em paralelo com seis línguas (Português, Español, Russian, Français, Italiano, Deutsch, English)* grant `PTDC/CLE-LLI/108948/2008` from *Fundação para a Ciência e a Tecnologia*; and by the project *Desenvolvemento e explotación de recursos integrados da lingua galega* grant `INCITE08PXIB302185PR` from *Consellería de Innovación e Industria, Xunta de Galicia*.

7. References

[1] P. Fung and K. Church, "Kvec: A new approach for aligning parallel texts," Kyoto, Japan, pp. 1096–1102, 1994.

[2] N. Varma, "Identifying word translations in parallel corpora using measures of association," Master's thesis, University of Minnesota, 2002.

[3] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[4] D. Hiemstra, "Using statistical methods to create a bilingual dictionary," Master's thesis, Department of Computer Science, University of Twente, August 1996.

[5] A. M. Simões and J. J. Almeida, "NATools – a statistical word aligner workbench," *Procesamiento del Lenguaje Natural*, vol. 31, pp. 217–224, September 2003. [Online]. Available: http://alfarrabio.di.uminho.pt/~albie/publications/sepln2003.pdf

[6] X. G. Guinovart and E. S. Fontenla, "Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos," *Procesamiento del Lenguaje Natural*, vol. 33, pp. 133–140, 2004. [Online]. Available: http://webs.uvigo.es/sli/arquivos/sepln04a.pdf

[7] H. M. Caseli, M. G. V. Nunes, and M. L. Forcada, "Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts," *Procesamiento del Lenguaje Natural*, September 2005.

[8] A. Simões and J. J. Almeida, "NatServer: a client-server architecture for building parallel corpora applications," *Procesamiento del Lenguaje Natural*, vol. 37, pp. 91–97, September 2006. [Online]. Available: http://alfarrabio.di.uminho.pt/~albie/publications/sepln06.pdf

[9] ——, "Bilingual terminology extraction based on translation patterns," *Procesamiento del Lenguaje Natural*, vol. 41, pp. 281–288, September 2008.

[10] X. G. Guinovart and A. Simões, "Terminology extraction from English-Portuguese and English-Galician parallel corpora based on probabilistic translation dictionaries and bilingual syntactic patterns," in *I Iberian SLTech 2009*, A. Teixeira, M. S. Dias, and D. Braga, Eds., Porto Salvo, Portugal, September, 3–4 2009, pp. 13–16. [Online]. Available: http://webs.uvigo.es/sli/arquivos/IberianSLT09.pdf

[11] A. Simões and J. J. Almeida, "Bilingual example segmentation based on Markers Hypothesis," in *I Iberian SLTech 2009*, A. Teixeira, M. S. Dias, and D. Braga, Eds., Porto Salvo, Portugal, September, 3–4 2009, pp. 95–98.

[12] W. Kraaij, "TNO at CLEF-2001: Comparing Translation Resources," *Lecture Notes in Computer Science*, pp. 78–93, 2002.

[13] ——, "Exploring transitive translation methods," in *Proceedings of DIR*, 2003.

[14] X. G. Guinovart, "A investigación en lexicografía e terminoloxía no Corpus Lingüístico da Universidade de Vigo (CLUVI) e no Corpus Técnico do Galego (CTG)," in *A lexicografía galega moderna. Recursos e perspectivas*, E. González, A. Santamarina, and X. Varela, Eds. Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, 2008, pp. 197–216. [Online]. Available: http://webs.uvigo.es/sli/arquivos/sli_ilg07.pdf

[15] P. Koehn, "EuroParl: A parallel corpus for statistical machine translation," in *Proceedings of MT-Summit*, 2005, pp. 79–86.

[16] X. G. Guinovart, E. D. Rodríguez, and A. Álvarez Lugrís, "Aplicacións da lexicografía bilingüe baseada en córpora na elaboración do Dicionario CLUVI inglés-galego," *Viceversa*, vol. 14, pp. 71–87, 2008. [Online]. Available: http://webs.uvigo.es/sli/arquivos/viceversa_clig2ed_2008.pdf

Spanish generation from Spanish Sign Language using a phrase-based translation system

Verónica López, Rubén San-Segundo, Raquel Martín, Juan Manuel Lucas, Julián David Echeverry

Speech Technology Group at Universidad Politécnica de Madrid
veronicalopez@die.upm.es

Abstract

This paper describes the development of a Spoken Spanish generator from Spanish Sign Language (LSE – Lengua de Signos Española) in a specific domain: the renewal of Identity Document and Driver's license. The system is composed of three modules. The first one is an interface where a deaf person can specify a sign sequence in sign-writing. The second one is a language translator for converting the sign sequence into a word sequence. Finally, the last module is a text to speech converter. Also, the paper describes the generation of a parallel corpus for the system development composed of more than 4,000 Spanish sentences and their LSE translations in the application domain. The paper is focused on the translation module that uses a statistical strategy with a phrase-based translation model, and this paper analyses the effect of the alignment configuration used during the process of word-based translation model generation. Finally, the best configuration gives a 3.90% mWER and a 0.9645 BLEU.

Index Terms: Automatic Statistic Translation, Sign Language Translation, Spoken Spanish Generation, Phrase-based Translator, Type of Word Alignment.

1. Introduction

There are approximately 70 million people with hearing deficiencies in the world (information from World Federation of the Deaf). Deafness brings about significant problems in communicating, because deaf people cannot hear and most of them are unable to use written languages, having big problems in understanding and expressing this way (they have problems with verb tenses, concordances of gender and number, etc., and they have difficulties when creating a mental image of abstract concepts). This fact can cause deaf people to have problems in accessing to information, education, job, social relationship, culture, etc., because they use a sign language for communicating and there is not a sufficient number of sign-language interpreters and communication systems for it.

In the same way, deaf people have problems when they want to access to public services, for example, to renew their Driver's License (DL) or Identity Document (ID). In general, government employees do not know LSE, so a deaf person needs an interpreter for accessing to this service.

In 2007, the Spanish Government accepted the Spanish Sign Language (LSE: Lengua de Signos Española) as one of the official languages in Spain, defining a plan to invest in resources in this language and that it becomes not only the natural language for deaf people, but also an instrument when communicating with hearing people, or accessing information. The translation system described in this paper is part of this government plan and its goal is to help deaf people to communicate with government employees in two specific domains: the renewal of ID and DL.

2. Spanish Sign Language

Spanish Sign Language (LSE), just like other sign languages, has a visual-gestural channel, but it also has grammatical characteristics of oral languages that shares with exclusive characteristics of sign languages. In linguistic terms, sign languages are as complex as oral languages, despite the common misconception that they are a "simplification" of oral languages. The main characteristics of LSE and the differences with Spanish are as follows [11]:

- Predication order: LSE has a SOV (subject-object-verb) order in opposite to SVO (subject-verb-object) Spanish order. For example:
  Spanish: *Juan ha comprado las entradas*
  LSE: *JUAN ENTRADAS COMPRAR*
- Gender in LSE is not usually specified.
- For specifying verb tenses, the verb tense can be added in parentheses next to the gloss, for example "*USAR (FUT.) (to use in future)*".
- For representing a negative sentence, it is added to the verb in infinitive, the gloss "*NO*", for example "*PODER NO*" (cannot).
- In LSE, also there are spelling for representing names or unknown words and this is indicated with "dl" previous to the spelled word, for example, "*dlJUAN*" for spelling the name "*Juan*".
- The use of classifiers is common in LSE: signs that indicate actions, places, etc., and that are denoted with the prefix "*CL*" and a letter that indicates the classifier's type (for example, place). Some classifiers are "*CLL-ACERCARSE*", "*CLD-GRANDE*", etc. There are not classifiers in Spanish.
- Spanish has an informative style (without topics) and LSE has a communicative style (with topics).
- In LSE, there can be concordances between verbs and subject, receiver or object and even subject and receiver, but in Spanish there can be only concordance between verb and subject.
- Articles are used in Spanish, but not in LSE
- Plural can be descriptive in LSE, but not in Spanish.
- There is a difference between absent and present third person in LSE, but there is not absent third person in Spanish.
- In LSE, there is the possibility of using double reference, not in Spanish.
- LSE is a language with ample flexibility, and homonymy between substantive and adjective is usual, so most nouns can be adjectives and vice versa. But there are few cases in Spanish.
- In Spanish, there is a copula in non-verbal predications (the verb "to be", *ser* and *estar* in Spanish), but there is not in LSE (except some locative predications).

- There is a difference between inclusive and exclusive quantifier in LSE, but not in Spanish.
- There are Spanish impersonal sentences with "*se*" pronoun, but not in LSE.
- Iconicity: signs resemble to concept that represent. If written LSE is analysed, glosses have semantic information principally.
- It is important to comment that LSE is more lexically flexible than Spanish, and it is perfect for generating periphrasis through its descriptive nature and because of this, LSE has fewer nouns than Spanish.
- LSE has less gloss per sentence (4.4) than Spanish (5.9).

3. State of the Art

Several groups have generated corpora for sign language research. Some examples are: a corpus composed of more than 300 hours from 100 speakers in Australian Sign Language [1]. The RWTH-BOSTON-400 Database that contains 843 sentences with about 400 different signs from 5 speakers in American Sign Language with English annotations [2]. The British Sign Language Corpus Project tries to create a machine-readable digital corpus of spontaneous and elicited British Sign Language (BSL) collected from deaf native signers and early learners across the United Kingdom [3]. And a corpus developed at Institute for Language and Speech Processing (ILSP) and that contains parts of free signing narration, as well as a considerable amount of grouped signed phrases and sentence level utterances [4].
The best performing translation systems are based on various types of statistical approaches ([5]; [6]), including example-based methods [7], finite-state transducers [8] and other data driven approaches. Another important effort in machine translation has been the organization of several Workshops on Statistical Machine Translation (SMT). As a result of these workshops, there are two free machine translation systems called Moses (http://www.statmt.org/moses/) and Joshua (http://cs.jhu.edu/~ccb/joshua/).
About speech generation from sign language, in the Computer Science department of the RWTH, Aachen University, P. Dreuw supervised by H. Ney is making a significant effort in recognizing continuous sign language with a new vision-based technology ([9]; [10]).
This paper describes the development of a Spanish Sign Language into Spoken Spanish translation system in a real domain: the Driver's License and Identity Document renewal. Specifically, the paper is focused on translation module between a sequence of written signs and written Spanish.

4. Database

In order to develop a translation system focused on the domain of renewal of ID and DL, a database has been generated. This database has been obtained with the collaboration of Local Government Offices where the mentioned services (renewal of ID and DL) are provided. During three weeks, the most frequent explanations (from government employees) and the most frequent questions (from the user) were taken down and more than 5,000 sentences were noted.
These 5,000 sentences were analysed because not all of them refer to ID or DL, so sentences were selected manually in order to develop a system in a specific domain. Finally, 1360 sentences were selected: 1,023 pronounced by government employees and 337 by users. These sentences were translated into LSE, both in text (sequence of glosses) and in video, and compiled in an excel file. This corpus was increased to 4,080 by incorporating different variants for Spanish sentences, maintaining the LSE translation.
The main features of the corpus are shown in Table 1.

	ID		DL	
Government employee	**Spanish**	**LSE**	**Spanish**	**LSE**
Sentence pairs	1,425		1,641	
Different sentences	1,236	389	1,413	199
Running words	8,490	6,282	17,113	12,741
Vocabulary	652	364	527	237
User	**Spanish**	**LSE**	**Spanish**	**LSE**
Sentence pairs	531		483	
Different sentences	458	139	389	93
Running words	2,768	1,950	3,130	2,283
Vocabulary	422	165	294	133

Table 1: *Main statistics of the corpus.*

For the system development, two types of files were generated from the database: text files and sign files. Text files are composed of Spanish sentences of the parallel corpus and sign files contain their LSE translations (LSE sentences made up of glosses –capital words who represent signs).
These pairs of files were divided randomly into three sets: training (75%), development (12.5%) and test (12.5%), carrying out a round-robin evaluating process. The results presented in this paper are the average of this round robin, increasing the reliability of results this way.

5. Spanish generation from LSE

The spoken Spanish generation system converts a sign sequence (LSE sequence) into spoken Spanish. It is composed of three modules (Figure 1).

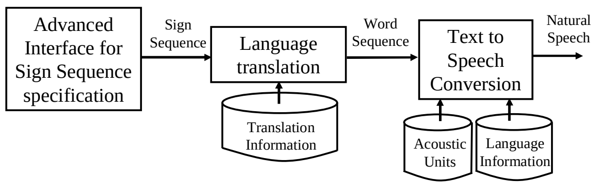


Figure 1: *Diagram of Spanish generation system.*

The first module is a visual interface for specifying the sign sequence. This interface includes several tools for sign specification: avatar for sign representation (to verify that sign corresponds to the gloss), prediction mechanisms, calendar and clock for date or time definitions, spelling, frequent questions, etc. With this visual interface the Deaf can build a sign sentence that will be translated into Spanish and spoken to a hearing person. The sign sequence is specified in glosses but signs can be searched by using specific sign characteristics in HamNoSys notation [12].
The second module converts a sign sequence into a word sequence with a statistical translation strategy.
The last module converts the word sequence into spoken Spanish by using a commercial Text to Speech converter. In this project, the Loquendo system has been used (http://www.loquendo.com/en/).
Visual interface is shown below in Figure 2.

Figure 2: *Visual interface for sign sequence specification.*

This paper describes the statistic translation system based on a phrase-based model.

5.1. Phrase-based translation

Phrase-based translation system uses a model of word sequences that is obtained from the alignment of parallel corpus (Figure 3). For the corpus alignment, *GIZA++* [13] program is used, and after, a set of phrases and their translation probabilities are obtained with *Phrase-Extract* and *Phrase-Score* programs.

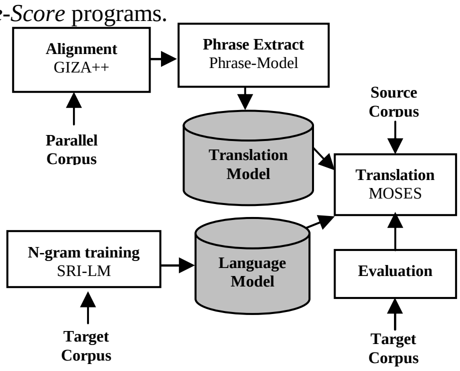


Figure 3: *Phrase-based translation module.*

Also, a 3-gram language model is incorporated, using the *SRI-LM* toolkit [14].
Translation model obtains $p(t|s)$ (probability of *t (target language)*) given *s (source language)*) with Bayes's theorem:

$$p(t|s) \approx p(s|t) \cdot p(t)$$

$p(s|t)$ is the probability of *s* given *t* (translation model) and $p(t)$ is the probability of seeing *t* (language model).
When translation and language models are generated, they have probability weights in translation that are not the best. Because of this, several translations are carried out with the development set in order to find probability weights that provide the best results. For translating, the *MOSES* decoder is used [15].
Finally, using the translation and language models and their probability weights, automatic translation is carried out with *MOSES* for evaluating the translation system.

5.2. Analysis of the alignment configuration

In order to generate the word alignment, *GIZA++* obtains the word alignment in both directions: source-target and target-source (LSE-Spanish and Spanish-LSE). Later, a final alignment is generated from a combination of previous alignments. Figure 4 shows different alignments between a pair of sentences in Spanish and LSE and their alignment points (each black box represents a word and a sign both aligned). The combination can be:

- **Source-Target (ST)**: Only the source-target (LSE-Spanish) alignment is considered. In this configuration, alignment is guided by signs: each sign in LSE is aligned with a Spanish word and it is possible that some word was unaligned.
- **Target-Source (TS)**: Target-source (Spanish-LSE) is the only considered alignment. In this configuration, alignment is guided by words: each Spanish word is aligned with a sign in LSE and it is possible that some sign was unaligned.
- **Union (U)**: In this case, alignment points of the union of both directions (source-target and target-source) are taken. This way, additional alignment points are obtained, having more examples for training the word translation model, but also, alignment quality is worse (more variability).
- **Intersection (I)**: In this case, alignment points of the intersection of both directions (source-target and target-source) are selected. This is the most strict configuration: less alignment points are obtained, but they are more reliable. This is not a good configuration if there is not a sufficient number of sentences for training.
- **Grow (G)**: In this configuration, alignment points of intersection are used to train the word translation model and also the adjoining points of union. This configuration is an intermediate solution between union and intersection, seeking a compromise between quality and quantity of alignment points.
- **Diagonal Grow (DG)**: In this configuration, alignment points of intersection are considered and also the adjoining points of union, but only adjoining points in diagonal.
- **Final Diagonal Grow (FDG)**: In this configuration, alignment points of intersection are taken and also the adjoining points of union, but only adjoining points in diagonal. And finally, if there is any word or sign unaligned, it is taken the corresponding union alignment point.

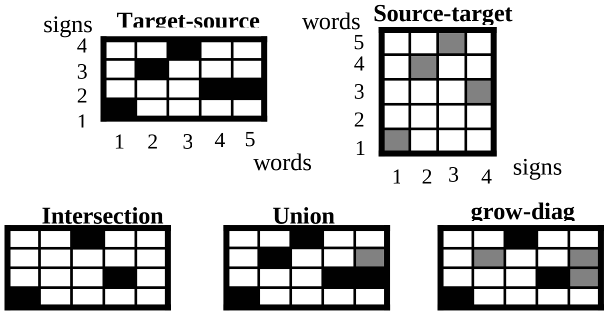


Figure 4: *Different alignment combinations.*

In order to analyse the effect of the alignment in final results, different alignment configurations were tested.

6. Evaluation

Table 2 and Figure 5 show the different results for each alignment configuration: mWER (multiple references Word Error Rate), BLEU (BiLingual Evaluation Understudy) and NIST. BLEU and NIST measures have been computed using the NIST tool (mteval.pl). Also, it is indicated the percentage of deletions (D), substitutions (S) and insertions (I) in translated sentences.

	mWER (%)	D (%)	S (%)	I (%)	BLEU	NIST
I	8.41	5.29	1.38	1.75	0.9252	11.7069
ST	6.52	4.28	1.09	1.14	0.9397	11.8033
DG	6.39	3.54	1.32	1.53	0.9430	11.8022
U	5.66	2.36	1.96	1.33	0.9459	11.7416
G	5.61	2.34	1.99	1.28	0.9459	11.7416
FDG	4.84	1.75	2.02	1.07	0.9520	11.7218
TS	**3.90**	**1.68**	**1.34**	**0.89**	**0.9645**	**11.9020**

Table 2: *Results of phrase-based translation system using different alignment configurations.*

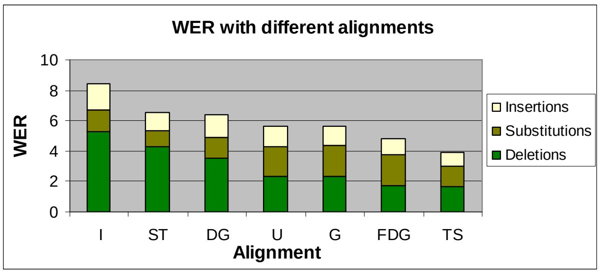


Figure 5: *Comparison of results of phrase-based translation system using different alignment configurations.*

As Table 2 and Figure 5 show, the best alignment is **target-source**: alignment is guided by words in this case. The main improvement is due to a less number of deletions and these deletions are important in translation because system translates from a language with less tokens per sentence (4.4 in LSE) to a language with more tokens per sentence (5.9 in Spanish). On the other hand, it can be observed that the worst result is given by intersection alignment, because important alignment points of target-source are deleted (looking at the Table 2, most mistakes are deletions). As additional points of target-source are added, results improve (deletions are reduced), and finally, with target-source the best result is obtained, giving a **3.90% mWER** and a **0.9645 BLEU**.

7. Conclusions

The paper has described the development of a spoken Spanish generation system from Spanish Sign Language. This system is focused on the application domain of renewal of Identity Document and Driver's license and it is composed of three modules: an interface for specifying the sign sequence, a statistical translation module for converting the sign sequence into a word sequence and a text to speech converter. The paper has been focused on this statistical translation module that uses a phrase-based translation model. Specifically, the paper has studied the alignment effect of the translation model in final results. The best alignment configuration is target-source where alignment is guided by words: less deletions in translated sentences are produced. Also, it can be observed that according as target-source alignment points are deleted, word error rate increases, because deletions in translated sentences increase and these deletions are important: system translates from LSE (with less tokens per sentence) to Spanish (with more tokens per sentence). Finally, it is obtained a 3.90% mWER and a 0.9645 BLEU.

8. Acknowledgements

The authors want to thank the eSIGN (Essential Sign Language Information on Government Networks) consortium for permitting the use of the eSIGN Editor and the 3D avatar in this research work. This work has been supported by Plan Avanza Exp Nº: PAV-070000-2007-567, ROBONAUTA (MEC ref: DPI2007-66846-c02-02) and SD-TEAM (MEC ref: TIN2008-06856-C05-03) projects.

9. References

[1] Johnston T., 2008. "Corpus linguistics and signed languages: no lemmata, no corpus". 3rd Workshop on the Representation and Processing of Sign Languages, June 1. 2008.

[2] Dreuw P., Neidle C., Athitsos V., Sclaroff S., and Ney H. 2008a. "Benchmark Databases for Video-Based Automatic Sign Language Recognition". LREC, Marrakech, Morocco.

[3] Schembri A.,2008 "British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox". Deafness Cognition and Language Research Centre, University College London. LREC 2008.

[4] Efthimiou E., and Fotinea, E., 2008 "GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI" LREC 2008.

[5] Och J., Ney. H., 2002. "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation". Annual Meeting of the Ass. For Computational Linguistics (ACL), Philadelphia, PA, pp. 295-302. 2002.

[6] Mariño J.B., Banchs R., Crego J.M., Gispert A., Lambert P., Fonollosa J.A., Costa-Jussà M., 2006. "N-gram-based Machine Translation", Computational Linguistics, Association for Computacional Linguistics. Vol. 32, nº 4, pp. 527-549.

[7] Sumita E., Y. Akiba, T. Doi et al. 2003. "A Corpus-Centered Approach to Spoken Language Translation". Conf. of the Europ. Chapter of the Ass. For Computational Linguistics (EACL), Budapest, Hungary. pp171-174. 2003.

[8] Casacuberta F., E. Vidal. 2004. "Machine Translation with Inferred Stochastic Finite-State Transducers". Computational Linguistics, Vol. 30, No. 2, pp. 205-225, June 2004.

[9] Dreuw P., Ney H., Martinez G., Crasborn O., Piater J., Miguel Moya J., and Wheatley M., 2010 "The SignSpeak Project - Bridging the Gap Between Signers and Speakers". In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, May 2010a.

[10] Dreuw P., Forster J., Gweth Y., Stein D., Ney H., Martinez G., Verges Llahi J., Crasborn O., Ormel E., Du W., Hoyoux T., Piater J., Moya Lazaro JM, and Wheatley M. 2010 "SignSpeak - Understanding, Recognition, and Translation of Sign Languages". In 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Valletta, Malta, May 2010b.

[11] Herrero, A. 2009. "Gramática didáctica de la Lengua de Signos Española (LSE)". 2009.

[12] Prillwitz, S., R. Leven, H. Zienert, T. Hanke, J. Henning, et-al. 1989. "Hamburg Notation System for Sign Languages – An introductory Guide". International Studies on Sign Language and the Communication of the Deaf, Vol. 5. Uni. of Hamburg.

[13] Och J., Ney H., 2000. "Improved Statistical Alignment Models". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China.

[14] Stolcke A., 2002. "SRILM – An Extensible Language Modelling Toolkit". Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver.

[15] Koehn, Philipp. 2010. "Statistical Machine Translation". Cambridge University Press.

Multiclass Classification for Morphology Generation in Statistical Machine Translation

Adolfo Hernández Huerta, Enric Monte Moreno, José B. Mariño Acebal

Department of Signal Theory and Communications
TALP Research Center (UPC)
Barcelona 08034, Spain
(adolfo.hernandez, enric.monte, jose.marino)@upc.edu

Abstract

We present a system for multiclass classification of simplified morphology of Spanish verbs within the framework of morphology generation for Statistical Machine Translation (SMT) from English into Spanish. In previous works it was proved that, when statistically translating from English into Spanish, the richness of morphology of the target language affects the translation models at training time by creating data sparseness. In order to determine the correct morphology of the Spanish translation we use a hierarchical set of classifiers through a Decision Directed Acyclic Graph (DDAG) structure, each decision-node operates with a classifier which is a Support Vector Machine (SVM). This structure is justified because it allows to introduce prior information about the difficulty of the task. The classification results are analyzed and commented.

Index Terms: morphology, machine learning, statistical machine translation

1. Introduction

Despite the fact that initially SMT systems ignored any linguistic analysis and worked at the surface level of word forms, there has been a growing effort to introduce linguistic knowledge into their statistical framework. It is clear that linguistic information has the potential to improve the performance of SMT systems, especially when limited amounts of parallel training data sets are available. However, incorporating linguistic, morphological, syntactic and semantic information into the statistical framework of SMT is a hard problem.

In particular, languages with rich morphology pose significant challenges for natural language processing. In highly inflected languages, the extensive use of inflection (to express agreement, gender, case, etc.), derivation, and composition leads to a huge vocabulary, forcing the translation model to learn different translation probability distributions for all inflected forms of nouns, adjectives or verbs, suffering thus from data sparseness. SMT systems estimated from parallel text are affected by this fact because it is impossible to train all forms from the training corpora. It is also important to point out that obtaining good performance in SMT systems when translating between languages with different morphological richness is a challenging task; especially when translating into a richer morphology language because the target language is represented by a larger vocabulary set, making decisions harder for SMT systems (e.g. higher perplexity in translation and target language models). However, different strategies have to be tackled when translating in the opposite direction (i.e. from a richer morphology language), where sparsity problems may arise in the source language (higher percentage of out-of-vocabulary (OOV) words, fewer translation examples for each input word, etc.).

Due to the above discussion, it is reasonable to find challenges related with morphology in SMT systems whose language pairs are English and any Romance-family language (Portuguese, Catalan, Galician, Italian, French, etc.) or pairs such as English and Arabic, Finnish, or German, to name a few. Then, the linguistic properties of the pair of languages and the translation direction pose severe limitations in most of the SMT tasks, such as word alignment and modeling. Several efforts are being done in the community to overcome such constraints by analyzing language specific problems and their impact on statistical translation as well as introducing some linguistic information in statistical models.

An important work was developed by Nießen and Ney [1], where, for a German-to-English task, several transformations of the source string are proposed, leading to an increased translation performance. These transformations include compound word separation, reordering of separated verb prefixes, and word mapping to word plus POS in order to distinguish articles from pronouns, among others. The same pair of languages and translation direction are used by Nießen and Ney [2] and Corston-Oliver and Gamon [3]. In the former, hierarchical lexicon models including base form and POS information are introduced, as well as other morphology-based data transformations. In the latter, inflectional normalization was achieved, leading to improvements in the perplexity of IBM translation models and reducing alignment errors. Aiming for a more general approach to deal with language-specific challenges, Koehn and Hoang [4] introduced factored translation models that can efficiently integrate morpho-syntactic information into phrase-based SMT. This framework adds in the model a vector of factors that represent different levels of annotation: word, lemma, POS, morphology, word class.

This work extends the research of de Gispert and Mariño [5], which dealt with the problem of the morphology derivation on Ngram-based Statistical Machine Translation (SMT) models from English into a morphology-rich language such as Spanish. It was shown that some Spanish morphology information could be introduced into simplified morphology translation hypotheses by means of an independent model. This approach is depicted in Figure 1. The translation system was proposed as a cascade-system integrated by a SMT system and a morphology generator. The SMT system consisted of simplified morphology translation models trained through a corpus with morphology simplification of words. From this study, several types of morphology simplification schemes were applied, depending

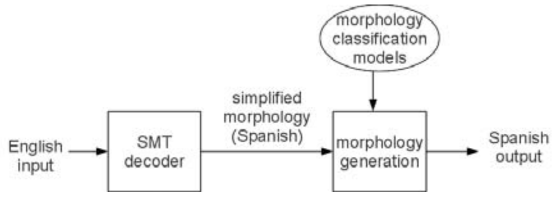


Fig. 1: *Translation architecture for SMT with morphology generation*

on which Part-Of-Speech category was modified (verbs, nouns, adjectives, etc.); it was concluded that the main source of potential improvement lied in verb form morphology. After the translation, Spanish morphology information is introduced into the simplified translation. For this purpose, a set of relevant features for each Spanish verb base form was defined in order to train statistical classifiers based on machine learning techniques, specifically through Adaboost, that used as base classifier a decision tree [6]. High accuracy scores were obtained when generating Spanish verb person, number and gender information, resulting in a significant improvement of final translation scores.

This paper is organized as follows. Section 2 briefly outlines the SMT system, the morphology generation and the multiclass classification approach. Section 3 reports and discusses the experimental results. Finally, Section 4 sums up the main conclusions from the paper.

2. System description

2.1. Ngram-based SMT system

The translation system implements a log-linear model in which a foreign language sentence $f_1^J = f_1, f_2 \ldots, f_J$ is translated into another language $e_1^I = e_1, e_2 \ldots, e_I$ by searching for the translation hypothesis $\hat{e}_1^I$ maximizing a log-linear combination of several feature models [7]:

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m \left(e_1^I, f_1^J \right) \right\} \quad (1)$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

The core part of the system constructed in that way is a translation model, which is based on bilingual n-grams. It actually constitutes an Ngram-based language model of bilingual units (called tuples), which approximates the joint probability between the languages under consideration. The procedure of tuples extraction from a word-to-word alignment according to certain constraints is explained in detail by Mariño et al. [8].

The Ngram-based approach differs from the phrase based SMT mainly by distinct representating of the bilingual units defined by word alignment and using a higher order HMM of the translation process. While regular phrase-based SMT considers context only for phrase reordering but not for translation, the N-gram based approach conditions translation decisions on previous translation decisions.

The translation system, besides the bilingual translation model, which consists of a 4-gram LM of tuples with Kneser-Ney discounting (estimated with SRI Language Modeling Toolkit[1]), implements a log-linear combination of five additional feature models:

- a target language model (a 4-gram model of words, estimated with Kneser-Ney smoothing);
- a POS source language model (a 4-gram model of tags with Good-Turing discounting);
- a POS target language model (a 4-gram model of tags with Good-Turing discounting);
- a word bonus model, which is used to compensate the system's preference for short output sentences;
- a source-to-target lexicon model and a target-to-source lexicon model, these models use word-to-word IBM Model 1 probabilities [9] to estimate the lexical weights for each tuple in the translation table.

Decisions on the particular LM configuration and smoothing technique were taken on the minimal-perplexity and maximal-BLEU bases.

The decoder (called MARIE), an open source tool[2], implementing a beam search strategy was used in the translation system.

Given the development set and references, the log-linear combination of weights was adjusted using a simplex optimization method (with the optimization criteria of the highest BLEU score) and an n-best re-ranking just as described in http://www.statmt.org/jhuws/. This strategy allows for a faster and more efficient adjustment of model weights by means of a double-loop optimization, which provides significant reduction of the number of translations that should be carried out.

2.2. Morphology simplification

In the SMT system, after standard word alignment and tuple extraction, target language words (Spanish) are substituted with their simplified morphology forms. Then the bilingual N-gram translation model is estimated with these new tuples. The result is a standard bilingual model translating English into simplified morphology Spanish. The morphology simplification module produces a set of samples whose correct morphology is known, as they belong to the training corpus; these samples can be used to estimate morphology classification models.

In the case of simplification of information about person and number for Spanish verbs, the verb form, for instance, *'apoyen'* is transformed into '*VMSPpn[apoyar]*', indicating simplified Part of Speech (POS) and base form. Under this simplification, the POS keeps information on word category (*'VM'*: Main Verb), mode and tense (*'SP'*: subjunctive, present), whereas *'p'* and *'n'* represent any person and number. Furthermore, as the correct person and number for this verb is known beforehand, it also serves as the class label during the training phase of the classifier that generates the morphology.

2.3. Multiclass classification

Morphology generation is implemented by means of classification models which, making use of a set of relevant features for each simplified morphology word and its context, generates its appropriate morphology. In order to tackle this task, we use a Decision Directed Acyclic Graph (DDAG), which combines many two-class classifiers into a multiclassification task. The use of this structure is justified because it allows to introduce

[1] http://www-speech.sri.com/projects/srilm/
[2] http://www.talp.cat/talp/index.php/ca/recursos/eines/marie

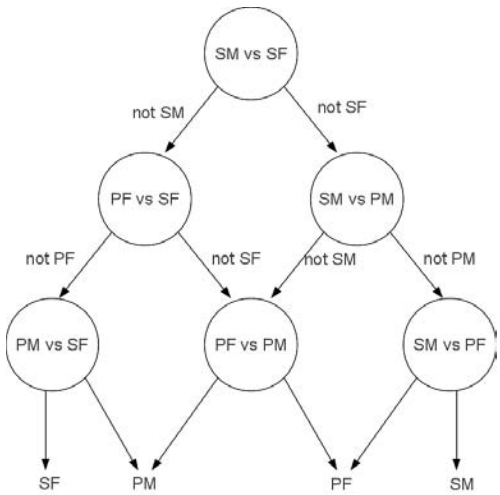


Fig. 2: *Decision DAG to find the best class out of four classes, it was the morphology related to gender and number*

EPPS corpus		sent.	words	vcb	avg.len.
train	Eng	1.40 M	39.29 M	121.07 k	28.02
	Spa		41.76 M	157.66 k	29.79
dev.	Eng	1996	58.63 k	6.54 k	29.37
test	Eng	1094	26.91 k	3.95 k	24.60

Tab. 1: *English-Spanish European Parliament corpus statistics*

information about the difficulty of the task. Instead of classifying one class versus all the others it does a pairwise comparison. As the set of classifiers are organized in a tree structure, the upper levels are assigned to the pairs of classes that have the lower error rate and also that the number of samples of each class is balanced (i.e. approximately the same number of examples per class).

The description of the structure is as follows. For an N-class problem, the DDAG contains N(N-1)/2 nodes, one for each pair of classes (one-vs-one classifier). A DAGSVM algorithm is proposed by Platt et al. [10], it proved to be superior to other multiclass SVM algorithms in both training and evaluation time. A DAGSVM places one-vs-one SVMs into the nodes of a DDAG. An example of a structure of the DDAG is shown in Figure 2.

3. Experiments

3.1. Database

This work was carried out on a large-data English-to-Spanish task, defined by a corpus containing official transcriptions of the European Parliament Plenary Sessions (EPPS), whose statistics are presented in Table 1. This corpus is available through the ACL-WMT evaluation campaign of year 2009[3].

3.2. Classification of Verb Forms

In order to generate morphology, two subcategories are distinguished: a) verb forms whose person and number (PN) information is missing (i.e. 1st person singular (1S), 2nd person singular (2S), and so on); b) verb forms whose number and gen-

[3] http://www.statmt.org/wmt09

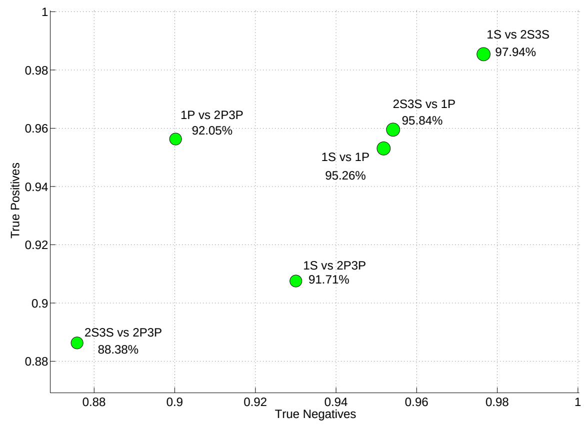


Fig. 3: *Spatial relationship between true positives, true negatives and accuracies for person and number*

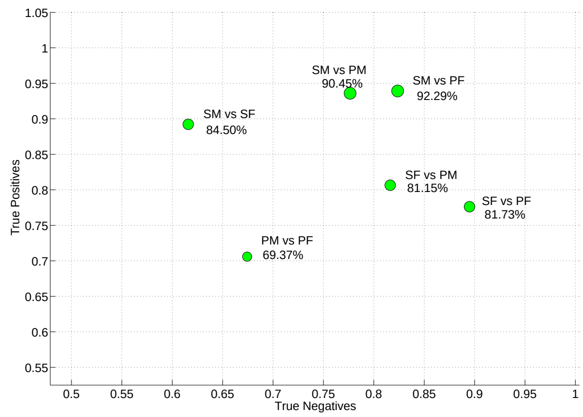


Fig. 4: *Spatial relationship between true positives, true negatives and accuracies for number and gender*

der (NG) is missing (i.e. past participle, which can also be regarded as adjective). Regarding the features, they were defined in lowercase text and extracted by a set of rules using word, POS tag, and base form information from both languages. Features take information from: a) bilingual model, i.e. current and previous tuples; b) target language, such as personal pronouns, presence of auxiliary verb 'haber' (for past participles), and verb form without 'tense' and 'mode' information; and c) source language, such as presence of English full verb form in the tuples, its POS tag, base form and related personal pronoun (if any, including reflexive pronouns, such as 'show them'), as well as presence of active or passive voice.

In the first case (PN), we trained the classifiers with 500k samples, while the second case (NG) the classifiers were trained with 300k samples. Finally, 10k samples were set apart for testing. The SVM^{Light} algorithm [11] is used for training[4] .

In order to keep a balance in the one-vs-one classifiers as discussed in section 2.3, we decided to join clases for 2nd person and 3rd person singular (2S and 3S, respectively), as well

[4] http://svmlight.joachims.org/

classifier	accuracy
1S vs 2S3S	97.94%
1S vs 1P	95.26%
1S vs 2P3P	91.71%
2S3S vs 1P	95.84%
2S3S vs 2P3P	88.38%
1P vs 2P3P	92.05%

Tab. 2: *Classification accuracies on Spanish verb person and number morphology information*

classifier	accuracy
SM vs SF	84.50%
SM vs PM	90.45%
SM vs PF	92.29%
SF vs PM	81.15%
SF vs PF	81.73%
PM vs PF	69.37%

Tab. 3: *Classification accuracies on Spanish verb number and gender morphology information*

as 2n person and 3rd person plural (2P and 3P, respectively). This is due to the low number of training samples for 2S and 2P was rather low. Finally, in both subcategories, 4 classes were defined (i.e. 6 binary classifiers). Accuracies for each classifier alone are shown in Tables 2 and 3.

In general, accuracies to classify person and number are higher than accuracies for number and gender. Classifier related to PM-vs-PF shows the lowest accuracy. Figures 3 and 4 reflect our results; the nearer a classifier is to the upper right part of the figures, the better its performance. These figures allow us to compare the performance of the different one-vs-one classifiers.

Table 4 presents the accuracies obtained in the multiclassification task. Results for PN are more satisfactory than those for NG, it is clear that classifier PM-vs-PF requires bigger improvement.

4. Conclusions

In this paper we presented a system for multiclass classification of simplified morphology for SMT. The DDAG structure provides good accuracy results to classify person and number of Spanish verbs; however classification accuracies need to be improved to better generate gender and number. Future work to improve the performance of the classifiers will be directed to optimize the set of features used by each classifier and to improve the representation of the features through coding alternatives to reduce the dimensionality of the vectors.

5. Acknowledgements

This work has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247762. This project was financed by the Ministry of Science and Innovation, ref. TEC2009-14094-C04-01. The first author is granted by a FI grant from the Catalan autonomous government.

morph. class.	DAG approach
Person & Number	85.70%
Number & Gender	72.31%

Tab. 4: *Classification accuracies of DDAG for PN and NG*

6. References

[1] S. Nießen and H. Ney, "Improving SMT quality with morpho-syntactic analysis," in *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbrücken, Germany, July 2000, pp. 1081–1085.

[2] S. Niessen and H. Ney, "Statistical machine translation with scarce resources using morpho-syntactic information," *Computational Linguistics*, vol. 30, no. 2, pp. 182–204, Jun 2004.

[3] S. Corston-Oliver and M. Gamon, "Normalizing German and English inflectional morphology to improve statistical word alignment," in *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, October 2004, pp. 48–57.

[4] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 868–876. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1091

[5] A. de Gispert and J. B. Mariño, "On the impact of morphology in English to Spanish statistical MT," *Speech Communication*, vol. 50, no. 11-12, pp. 1034–1046, 2008.

[6] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998, pp. 80–91. [Online]. Available: citeseer.nj.nec.com/schapire99improved.html

[7] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 76–85, 1990.

[8] J. Mariño, R. Banchs, J. Crego, A. de Gispert, P. Lambert, J. Fonollosa, and M. Costa-jussà, "N-gram-based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.

[9] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, 2004.

[10] J. C. Platt, N. Cristianini, and J. Shawe-taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems*. MIT Press, 2000, pp. 547–553.

[11] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, ch. 11, pp. 169–184.

Crowd-sourcing platform for large-scale speech data collection

João Freitas[1], *António Calado*[1], *Daniela Braga*[1], *Pedro Silva*[1], *Miguel Sales Dias*[1]

[1] Microsoft Language Development Center, Portugal
{i-joaof, i-antonc, i-dbraga, i-pedros, Miguel.Dias}@microsoft.com

Abstract

This paper presents an online platform based on crowd sourcing for speech data collection, named YourSpeech. This platform aims at collecting desktop speech data at negligible costs for any language, in order to provide larger training data for Automatic Speech Recognition (ASR) systems. YourSpeech provides means for users to donate their speech through a quiz game and a through a platform that allows the deployment of a personalized TTS (Text-to-Speech) system. We have already collected more than 25 hours of pure speech for European Portuguese (EP) and achieved a Word Error Rate (WER) of 1% over 10% of the collected corpus.

Index Terms: Speech data, crowd sourcing, speech donation, Text-to-Speech, Automatic Speech Recognition

1. Introduction

It is well known that Automatic Speech Recognition systems based on statistical methods require vast amounts of transcribed and annotated speech data in order to achieve acceptable accuracy rates. Acquiring a lot of speech data is particularly difficult when addressing less-resourced languages or even any other language that is not amongst the big five in terms of market economic relevance (English, Spanish, French, German and Italian). The main reason for this is that these corpora are expensive and recruiting speakers has proven to be quite costly [1] and hard to manage. Besides, some speech databases lack quality because of the bad recording conditions, sample rates inconsistency, erroneous, inconsistent or inexistent transcription, etc. [2].

This paper describes a solution to tackle this issue by using a crowd-sourcing approach. Crowd-sourcing is a term used to describe the leveraging of vast amounts of people to achieve a specific goal in a collaborative manner over the Internet. Many crowd-sourcing initiatives have been made possible due to the availability of Web 2.0 technologies, which enable massive collaboration projects to take place. Crowd-sourcing can be considered as a distributed process for the resolution of problems. Typically the process is as follows: an entity has a problem and needs to solve it in a cost effective way. The entity publishes the problem in the web and usually provides the tools to solve it. Users (the crowd) respond to the call and propose solutions to the problem. The publishing entity chooses the winning solution and rewards the user/users accordingly. Rewards vary from money incentives to just public recognition. The publishing entity will own the final winning solution. Multiple solutions can be found across the web in order to digitize old books [2], transcribe speech [4], classify tunes [5], classify galaxies from the Sloan sky survey [6], find ideas for proposed problems [7], image [8][9] and video [10] tagging and even build a summary of the entire Human knowledge [11].

The Yourspeech platform (http://pt.yourspeech.net) at MSN [12] aims at collecting speech at negligible costs for any language. The concept behind this system is to provide the user with an entertainment reward in exchange for his/her speech. This collection is based on crowd sourcing approaches [13] and invites the users to aid in the development of new ASR technology, while at the same time they are entertained by playing a quiz game (JustSayIt), or by obtaining audio files containing phrases pronounced by their own synthetic voices. European Portuguese is the language used in this prototype, but our goal is to scale it to other languages.

This paper is organized as follows: section 2 describes the system architecture, how the quiz game is built and how the personalized TTS voices are produced. In section 3, the media repercussion and users' experience and feedback is discussed. In section 4, the current results are depicted and in sections 5 and 6, future work and conclusions are presented respectively.

2. System description

The system architecture (Figure 1) is based on the client/server paradigm. The client application accesses the platform through a website and it is uniquely identified by the user's Windows Live ID. Once there, the user chooses to play the quiz game or create the user's own personalized synthetic voice. In order to access the client's operating system resources such as recording and playback devices, an ActiveX control is installed in the local machine. This control provides access to the Windows audio pipeline, thus enabling audio recording using any of the installed audio input devices.

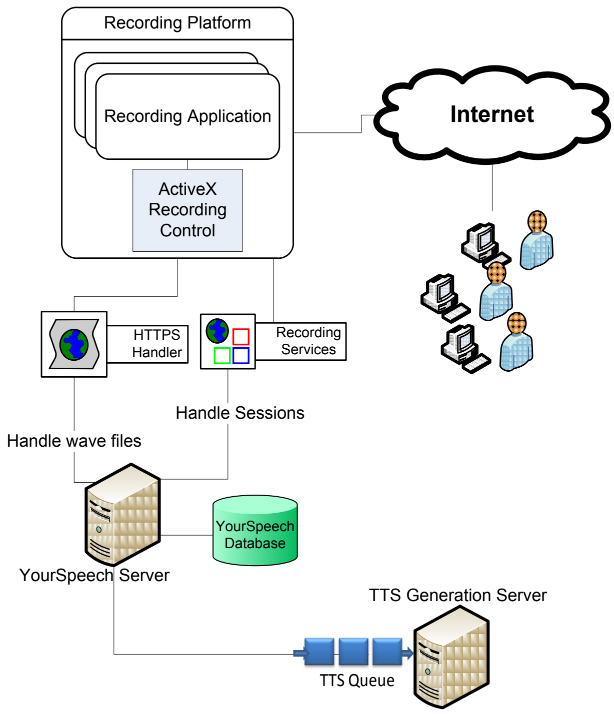


Figure 1: *High level system architecture diagram.*

After installing the ActiveX control, the client undergoes a setup phase in order to select the recording device and to guarantee the quality of the recorded audio. To perform this quality analysis, a Signal-to-Noise Ratio (SNR) value and server-side recognition accuracy rates' results are calculated before allowing the user to donate speech. After successfully passing the setup phase, the user is ready to donate speech.

2.1. Quiz game

When the user chooses the quiz game branch, a Rich Internet Application (RIA) based on Silverlight is loaded (Figure 2) and the game is presented to the user. The difference from this quiz to other quizzes found across the web [14] is that the questions are read by the default Text-to-speech voice installed in the client system and the answers are recognized by the correspondent speech recognition engine installed in the server. After the answer is spoken, the audio recording automatically stops and a wave file is sent to the server for recognition.

At the server side, a dynamic grammar with the answers is generated and fed to the engine. If the recognized answer matches the correct answer the user scores points accordingly to the answer's difficulty. Each quiz contains 18 thematic and generic questions split by easy, medium and hard difficulty, randomly extracted from a total pool of 160 questions. This pool of questions was entirely generated taking into account the phonetic richness of the answers as well as content variability, i.e. we tried to create questions from various common themes such as, sports, history, geography, literature, mathematics, physics, etc.

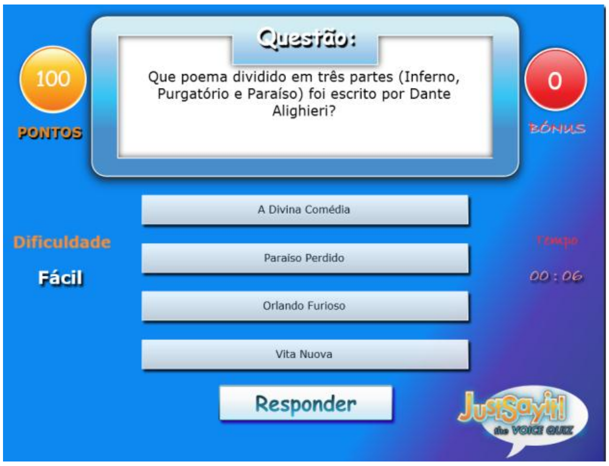


Figure 2: *Quiz game.*

2.2. Personalized TTS

When the user selects the Personalized TTS option, a different Silverlight based interface is presented (Figure 3). A minimum of 200 recorded sentences, taken from a phonetically rich set of prompts, are required in order to guarantee complete phone coverage and the success of the TTS generation process. When the required number of sentences is recorded, the user may choose to generate his/her personalized voice. Upon the generation of his/her personalized voice, the user is able to create and download audio files containing his/her synthesized voice uttering any input text of his/her choice. The quality of the voice rises as the number of recorded sentences increases.

The personalized TTS system relies on a secondary server to generate the personalized voices. This server receives requests to generate a user personalized voice over Microsoft Message Queuing (MSMQ). Tests indicate that the server is able to generate over 20 voices simultaneously, however the performance impact increases the generation time for more than 24 hours.

Figure 3: *Recording platform for the Personalized TTS option.*

Due to this fact, the server only generates 12 voices simultaneously, leaving other voice requests on hold. The process of generating a voice with 200 recorded prompts takes approximately 20 minutes. A synthetic voice generated with 200 prompts is intelligible, but may lack naturalness, because more prompts are required and its quality may be degraded, since the audio capture may be done using a common laptop and simple recording hardware and software.

The TTS system used to generate the user's voices is based on HMM's [15]. The front-end is dictionary-based, being composed by a lexicon with around 135000 words, phonetically annotated by a professional linguist with phonetic transcriptions, stress marks and syllable boundaries, and with Part-of-Speech (POS) information. The stress and syllable marking was automatically assigned using linguistic rule-based algorithms, specially developed for the European Portuguese language. The front-end is also composed by the text analysis module, which involves the sentence separator and word breaker components, including several other files, such as phone set and features and the POS tags set. It also includes a rule-based TN (Text Normalization) module [16], as well as a homograph ambiguity (also polyphony) resolution algorithm [16][18], a stochastic-based LTS (Letter-to-Sound) converter, used to predict phonetic transcriptions for out-of-vocabulary words and the prosody models, which are data driven using a prosody tagged corpus of 2000 sentences and a POS tagger who provides morpho-syntactic contextual information. The front-end outputs phonetic transcriptions that are subsequently input of the TTS runtime engine or back-end, which then outputs synthetic voice.

A voice font is built with the users' voice samples recorded through Yourspeech online platform. The recorded waves must be coincident with the scripts used in the EP HTS-based synthetic voice. The voice font creation involves wave processing, automatic phonetic annotation using ASR acoustic models, alignment with orthographic input and compiling. In order to be able to dynamically create voices, the whole pipeline process was automated. Figure 4 illustrates the system workflow.

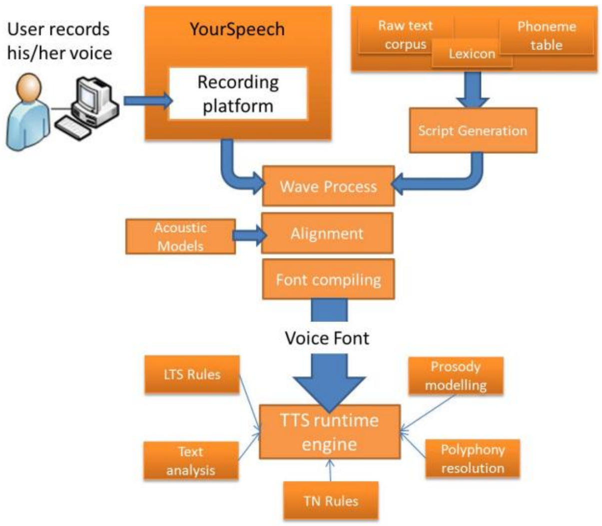


Figure 4: *Synthetic voice creation platform using YourSpeech.*

3. Media and User feedback

The success of this project is also dependent on the number of users that visit the site and actually record at least one prompt. For that reason, the user interface should attract the user's curiosity, but special emphasis should also be given to the dissemination and advertisement of the website. Figure 5 illustrates a clear correlation between media publication and number of visitors in the website. This graphic shows that YourSpeech started to be disseminated in some important European Portuguese blogs.

Later, it was integrated into MSN website and the highest peak of visits occurred when MSN announced and featured the Quiz Game in their entertainment area.

After that, YourSpeech appeared in several magazines publications [19][20][21], online TV stations [22] and it was referred in the television cable channel news TVI24 [23]. Site traffic statistics indicate 8682 visits and 1308 registered users.

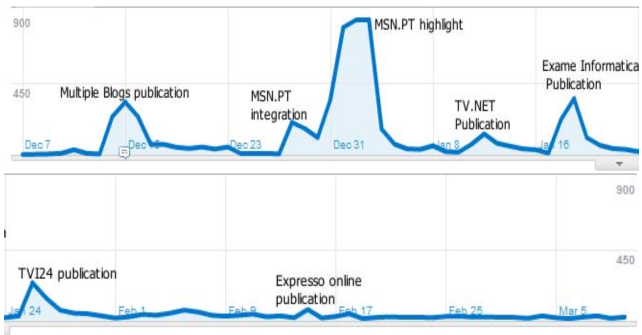


Figure 5: *Website visits from December to March.*

The website also allows users to comment on their experience. User feedback was extremely positive being reported that: "The game didn't have any issues in recognizing the answers, even when I misspelled some word"; "Good initiative! Keep up the good job and I'm looking forward for the final result"; "Very interesting game and a good cause for the Portuguese people"; "Excellent application. Good job".

4. Results

The YourSpeech platform went live at 14th of December 2009. Table 1 shows the speech/session results obtained until April 14th 2010.

Table 1. *Speech/Session results*

	Quiz Game	Personalized TTS	Total
Pure Speech (hours)	3.87	21.4	25.27
Total audio (hours)	11.9	48	59.9
Completed Sessions	473	94	567
Incomplete Sessions	205	223	428
Utterances	18300	9463	27763

Based on the results shown on Table 1, we can see that the Quiz Game is only responsible for 15% of the accumulated amount of pure speech, although it has 83% of the completed sessions. The reason for this to happen is due to the amount of time used when playing the Quiz Game (typically around 5 minutes per session) when compared with the time required to record the minimum 200 sentences for the personalized TTS voice (typically around 20 minutes per session).

From the 93 completed sessions, the total number of personalized TTS voices generated with success was 63.

A quality analysis was performed on the audio that has been collected during the first month, in order to check how well the speakers read what was proposed. The analysis was done with the help of a native Portuguese transcriber that listened and transcribed a sample of the collected corpus. The transcription process discards wrong utterances and corrects misspelled ones. The achieved results are described on Table 2.

Table 2. *Word Error Rate*

	Quiz Game	Personalized TTS	Total
WER	10.3%	0.05%	1%
Insertions	79	46	125
Deletions	92	103	195
Substitutions	36	47	83
Words	2010	40119	42129

Based the results from Table 2, we can see that the WER for the Quiz Game is much higher that for the Personalized TTS mostly due the number of insertions. Users tend to add articles or other words to the expected answer in order to complete their answer. In the personalized TTS system, the opposite happens. The user is told to read a sentence and we have seen that it is more common for users to omit or misspell words.

5. Future work

This platform has proven that crowd-sourcing can be a reliable and powerful tool to collect speech with a very small cost if the right motivation is provided and if a good marketing and advertisement structure is in place. Future steps include: transcribe and annotate all the collected corpora, retrain existent acoustic models by adding the collected data and verify any changes in the ASR accuracy rate.

We would also like to create content-specific games that are focused on certain groups of words (e.g. city names, numbers, etc.) in order to have acoustic models specialized in specific grammar types.

Improvements to the platform include: increase the number of questions available in the quiz, provide a better user experience by using Silverlight 4 and its built-in microphone recording features; have the platform available beyond the browser, this is, online and offline or migrate totally or partially the platform to Microsoft's Azure platform [24] in order to use cloud computing. It is also planned to use this concept in order to collect other languages.

6. Conclusions

YourSpeech demonstrates how crowd-sourcing can be used to expand speech resources. This platform is essentially divided into two applications: a quiz game with a speech interface and a recording platform that allows generating the user's own personalized synthetic voice.

The quiz game attracts more speakers (460 completed sessions) than the Personalized TTS application (90 completed sessions); however a session of the latter produces 24 times more pure speech than the other. The quiz game proved to be a lure for the personalized TTS system. At the time of writing of this paper, YourSpeech is online for 4 months and we have collected more than 25 hours of pure speech in European Portuguese.

After manually analyzing 10% of the collected corpus, we got 1% total WER, which shows that our data collection approach is promising and effective. According to the experience obtained from the European Portuguese campaign, YourSpeech is a viable platform for obtaining speech data at marginal costs given the fact that appropriate marketing and advertisement actions are taken.

This concept can also be expanded to a multi-lingual platform and applied using more sophisticated games.

7. Acknowledgements

The authors would like to thank to the Public Relations department from Microsoft Portugal for all the marketing support and to all YourSpeech users for donating their speech and feedback.

8. References

[1] Calado, A., Freitas, J., Braga, D., Dias, M., "Multi-Language Telephony Speech Data Collection and Annotation". in: Braga et al. (eds.) Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies, September 10, 2008, Curia, Portugal, 2008.

[2] Neto, N., Patrick, C., Adami, A.G., Klautau, A. G., "Spoltech and ogi-22 baseline systems for speech recognition in Brazilian portuguese", in Teixeira, A., Lima, V., Oliveira, L., Quaresma, P. (eds) Propor 2008, LNCS (LNAI), vol. 5190, pp. 256-259, Springer, Heidelberg, 2008.

[3] Re-captcha website, Online: http://recaptcha.net/ accessed on 20th April 2010.

[4] Google new account creation, Online: http://mail.google.com/mail/signup accessed on 20th April 2010.

[5] Tag a Tune, Online: http://www.gwap.com/gwap/gamesPreview/tagatune/ accessed on 20 April 2010.

[6] Galaxy zoo, Online: http://www.galaxyzoo.org/ accessed on 20th April 2010.

[7] Idea Bounty, Online: http://www.ideabounty.com/ accessed on 20th April 2010.

[8] Squigl, Online: http://www.gwap.com/gwap/gamesPreview/squigl/ accessed on 20th April 2010.

[9] Google Image labeler, Online: http://images.google.com/imagelabeler/ accessed on 20th April 2010.

[10] PopVideo, Online: http://www.gwap.com/gwap/gamesPreview/popvideo/ accessed on 20th April 2010.

[11] Wikipedia.org, Online: http://en.wikipedia.org/wiki/Wikipedia accessed on 20th April 2010.

[12] MSN Portugal, Online: http://pt.msn.com/ accessed on 20th April 2010.

[13] Brabham, D. C., "Crowdsourcing as a Model for Problem Solving: An Introduction and Cases", in: Convergence The International Journal of Research into New Media Technologies, Vol. 14(1), pp. 75-90, 2008.

[14] Funtrivia website, Online: http://www.funtrivia.com/, accessed on 20th April 2010

[15] Braga, D., Silva, P., Ribeiro, M., Henriques, M. and Dias, M. "HMM-based Brazilian Portuguese TTS", Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies, September 10, 2008, Curia, Portugal,2008.

[16] Ribeiro, M., Braga, D., Henriques, M., Dias, M., Rahmel, H., "Resolução de Ambiguidades na Normalização de Texto em Português Europeu", in Actas do XXIV Encontro Nacional da Associação Portuguesa de Linguística. Textos Seleccionados, Braga, Portugal, 2009.

[17] Braga, D., "Algoritmos de Processamento da Linguagem Natural para Sistemas de Conversão Texto-Fala em Português" PhD Thesis. A Coruña University, Spain, 2008.

[18] Braga, D.; Coelho, L.; Resende Jr., F. G. V., "Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems", in Proceedings of Interspeech 2007. Antwerpen, Belgium. pp.1761-1764, 2007.

[19] Interview for Exame Informática (top business magazine in Portugal) magazine. Online: http://aeiou.exameinformatica.pt/a-microsoft-quer-por-os-portugueses-a-falar-para-o-computador-video=f1004730 accessed 20th April 2010.

[20] Exame Informática article. Online:http://www.pt.cision.com/O4KPTWeb/ClientUser/GetClippingDetails.aspx?id=8b7e0270-395c-4a9c-a8fc-4429c32d9b72&analises=1 accessed 20th April 2010.

[21] Expresso Online article. Online: http://aeiou.expresso.pt/gen.pl?p=stories&op=view&fokey=ex.stories/563758 accessed 20th April 2010.

[22] Interview for TV.net (on-line TV station in Portuguese). Online: http://tvnet.sapo.pt/noticias/video_detalhes.php?id=53031 accessed 20th April 2010.

[23] TVI24 interview. Online: http://www.agenciafinanceira.iol.pt/consola.html?id=1132691 accessed 20th April 2010.

[24] Microsoft Azure: http://www.microsoft.com/windowsazure/ accessed 20th April 2010

A Prototype of Distributed Speech Technologies for the Development of Websites Accessible to the Blind Community

Oscar Saz, Eduardo Lleida, José-Enrique García, Alfonso Ortega

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Spain
{oskarsaz,lleida,jegarlai,ortega}@unizar.es

Abstract

This paper presents a proposal for the use of distributed speech technologies like Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) synthesis to create a paradigm of web development which integrates voice as input and output interface in the navigation experience. This enhanced usability is aimed to provide accessibility to the Internet for the community of blind and visually impaired people, whose access nowadays is difficult and costly. The novel concept in this work is to provide the web developer all the control on the use of the speech synthesis and recognition, i.e. designing the voice experience within the website through the use of a simple set of tags, rules and functions. A pilot experience with Aragón Radio 2, the online radio channel of the Aragonese Corporation of Radio and Television (CARTV), has shown the arising possibilities for this technology and opens the interest for further use of them in collaboration with different institutions.

Index Terms: visual impairments, technical aids, web accessibility, speech synthesis, speech recognition

1. Introduction

Providing universal accessibility to the contents of the Internet is a major issue for all the agents involved in the evolution of the World Wide Web (WWW); the World Wide Web Consortium (W3C) is currently working through the Web Accessibility Initiative (WAI) on defining different scenarios of inaccessibility to web sites and trying to create standards of web development to eliminate barriers for the handicapped community [1]. Due to the visual nature of the web, blind and visually impaired people are the community which finds the largest barrier in their access to this new gate to knowledge.

Screen readers have been the usual tools for blind people to access computers and, furthermore, the Internet. Some of the most popular systems, like JAWS Screen Reader [2], GW Windows-Eyes [3] or the late IBM Homepage Reader, have been running up for many years, achieving great success and providing new versions and languages along time. These systems take all the textual information presented in the screen and provide synthesized speech for it. The blind person can use keyboard commands to navigate through the different elements in the computer screen and to configure the system. Nowadays, more novel solutions are appearing from the academic world trying some of the disadvantages of these commercial software, especially the high price and the need of installing them in a specific platform (mainly Microsoft Windows). For instance, WebAnywhere [4] is a novel solution which provides speech synthesis on the Internet; when a URL address is provided, a Macromedia Flash application connects to a remote server which synthesizes the sequence of texts within the website.

However, the experience of navigating the web with these systems is not as easy as it may seem. Professionals developing websites hardly think on how their site will be interpreted by these speech-synthesis-based applications. This means that the order in which the different texts in a website are synthesized may provide a different feeling than the one when visually inspecting the site. Furthermore, elements which are very common in the development of websites like side bars, search indexes, etc. produce a lot of information when synthesizing the page, which makes difficult for the blind person to separate the important information in the site from all the secondary elements within.

This work proposes to give web developers the tools to not only decide the visual appearance of their websites, but also to define the parameters of a voice input-voice output interaction with the webpage. The inclusion of an active element, i.e. Java Applet, allows the use of distributed services for Text-to-Speech (TTS) synthesis and Automatic Speech Recognition (ASR) as part of the interaction; these elements being handled by the developer via Javascript functions. With the use of a set of tags inserted in the Hyper-Text Markup Language (HTML) document of the website, it is possible to define which elements will be synthesized, to create dependencies among them, to use alternative texts to synthesize and to define a simple oral control of the most usual commands. The prototype for a pilot study will show the possibilities of this new proposal and will promote the development of websites based on this technology.

The paper is organized as follows: Section 2 will describe the use of distributed speech technologies and the active element which permits to include them in the development of a website. Section 3 will describe the HTML tagging system created for the design of the audio-accessible web, as well as the set of keyboard commands which the blind person can use to navigate. Posteriorly, the pilot study with the website of Aragón Radio 2 will be described in Section 4. Finally, Section 5 will present the future lines of work and conclusions to this work.

2. Distributed Speech Technologies

The use of distributed architectures is becoming more and more usual nowadays to take advantage of technologies with a high computational cost on small portable devices which incorporate a fast and reliable network connection. It also helps providing software services on devices and platforms where it is not possible or desirable to install a stand-alone application.

In the specific case of speech technologies like ASR and

TTS, distributed architectures have allowed the massive introduction of these technologies in devices like PDAs or smartphones. In these cases, the client (small device) only captures the audio from the user and plays the synthesized audio signals while the server carries on all the signal and language processing required, including Large Vocabulary Continuous Speech Recognition (LVCSR) or Hidden Markov Models (HMM)-based TTS. Commercial services of the so called Voice Search are operative nowadays achieving a certain success [5].

The EDECÁN Consortium (TIN-2005-08660-C04-01) gave a proposal for the use of speech technologies in distributed architectures [6]. This proposal had a central communications manager which interconnected different services in different remote machines. The services included ASR, TTS, speaker adaptation, speaker verification, dialog managers and any other service which was adapted to the EDECÁN protocol. With this architecture, any client requiring specific speech technologies systems (for instance, to create a full dialog system) would connect to the communications manager and the manager would connect with the services required by the client. This architecture served as the basis of the proposed system for distributed speech technologies on the web.

2.1. Speech Technologies on the Web

The use of advanced technological resources on the web is another case where distributed architectures are shown to be extremely useful. In these cases, using applications directly installed on the user's computer might be uncomfortable for the user and complicated for the developer in order to provide updated versions for all the possible platforms (Microsoft Windows, MAC OS, Linux) and web browsers (Microsoft Internet Explorer or Netscape-based browsers). Hence, the same network connection used to access the remote website can be used to access a resource. Systems like XHTML+Voice (X+V) have tried to provide standards for creating spoken dialogs in the web using the potentially of VoiceXML [7],but the need to include the X+V capabilities on the client side and VoiceXML tagging on the website by the developer have limited its possibilities to become a universal standard.

For the remote use of speech technologies on the web, a schematic diagram like in Figure 1 is proposed. The user accesses to the website via the WWW protocol and the page connects in a client-server protocol with the remote server where all the main processing is done. The page does not require any specific coding aside from the standard HTML, PHP or Javascript languages. The use of speech synthesis and recognition in the web is performed with the inclusion of a Voice-input Voice-output (ViVo) Java Applet in the web source code. The utilities that are provided by this Applet are ASR and TTS.

The ASR is performed with the vocabulary indicated by the website, either in a separate file or as a String. The visual interface of the Applet permits a push-to-talk interface where the user presses a button in the Applet to launch the recognition. Otherwise, the developer can start the recognition when a certain function of the code is called by an event in the page. An audio reinforcement of the start of the recognition phase in the form of a 'beep' can be used to help the user. The result of the ASR is collected by an specific function in Javascript.

The TTS is able to synthesize any utterance in Spanish represented by a String. It provides different voices in Castilian-Spanish and Aragonese languages which can be selected at the moment of invoking the TTS procedure.

A set of Javascript functions are provided to make use of

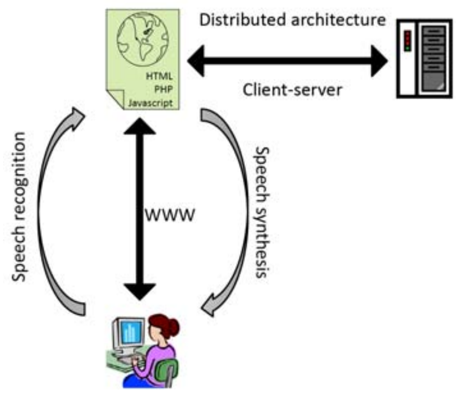


Figure 1: *Scheme of distributed voice interaction on the web.*

these functionalities:

- UZSinte(String sentence, String speaker): Launches a TTS instance of a "sentence" using a given "speaker".
- UZSinteStop(): Finishes reproduction of the TTS.
- UZStartRecoGrammar(String url-grammar): Launches an ASR instance using the grammar in a URL.
- UZStartRecoGrammarText(String grammar): Launches an ASR instance using the provided grammar.
- UZStopReco(): Stops the ASR procedures.

The Applet connects with the remote server indicated as a parameter in the HTML code, which performs all the ASR and TTS processing. When ASR is required, the Applet captures the audio from the standard audio input and performs the extraction of the Mel Frequency Cepstral Coefficients (MFCC) features from the speech signal. These MFCC features are quantized and the Applet sends them to the server with with a reduced bit rate without losing ASR performance [8]. When the server decodes the uttered sentence, it sends back the results to the Applet as a string, which is post-processed by a recoend() function which performs the actions programmed by the developer.

When TTS is called, a string is sent to the server with the sentence to synthesize and the desired voice to use. The TTS process is performed on the remote server and, when finished, the audio is sent as a stream to the Applet which plays it through the standard audio output.

3. Proposal for Audio-Accessible Web Development

The functionality for providing the accessibility on the website with the use of the ViVo Java Applet is based on two elements which define the interaction: The HTML tags defined the elements of the site to be synthesized and that can be placed freely by the web developer; and the set of keyboard commands which serve the blind user to navigate thought those different elements and to activate and deactivate different functionalities.

3.1. HTML Tags for Voice Web Interaction

The way in which a web developer provides TTS of a certain section of the website is made by adding attributes to certain HTML tags. Three HTML elements can have synthesized speech associated to them: , <p> (paragraph) and <a> (link). Table 1 shows the templates to apply this feature in these 3 elements with different possibilities. An element with synthesized speech is marked with the attribute $class = \text{“}headings - sinte''$ and the text to be synthesized is included in the attribute $title = \text{“}Text\ to\ synthesize''$. Links and paragraphs allow for the direct synthesis of the text in the given paragraph or link. The use of spans permits to summarize a large section of the section in a single synthesized text.

Table 1: *HTML tags for voice web interaction.*

span	 HTML code
p	<p class="headings-sinte"> Text to appear on screen and synthesize </p>
a	 Text in link
a	 Text in link and synthesize

One of the more important features that is allowed by the system is to define two levels of relevance within the different elements to be synthesized. This way, a major heading can contain subheadings which will only be synthesized if the user decides so with a certain key combination. The HTML tagging for this feature is as shown in Table 2 where a major <span> element contains a paragraph and a link which are marked with the attribute $class = \text{“}subheadings - sinte''$.

Table 2: *Organizational tree of HTML tags.*

```
<span class="headings-sinte" title="Top level text">
    HTML code
    <p class="subheadings-sinte">
        Second level text
    </p>
    HTML code
    <a class="subheadings-sinte">
        Second level text
    </a>
    HTML code
</span>
```

When the TTS voice reads the main element, the system indicates orally that there are a certain number of elements and asks whether the user wants to read them or not. At any moment, the user can come back to the initial list of main elements with a keyboard command.

The web developer also has the chance to provide a set of oral commands which create oral shortcuts to the most usual actions on the site. The HTML syntax to define these commands is as seen on Table 3, which is interpreted by the system and extracts the "Command utterances", that is the sentence to be spoken by the user to be recognized, and the command to be performed as the function toDoOnRecognition(){}.

Table 3: *HTML tags for speech commands.*

```
<span class="commands-reco" title="Command utterance"
onclick="function toDoOnRecognition(){}">
    HTML code
</span>
```

The system is able to process all these tags in the body of the HTML code and create a grammar with all of them to be used in the recognition stage. When a valid sentence is pronounced, the function or code to be performed is run automatically.

3.2. Use of keyboard commands

Keyboard commands are the main basis that blind people use when accessing computer systems and the Internet. Different combinations of keys allows to move through the different elements in the screen and to configure and use specific software.

The basis that permits the use of the ViVo Java Applet in the accessibility of web sites by blind people are a series of Javascript functions which allow the control of the Applet with keyboard commands and which have already been developed and are ready to be provided to web developers.

Table 4 presents all the available commands for the website control. They are divided into control commands which activate or deactivate the voice aids, synthesis and recognition, or provide an auditive help about the handling of the system; and into commands which serve for navigating around the website, moving back and forth through the different items which are defined to be synthesized.

Table 4: *Keyboard commands for voice web interaction.*

TTS handling commands	
Ctrl+Q	Reads the synthesis help
Ctrl+A	Activates the synthesis of the site
Ctrl+D	De-activates the synthesis of the site
Ctrl+(Right arrow)	Reads next element in list
Tab	Reads next element in list
Ctrl+(Left arrow)	Reads previous element in list
Shift+Tab	Reads previous element in list
Ctrl+(Up arrow)	Reads first element in list
Ctrl+Home	Reads first element in list
Ctrl+(Down arrow)	Reads current element in list
Ctrl+End	Reads last element in list
Ctrl+Z	Switches levels of in the elements tree
ASR handling commands	
Ctrl+E	Reads the recognition help
Ctrl+R	Starts recognition of the user's speech

4. Pilot Study with Aragón Radio 2

A prototype for an audio-accessible website was created in collaboration with the Aragonese Corporation of Radio and Television (CARTV). CARTV has recently launched an on-line radio channel on the Internet (http://www.aragonradio2.com) with the aim of creating a more participative and accessible way of communication with its listeners. The intention of the prototype was to demonstrate the possibilities that the novel techniques explained in this work had for providing Internet access to the blind community [1].

The prototype was build on a static version of the website retrieved on January 2010. The visual appearance of the prototype site is the same to the original one, as seen in Figure 2. The 4 main sites of the AR2 website were adapted to the system (Home, News, Public Service and Podcasts) as well as the radio and podcast players (which work as pop up windows).

The highlighted elements in Figure 2 show the initial part of the organizational tree of the website for synthesis purposes.

[1] http://dihana.cps.unizar.es/ alborada/ar2/ar2_frames.htm

Figure 2: *Appareance of the pilot study in the Aragón Radio 2 prototype website (synthesis elements are highlighted).*

Elements in dotted line are those which are at the top level of the organizational tree and which are read sequentially until the user decides to enter into some of the subheadings of a main element with the Ctrl+Z command (continuous line in Figure 2). Table 5 provides the contents available for synthesis in the homepage of the prototype and the sub-contents within them.

Table 5: *Information structure in the Aragón Radio 2 prototype.*

Main contents	Sub contents
Main menus	Home page News page Public services page Podcasts page
Latest news	Text of first new Text of second new Text of third new, etc.
Radio launcher	Aragón Radio Aragón Radio 2
Latest podcasts	Description of first podcast Description of second podcast Description of third podcast, etc.
Daily programs	Name of the first program Name of the second program Name of the third program, etc.

With the proposed systems, the understanding of the key elements in the website is better and long lists (programs, podcasts, latest news, etc.) are kept in a secondary level that the user can read when really interested in doing so. Furthermore, elements with little relevance (banners, self-advertisements) are omitted to avoid the user losing focus on the relevant elements.

For providing enhanced interaction, the most common actions in the website were also allowed via ASR. The commands that were incorporated included:

- Navigate to the home page
- Navigate to news page
- Navigate to the public service page
- Navigate to the podcasts page
- Listen to Aragón Radio
- Listen to Aragón Radio 2
- Play radio (when radio pop-up is on)
- Stop radio (when radio pop-up is on)
- Switch radio channel (when radio pop-up is on)

A preliminary evaluation by the experts in technical aids of the Spanish National Association for the Blind (ONCE) gave a positive review on the perspectives of this new technology. Some of their proposals regarding the functionality and use of the system were incorporated in the described system and future evaluation with a set of users will be made to assess how helpful this technology can actually be.

5. Conclusions and Further Work

The conclusion of this work focuses on the great possibilities that the use of speech technologies may have for providing accessibility on the Internet for blind people. The preliminary evaluation of the pilot study carried out showed a great potential of the technology and encouraged to keep working in this line. Among the main features of the proposed system are: Cross-platform use of speech synthesis and recognition, low bandwidth required due to the use of efficient coding features of speech features, absolute power for the web developer to design the audio appearance of the website and full accessibility via simple keyboard commands.

Aragón Radio 2 plans to fully incorporate this system in its website in the near future as part of their effort on providing universal accessibility to their resources. The auditive aid will be started via a keyboard command which can be easily used by a blind user. That way, people without visual impairments will maintain the same experience navigating the site, while their blind peers will have available the new web experience.

6. Acknowledgements

This work was supported under TIN-2008-06856-C05-04 from MEC of the Spanish government.

7. References

[1] W3C. (2010, May) Introduction to "How People with Disabilities Use the Web". [Online]. Available: http://www.w3.org/WAI/intro/people-use-web

[2] (2006) Jaws 8.0 for Windows, Freedom Scientific. [Online]. Available: http://www.freedomscientific.com

[3] (2006) GW Micro Windows Eyes. [Online]. Available: http://www.gwmicro.com/Window-Eyes/

[4] J.-P. Bigham, C.-M. Prince, and R.-E. Ladner, "WebAnywhere: A screen reader on-the-go," in *Proceedings of the Cross-Disciplinary Conference on Web Accessibility*, Beijing, China, 2008.

[5] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero, "An introduction to voice search," *IEEE Signal Processing Magazine*, pp. 29–38, May 2008.

[6] J.-E. García, A. Ortega, A. Miguel, and E. Lleida, "Arquitectura distrubuida para el desarrollo de sistemas de diálogo hablado: EDECÁN," in *Proceedings of the VJTH*, Bilbao, Spain, November 2008, pp. 171–174.

[7] R. Beasley, J. O'Reilly, K.-M. Farley, and L.-H. Squire, *Voice Application Development with VoiceXML*. Indianapolis, IN, USA: SAMS, 2001.

[8] J.-E. García, A. Ortega, A. Miguel, and E. Lleida, "Differential vector quantization of feature vectors for distributed speech recognition," in *Proceedings of Eurospeech*, Brighton, UK, September 2009, pp. 2587–2590.

Poster Session 3

A Discriminative Text Categorization Technique for Language Identification built into a PPRLM System

M. A. Caraballo, L. F. D'Haro, R. Cordoba, R. San-Segundo, J.M. Pardo

Speech Technology Group. Dept. of Electronic Engineering. Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040-Madrid, Spain
{macaraballo,lfdharo,cordoba,lapiz,pardo}@die.upm.es

Abstract

In this paper we describe a state-of-the-art language identification system based on a parallel phone recognizer, the same as in PPRLM, but instead of using as phonotactic constraints traditional n-gram language models we use a new language model which is created using a ranking with the most frequent and discriminative n-grams between languages. Then, the distance between the ranking for the input sentence and the ranking for each language is computed, based on the difference in relative positions for each n-gram. The advantage of the proposed ranking is that it is able to model reliably longer span information than in traditional language models and that with less training data it is able to obtain more reliable estimations. In the paper, we describe the modifications that we have made to the original ranking technique, i.e., different discriminative formulas to establish the ranking, variations of the template size and a penalty for out-of-rank n-grams. Results are presented on a new and larger database. The test database has been significantly increased using cross-fold validation for more reliable results.

Index Te rms: Language Identification, n-gram frequency ranking, text categorization, PPRLM

1. Introduction

Currently, one of the most used technique in Language identification (LID) is the phone-based approach, like Parallel phone recognition followed by language modeling (PPRLM)[1]. In PPRLM, the language is classified based on statistical characteristics extracted from the sequence of recognized allophones.

In spite of the high LID accuracy results obtained by PPRLM, the accuracy is reduced due to the presence of bias in the scores generated by each recognizer and because PPRLM does not model correctly long-span dependencies (i.e. to use high order n-gram language models) probably due to an unreliable estimation of the n-gram probabilities. In order to solve the first problem, we decided to use a GMM classifier and a normalization procedure called differential scores. Regarding the second problem, we decided to use a ranking of occurrences of each n-gram with higher n-grams, in a similar way to [2] and [3] where the ranking is applied to written text. Although the information source is very similar to PPRLM (frequency of occurrence of n-grams), results are much better, as we will see.

This paper is a continuation of the work done in [4] and [5] but tested on a new database with more languages and including new modifications to the ranking algorithm. Section 2 describes the system setup and basic techniques. In Section 3 the basic n-gram ranking technique and the new discriminative n-gram ranking are described, together with the results considering all the new alternatives considered. Finally, conclusions and future works are presented in Section 4.

2. System description

2.1. Database

For this work we have used the C-ORAL-ROM database [6], which consists of spontaneous speech for 4 main Romance Languages: Spanish, French, Portuguese, and Italian. This database is made of 772 spoken texts with more than 120 hours of speech and around 300K words for each language. The database transcriptions and annotations were validated by both external and internal reviewers. The database includes recordings in two different types: formal and informal (equally distributed). The formal recordings consist of three different contexts: natural (e.g. political speech, teaching, preaching, etc.), media (e.g. talk shows, news, scientific press, etc), and telephone (e.g. private and human-machine). The informal recordings include monologues, dialogues, and conversations in familiar and public contexts.

Next, we describe the main changes that we made to the database in order to adapt it to our experiments and recognition system: a) Most of the sound files were sampled to 22,050 Hz @ 16 bits and some others to 11 KHz @ 16 bits, all of them were sub-sampled to 8 KHz @ 16 bits in order to use them with the acoustic models of our recognizer. b) Some recordings in the database were too long (i.e. longer than 10 minutes) so they were splitted into shorter files. This way, we also eliminated noised and difficult to recognize sections, c) Finally, we generated random recording lists in order to avoid any kind of bias at training. Table 1 shows the number of sentences in the database that we have finally used. The average sentence length is 6.2 seconds.

	Spanish	French	Italian	Portuguese
Sentences	17634	16474	19074	17946

Table 1: *Number of sentences by language*

2.2. General conditions of the experiments

The system uses a front-end with PLP coefficients derived from a mel-scale filter bank (MF-PLP), with 13 coefficients including c0 and their first and second-order differentials, giving a total of 3 streams and 39 parameters per frame. We have used two phoneme recognizers, for Spanish and English, with context-independent continuous HMM models. For Spanish, we have considered 49 allophones and, for English, 61 allophones, all with 3 states. All models use 10 Gaussians densities per state per stream.

The performance of phoneme recognizers is very low for several reasons: a) there is a mismatch between the recognizers' languages and the 4 languages to be identified; b) the recordings still contain different kind of noises, background music, etc., and very spontaneous speech; c) the acoustic models were not adapted to this database. So, there is a clear mismatch in the languages and in the channel conditions. The good thing of using this setup is that improvements obtained with our techniques will be more evident, as we will see.

In order to increase the reliability of the results presented in the next sections, we performed a cross-fold validation, dividing all the available material in 9 subsets: 5 subsets to estimate the LMs, 2 subsets to estimate the Gaussian classifier, 1 subset for development, and 1 subset for test.

2.3. Description of PPRLM

Nowadays, PPRLM is the most popular approach to language identification. The main objective of PPRLM is to model the frequency of occurrence of different allophone sequences in each language. The technique can be divided into two stages. First, several parallel phone recognizers take the speech utterance and outputs a sequence of allophones corresponding to the phone sets used for each one. Second, the sequence of allophones is used as input to a bank of n-gram language models (LM) in order to capture phonotactics information. In this stage, the language model scores the probability that the sequence of allophones corresponds to a given language.

The main advantages of PPRLM are: a) Since it uses many recognizers, it is possible to cover most of the phonetic realizations of every language. b) It is possible to have phone recognizers modeled for languages different to the languages that we have to identify, which is especially useful in situations when the training data is not enough to obtain reliable language dependant models. On the other hand, PPRLM presents two major weaknesses: a) The presence of bias in the log-likelihood scores generated by each combination of the *N* recognizers and *M* language models and, b) the data sparcity and limitations of the n-grams LMs to model long span information.

The bias problem is mainly due to the differences between the allophone dictionaries and training data used by each recognizer [1]. In [7] two solutions for this problem are described. The first solution is called bias removal; it consists on a normalization procedure using as LM score the calculated score minus the average score in the training data. Then, the language is identified using a Maximum Likelihood Classifier. The second solution is to use another kind of classifier, such as Gaussian, K nearest-neighbor, or Support Vector Machine (SVM) classifiers. The advantage of using these classifiers is that the classification is not based on using an absolute discriminant function, and therefore it is not affected by the bias. In our system, given the good results obtained in [8], we decided to continue using a Gaussian Classifier. These classifiers also benefit from applying normalization of the scores (e.g., the T-norm normalization). In our system, we use what we call "differential scores", which is a similar normalization.

Regarding the problems with the LMs, the data sparcity is difficult to solve because it would require new training data (i.e., obtaining new recordings or using an external corpora with the same dictionary of phonemes used in our platform) consisting of a sequence of recognized phonemes. Regarding solutions for the problem of including long span information to the language models, in [9] they describe slight improvements on the LID rate when using the skip-gram technique, and in [3] they present LID experiments on written text for six languages using three different kinds of LM: Markov models, trigram frequency vectors, and n-gram text categorization, with good results for the last technique. In our case, we have used and extended the n-gram text categorization technique [2].

2.4. Gaussian classifier for LID

As mentioned above, the general PPRLM approach has a bias problem in the log-likelihood score for the languages considered. To tackle this issue, we proposed in [10] to use a Gaussian classifier instead of the usual decision formula applied in PPRLM. With all the scores provided by every LM in the PPRLM module we prepare a score vector. With all the sentences in the training database we estimate a Gaussian distribution each language. In recognition, the distance between the input vector of LM scores and the Gaussian distributions for every language is computed, using a diagonal covariance matrix, and the distribution which is closer to the input vector is the one selected as identified language. Besides, the Gaussian classifier allows us to increase the number of Gaussians to better model the distribution that represents our classes.

One important conclusion of our work in [10] is that, instead of absolute values, we need to use differential scores: the difference between the score obtained by one LM and the average score obtained by the other 'competing' languages: ($SC_i' = SC_i - Aver(SC_j, j \neq i)$) in Figure 1. We applied it to unigram, bigram and trigram separately, with 8 scores x 3 n-grams = 24 features in total in the feature vector.

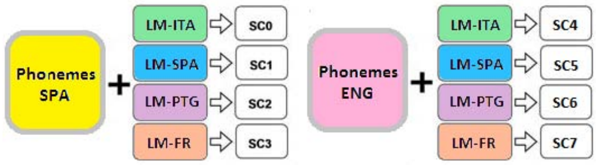


Figure 1: *PPRLM scores used for the LID system*

The average result in LID for PPRLM is **35.8%** error rate. It is a bad result, but, as we mentioned in Sections 2.1 and 2.2, the performance of the acoustic models is really poor and the sentences average length is short.

3. N-Gram Frequency Ranking

In this section we will describe the original text categorization technique and the modifications that we have made to improve it, as well as the selection of the most discriminative n-grams.

3.1. Description of the Basic Technique

In [2], an interesting technique that combines local information (n-grams) and long-span information (collected counts from the whole utterance) is described. In summary, for training the original technique proposes the creation of a ranked template with the N (typically 400) most frequents n-grams (up to n-grams of order five) of the character sequences in the train corpus for each language sorted by occurrence and then orthographically in case two or more n-grams contain the same occurrence (e.g., positions 10 and 11 in Figure 2).

During the evaluation, a dynamic ranked template is created for the phoneme sequence of the recognized sentence following the same procedure. Then a distance measure (OOP, Out-Of-Place) is applied between the input sentence template and each language dependent template previously trained. The distance for a given ranking T is calculated using Eq. 1.

$$d^T = \frac{1}{L}\sum_{i=1}^{L} abs(pos\ w_i - pos\ w_i^T) \qquad \text{Eq. 1}$$

Where *L* is the number of n-grams generated for the input sentence. If an n-gram does not appear in the global ranking (meaning that it has not appeared in training or it is not in the top n-grams selected) it is assigned a maximum distance: the size of the ranking. The selected language is the one that presents the higher correlation between templates (i.e., the lower distance).

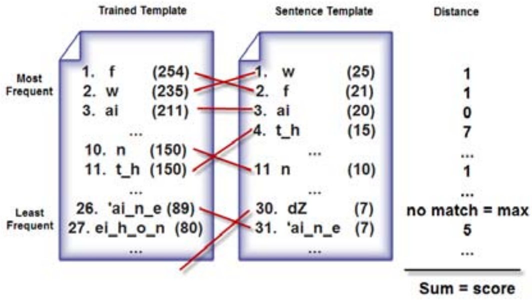


Figure 2: *Example and calculation of distance score using a ranking of n-grams as proposed by [2]*

Figure 2 shows an example of one of the templates created in our system using the English phoneme set and the template created for the unknown sentence.

3.2. Our baseline for N-Gram Ranking

In [5] we described several modifications that we made on the basic technique proposed in [2]. Below we provide a brief description of the most important ones.

Our first variation is what we called the "golf score". As the number of occurrences of the n-grams in the input sentence is very low, most n-grams have the same number of occurrences and should have the same position in the ranking. It is the same as a ranking in golf (the sport): all players with the same number of strokes share the same position. Figure 3 shows an example of the modification applied to the original template using the proposed "golf" score. Using this technique we obtained a 2.5% relative improvement.

For the second modification we thought that having only one global ranking was not efficient since, in general, the top positions were always devoted to unigrams & bigrams, which we already knew that were less discriminative for LID. So, we decided to have different rankings for each n-gram order (besides that, the procedure is the same). As the ranking size for unigram and bigram will be different between languages, we need an additional normalization in the distance measure, i.e., we divide it by the number of items in the set for that n-gram order. We also increased the template size to an optimum of 3000, which is the baseline for this paper.

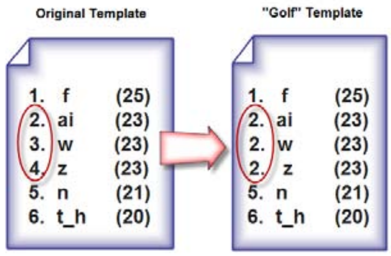


Figure 3*: Ranking template modification with "golf score"*

3.3. N-Gram Discriminative Ranking

Inspired in the work of [11], where better LID results could be obtained using the most discriminative units, we thought that we should introduce the same concept in the ranking creation process; therefore, we decided to give more relevance (higher positions) in the ranking to the items that are actually more specific to the language that is being identified, i.e. n-grams with a high frequency in one language but with zero or low frequency in the competing languages.

In our work we propose a variation of tf-idf. After the original global rankings are created, we have the number of occurrences of each n-gram: $n_1(w)$ = occurrences of n-gram w in the current language, and $n_2(w)$ = the average occurrences of w in the competing languages, where T are the ranking templates created for each language.

$$N_1 = \sum_{\forall w: w \in T_1} n_1(w) \qquad N_2 = \frac{1}{|T-1|} \sum_{\forall w: w \in T; T \neq T_1} n_2(w) \qquad \text{Eq. 2}$$

As the number of total occurrences will be different for each language and n-gram order, before the subtraction a normalization is needed to have comparable amounts. Being N_1 the sum of all occurrences for the current language and N_2 the average for the competing languages (see Eq. 2):

$$n_1'(w) = \frac{n_1(w) \times N_2}{N_1 + N_2} \qquad n_2'(w) = \frac{n_2(w) \times N_1}{N_1 + N_2} \qquad \text{Eq. 3}$$

Another important issue is to use a threshold value for these normalized values (Eq. 3), i.e., to discard the n-grams that were below a threshold as non-representative. In our case, we obtained an optimum using 9-9-3-3-2 (threshold values for each n-gram starting at unigram). Then, we considered several alternative formulas with the same philosophy as tf-idf for the final number of occurrences used to assign the final position in the ranking (which we will call n_1'').

1	$n_1'' = (n_1' - n_2') / (n_1' + n_2')$
2	$n_1'' = n_1' * (n_1' - n_2') / (n_1' + n_2')$
3	$n_1'' = \log(n_1') * (n_1' - n_2') / (n_1' + n_2')$
4	$n_1'' = \mathrm{sqrt}(n_1') * (n_1' - n_2') / (n_1' + n_2')$
5	$n_1'' = n_1' * (n_1' - n_2') / (n_1' + n_2')^2$
6	$n_1'' = \mathrm{abs}\,(n_2' - n_1) / \text{sum all lang}(n_1)$

In Table 2 we can see the results in LID error rates for the 6 different formulas considered. We present the results for each n-gram order alone and, in the final column, the result for the fusion of all n-gram classifiers. This way, we can see the relevance of each n-gram alone. The first line is our baseline experiment: no discriminative ranking, "golf score", independent templates for each n-gram with 3,000 units.

Formula	1-gram	2-gram	3-gram	4-gram	5-gram	All
No discrim.	74.4	43.1	38.7	44.3	58.6	34.40
1	52.6	37.6	32.9	34.4	49.6	**24.93**
2	53.8	40.6	35.8	39.2	56.9	32.93
3	53.4	38.6	32.4	35.3	54.8	29.10
4	53.4	39.7	34.2	35.9	56.1	30.38
5	**52.6 37.**	**6**	**32.8 34.**	**4**	**49.6**	**24.91**
6	52.7	39.3	32.8	34.4	49.4	25.23

Table 2: *Error rates for the different formulas.*

We can see that the discriminative ranking means an outstanding improvement (from 34.40 to 24.91, 27.6% relative improvement) and without it, results are similar to PPLRM, although slightly better (34.4 vs. 35.8%). We also observe in the table that, as could be expected, the trigram is the most powerful classifier. But what is extremely interesting is that the 4-gram is very close in performance, so it is a clear advantage over PPRLM, where obtaining reliable estimates for 4-gram is difficult and requires a huge training database.

The best result corresponds to the formula 5. Its advantage is that it normalizes the values between 1 and -1: 1 means that the n-gram appears in the current language but not in the other competing ones (n_2'=0), indicating that it is especially relevant for that language; -1 means just the opposite (n_1'=0), so the n-gram does not appear in the current language.

3.4. Influence of the template size

In Table 3 we can see the effect of the template size for the best configuration (formula 5 in previous section). So, the best results correspond to a template size equal to 5000, although it begins to saturate, and the improvements are only obvious in 4-gram and 5-gram, which could be expected because is where more units are left out the template. So, an alternative that we are considering is to have different sizes for each n-gram. Obviously, this optimum will depend in the number of allophones in each language, so some fine tuning will be needed for another setup.

Template size	1-gram	2-gram	3-gram	4-gram	5-gram	All
500	53.6	40.0	52.6	57.9	66.7	36.70
1000	53.0	38.8	44.8	47.5	61.0	31.62
2000	52.8	37.9	36.5	39.5	53.6	27.13
3000	52.7	37.6	32.8	34.4	49.6	24.91
4000	52.7	37.5	32.8	**34.0**	48.2	24.70
5000	52.7	37.5	32.8	34.0	**48.0**	**24.68**

Table 3: *Error rates for different template sizes.*

3.5. Influence of out-of-rank n-grams

One issue that we have to take into account is that for high order n-grams the amount of out-of-rank units increases. Our first approach was to assign these units the last position in the template (the template size). But it is clear that some penalty can be applied for those cases, so we decided to multiply the last position by a factor greater than 1 for out-of-rank units. In Table 4 we can see the results. The baseline uses 3,000 units.

We can see that there is an optimum for the penalty 1.7, with improvements from 3-gram to 5-gram, as could be expected (unigram and bigram have almost no out-of-rank units). Obviously, improvements saturate for large penalties. Another interesting result is that this penalty is more effective than just increasing the template size (which could be an alternative): in 3-gram, 31.8 (1.7 penalty) vs. 32.8 (5,000 units) in Table 3. And we obtain similar gains for 4-gram and 5-gram (slightly less). The probable reason is that just increasing the template size includes very unreliable n-grams, especially for trigram.

Penalty factor	1-gram	2-gram	3-gram	4-gram	5-gram	All
1.0 (base)	52.6	37.6	32.8	34.4	49.6	24.91
1.35	52.6	37.6	31.8	33.7	48.7	24.57
1.7	52.6	37.5	31.8	**33.4**	48.3	**24.47**
2.0	52.6	37.5	31.8	33.5	48.1	24.48
2.5	52.6	37.4	31.9	33.5	**47.9**	24.51
3.0	52.6	37.4	31.9	33.5	47.9	24.57

Table 4: *Error rates for penalties for out-of-rank units.*

3.6. Stratified rankings

After examining the rankings obtained, we considered the possibility of grouping n-grams with close values in n_1'' value considered for the ranking, so that we "smooth" the ranking.

Total units	Units/ cluster	1-gram	2-gram	3-gram	4-gram	5-gram	All
3000	1 (base)	52.6	37.6	32.8	34.4	49.6	24.91
3000	2	52.6	37.5	31.8	33.4	48.0	**24.54**
3000	3	52.6	37.5	31.9	33.5	47.8	24.55
3000	4	52.6	37.5	32.0	33.5	47.8	24.62
4000	2	52.6	37.5	31.8	33.5	47.6	24.62
6000	2	52.6	37.5	31.8	33.5	47.5	24.62

Table 5: *Error rates for penalties for out-of-rank units.*

In Table 5 we can see the results for different template sizes and number of units in each cluster. E.g. last row means a 3000 cluster template size with 2 units/cluster. Again we can see some improvements, especially for 4-gram and 5-gram.

4. Conclusions and Future Work

We have demonstrated that the n-gram Frequency Ranking approach overcomes PPRLM thanks to the longer span that can be modeled, especially for the great effect of the 4-gram, and partially of the 5-gram. To obtain this improvement, the following issues have been crucial:

- n-gram discriminative rankings with the normalized value for the number of occurrences are able to overcome PPRLM (31.6% relative improvement, 24.47 vs. 35.8).
- The ranking size should be between 3,000 and 5,000 depending on the n-gram order.
- Applying a penalty to out-of-rank n-grams may provide up to 1.7% relative improvement.
- Similar gains can be obtained with the stratified rankings.

As future work, we will consider different template sizes and penalties for the different n-grams to achieve the best result possible in the fusion of all of them.

5. Acknowledgements

This work has been supported by SD-TEAM (TIN2008-06856-C05-03), ROBONAUTA (DPI2007-66846-c02-02), and MA2VICMR (S2009/TIC-1542).

6. References

[1] Zissman, M.A., "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech&Audio Proc., v. 4, pp. 31-44, 1996.

[2] Cavnar, W. B. and Trenkle, J. M., "N-Gram-Based Text Categorization", Proc. 3rd Symposium on Document Analysis & Information Retrieval, pp. 161-175, 1994.

[3] Vatanen, t., Väyrynen, J. and Virpioja, S. "Language Identification of Short Text Segments with N-gram Models". Int. Conf. on. Language Resources and Evaluation (LREC'10), 2010.

[4] Cordoba, R., D'Haro, L.F., et al. "n-gram Frequency Ranking with additional sources of information in a multiple-Gaussian classifier for Language Identification". V Jornadas de Tecnología del Habla, pp. 49-52, 2008. Bilbao, Spain.

[5] Cordoba, R., D'Haro, L.F., et al. "Language Identification based on n-gram Frequency Ranking". Interspeech 2007, pp. 354- 357.

[6] Cresti, E. et al. "The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages". IV Int. Conf. on Language Resources and Evaluation, 2004.

[7] Ramasubramaniam, V., Sai Jayram, A. K. V., and Sreenivas, T. V. "Language Identification using Parallel Phone Recognition. Workshop on Spoken Language Processing". pp. 109-116. 2003.

[8] Cordoba, et al. "Cross-Task and Speaker Adaptation in a Speech Recognition System for Air Traffic Control". IEEE Aerospace and Electronic Systems Magazine, Vol. 21, No 9, pp. 12-17. 2006.

[9] Navratil, and J. Zühlke, W. "Double bigram-decoding in phonotactic language identification". ICASSP, Vol. 2, pp. 1115–1118. 1997.

[10] Córdoba, R., et al. "Integration of acoustic information and PPRLM scores in a multiple-Gaussian classifier for Language Identification". IEEE Odyssey 2006.

[11] Nagarajan, T., and Murthy, H. A. "Language Identification Using Parallel Syllable-Like Unit Recognition". ICASSP, pp. I-401-404. 2004.

Speech/Music classification by using the C4.5 decision tree algorithm

Diego Castán, Alfonso Ortega, Eduardo Lleida

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
dcastan@unizar.es, ortega@unizar.es, lleida@unizar.es

Abstract

In this work a study about different features for classification of audio frames into speech or music is presented. This paper focuses on the following set of features: High Zero-Crossing Rate Ratio (HZZCR), Variation of Spectral Flux (VSF), Low Short Time Energy Ratio (LSTER), Amplitude Modulation Ratio (AMR), Mel-Frequency Cepstrum Coefficients Variation (Var.MFCC) and Minimum-Energy Tracking (MET). In addition, we propose the use of a system based on a decision tree in order to combine the proposed set of features getting an improvement in the number of correct classifications. Experimental results on a broadcast radio database are presented showing that the selected features along with the use of the decision tree classifier allows the segregation of speech from music with a high degree of accuracy.

Index Terms: Speech/Music classification, time-domain features, frequency-domain features, cepstral-domain features and C4.5 decision tree.

1. Introduction

The growth and development of the Internet and Information and Communication Technologies (ICTs) in recent years has led to a dramatic increase of the number of multimedia documents. Therefore it is necessary to develop efficient systems that allow the organization, search and manipulation for a proper classification and information storage. Many of the current indexing systems seek for the desired information using tags that were defined by the users so the semantic description may be limited to a few words or phrases. As an alternative to this type of systems there exists a new field of research, Automatic Audio Content Analysis (ACA) that tries to develop information retrieval systems that analyze the audio data and extract the information directly from the audio signals. An example of automatic indexing system is given in Figure 1.

In our case, we are involved in developing an automatic system able to classify broadcast radio data which often includes sections with different types of signals like music, voice or music and voice. To develop our system we focus on broadcast ratio data transferred by "Aragón Radio", the public radio station in Aragón. As a first step, a fundamental task is to segment this type of signals and classify those segments into different classes according to acoustical criteria. After this preprocessing different subsystems can be applied like speaker recognition and automatic transcription of sections classified as voice.

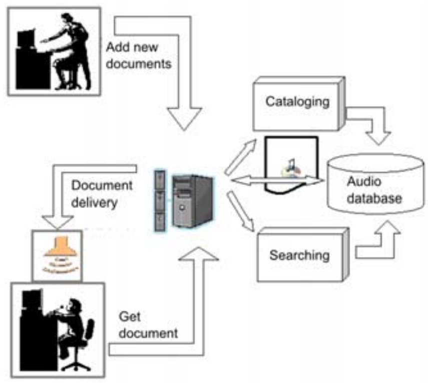


Figure 1: *Indexing system*

One of the difficulties in speech/music classification is to create a robust model for the identification of music signals. Speech is composed of a selection of typical sounds, therefore, it can be represented quite well by using simple models. However, music signals are composed of a big amount of different sounds produced by many different instruments with very different nature and often by many simultaneous sources. So music is defined by different styles and it can become difficult to create a single robust model. Moreover, the difficulty of our goal increases considering that we want to identify sections where speech and music overlap.

Former studies offer a rich set of papers about speech/music classification with a variety of applications. They describe which acoustic features can be used to emphasize the differences between music and speech. [1] presents a review of different solutions and the acoustic features used in each one of them. It can be found that many of the proposed solutions use the same features by varying the length of the feature vectors and the classification algorithm.

The solution proposed in this paper uses a set of six features. The combination of them is made by means of a C4.5 binary decision tree to provide the minimum classification error. The reduced dimensionality of the feature vector allows this algorithm to be used in real time. In addition, the model has been trained to identify the three most common classes that can be found on broadcast radio data: Music (M), speech (V) and the overlapping of both (A).

The authors would like to thank Aragón Radio for providing the data. This work has been partially supported by the national project TIN2008-06856-C05-04

The paper is arranged as follows: section 2 discusses the set of features used in the system, section 3 describes the classification algorithm, section 4 provides an evaluation of the system and is followed by the conclusion in section 5.

2. Features

Music and speech are described differently in both time and frequency domains. Speech signals are stationary for short periods of time (between 5 and 100 ms) while music signals should have larger stationary periods (around 200 ms). This means that speech has remarkable energy changes due to the alternation between voiced and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure.

In the spectral domain, speech signals present a rapidly and constantly changing energy distribution while music has a tonal structure which results in a more structured spectral distribution. We propose here a system that combines both temporal and spectral characteristics in order to be able to get the benefits from both descriptions.

2.1. Amplitude Modulation Ratio (AMR)

This feature calculates the relationship between local minima and local maxima in the envelope signal as detailed in (1). The envelope is obtained by filtering the signal with a lowpass filter with 25 Hz cutoff frequency. The alternation of high energy and low energy segments (vowels and consonants) in a speech signal causes the amplitude modulation ratio to be higher for speech and lower for music signals.

$$ID = \frac{V_{max} - V_{min}}{V_{max} + V_{min}} \qquad (1)$$

2.2. High Zero-Crossing Rate Ratio (HZZCR)

One of the most widely used acoustic feature in music/speech classification is the Zero-Crossing Rate (ZCR) [2]. However, there are several variants of this feature like the High Ratio Zero Crossing Rate (HZCRR) [3]. HZCRR is defined as the ratio of the number of frames whose ZCR are above 1.5 times the average zero-crossing rate as described in (2). The advantage of HZCRR against ZCR is that HZCRR provides a more robust description of music and speech by removing the small variations that may occur with ZCR. Hence, for speech signal, HZCRR variation will be greater than that of music.

$$HZCRR = \frac{1}{2N} \sum_{n=0}^{N-1} \left[sgn(ZCR(n) - 1.5\overline{ZCR}) + 1 \right] \qquad (2)$$

To obtain the HZCRR, first we calculate the ZCR in 20 ms windows and 10 ms frame shift. HZCRR is calculated then in 1 second windows. A median filter is applied to smooth the resulting values.

2.3. Variation of Spectral Flux (VSF)

The spectral flux provides an idea about the changes that occur in the shape of the spectrum frame by frame [4]. Speech exhibits an alternating sequence of noise-like segment with some others that present a more stationary behaviour. On the other hand, music presents a tonal structure due to a succession of periods of relative stability (notes or chords). In other words, the speech signal is distributed along the spectrum in a more random way than music does. The Spectrum Flux can be defined as the ordinary Euclidean norm of the delta spectrum magnitude which is calculated as:

$$SF = \|\mathbf{S}_i - \mathbf{S}_{i-1}\| = \frac{1}{N} \left(\sum_{k=0}^{N-1} \left(S_i(k) - S_{i-1}(k) \right)^2 \right)^{\frac{1}{2}}, \qquad (3)$$

where S_i is the spectrum magnitude vector of frame *i*, which is defined as the DFT with a frame size of 20 ms and frame shift of 10 ms. Finally, the variation is calculated every 0.2 seconds.

2.4. Low Short Time Energy Ratio (LSTER)

The features based on the energy of the input signal are very popular for speech/music classification. A reasonable generalization is that speech follows a pattern of high-energy periods for voiced sounds followed by low-energy segments for unvoiced sounds. On the other hand, the envelope of music is less likely to exhibit this behaviour. In this solution we use the Low Short Time Energy Ratio [5] that is defined as the ratio of the number of frames whose short time energy is less than 0.5 times the average short-time energy,

$$LSTER = \frac{1}{2N} \sum_{n=0}^{N-1} \left[sgn(0.5\overline{STE} - STE(n)) + 1 \right] \qquad (4)$$

In the same way we did for the HZCRR, we calculate the energy in 20 ms windows with a frame shift of 10 ms. LSTER is calculated in 1 second windows. A median filter is applied to smooth the resulting values.

2.5. Mel-Frequency Cepstrum Coefficients Variation (Var.MFCC)

It is well known that the mel frequency cepstral coefficients are a compact and efficient representation of speech [6].

In our approach, MFCCs are extracted every 10 ms using 20 ms windows with a 40 channel filter-bank. Then, we used 13 coefficients that are summed and we calculate the variation every 0.1 seconds.

$$\sigma^2_{sumMFCC_i} = \frac{1}{N} \sum_{k=0}^{N-1} \left(sumMFCC_i(k) - \overline{sumMFCC_i} \right) \qquad (5)$$

2.6. Minimum-Energy Tracking (MET)

This feature tries to describe the natural pauses that can be found in speech due to the nature of the human speech production mechanisms. These segments are characterized by the reduction of energy (represented by MFCC coefficient C_0) below a certain threshold. If C_0 is above the threshold for longer than 1.5 seconds, the frame will be classified as music, otherwise it will be classified as speech. This feature has certain limitations in sections of speech with background music and if the music level is high, the section can be classified as music. However, on those sections where only speech is present, this feature provides good results.

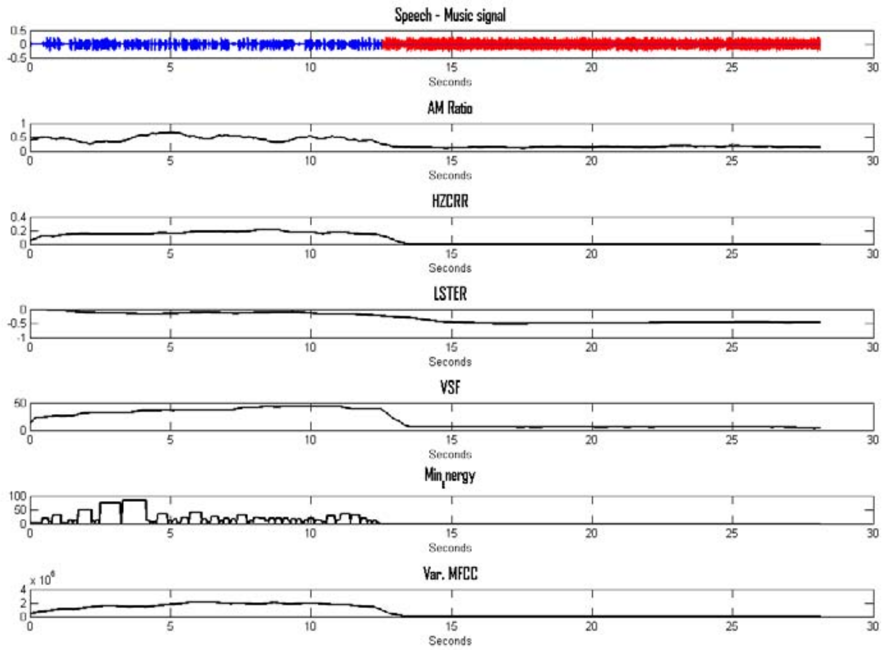


Figure 2: Behavior of the features with speech signal (blue) and music signal (red)

Finally, in Figure 2 the evolution of different features in a signal, where half of the samples are speech and the other half belong to music, is presented. Due to the asynchronous nature of some of the selected features like amplitude modulation ratio and minimum-energy tracking we have proceeded to the synchronization feature vector resampling all signals to 100 Hz.

3. C4.5 classification tree

When comparing different speech/music classification systems previously reported in the literature, the main difference that can be found among them is the classification algorithm they use. Artifitial Neural Networks (ANN), K Nearest-Neighbors (KNN) and Gaussian Mixture Models (GMM) are widely used [1].

In addition, decision tree methods have also been used for speech/music classification. In axis-parallel decision tree methods, a binary tree is constructed in which at each node a single parameter is compared to some constant. If the feature value is greater than the threshold, the right branch of the tree is taken; if the value is smaller, the left branch is followed. After a series of these tests, one reaches a leaf node of the tree where all the objects are labeled as belonging to a particular class. These are called axis-parallel trees because they correspond to partitioning the parameter space with a set of hyperplanes that are parallel to all of the feature axes except for the one being tested.

The C4.5 algorithm builds binary decision trees from a set of training data using the concept of *information entropy* that was developed by Quinlan in 1993 [7]. The training data is a set $S = s_1, s_2, ...$ of already classified samples. Each sample $s_i = x_1, x_2, ...$ is a vector where $x_1, x_2, ...$ represent our five features of the sample. The training data is augmented with a vector $C = c_1, c_2, ...$ where $c_1, c_2, ...$ represent the class to which each sample belongs. In our approach, the classes are *Speech*, *Music* and *Both*.

The algorithm considers all possible tests that can divide the set of data and select the test that contains greater information gain. For each continuous attribute binary testing is performed. At each node, the system must decide which test should be chosen to split the data. In our case, the features are continuous numerical values so the algorithm searches a threshold to get the best result as shown in Figure 3.

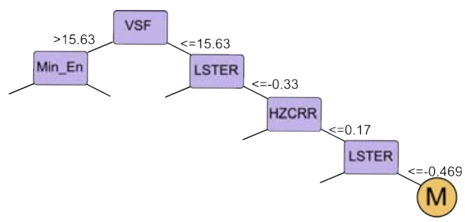


Figure 3: *C4.5 classification tree*

An additional constraint has been applied successfully: for any division, at least two of the subsets Ss_i must contain a reasonable number of cases. It uses a technique known as *Gain Ratio*. It consist on an information based measure that considers different numbers (and different probabilities) of the test results.

4. Evaluation and Results

The proposed approach has been evaluated by using an audio database supplied by Aragón Radio. From this database 20 tracks that alternate segments of music, speech or both, have been selected and manually annotated. Each segment has a duration between 10 and 30 seconds an the total amount of data is around one hour. This corpus contains clips with one voice, two voices, spots, speech-music mixed, instrumental music, songs, capella music and many other musical styles.

Feature	Accuracy
Var.MFCC	88.37%
HZCRR	74.34%
MET.	83.35%
LSTER	78.51%
VSF	79.11%
AMR	83.06%

Table 1: Features results

To check the reliability of the features in the classification of music and speech, we proceeded to the evaluation of those on the database described above. In this first experiment the overlapping of music and speech was not taken into account, the purpose was only to asses wether music or speech was present by making a simple threshold comparison for each one of the features. The results are presented in Table 1. As it can be seen, the variation of the MFCC coefficients is the feature that best segregate speech from music. It is also interesting to note that energy-based features and spectral features obtained around 80% of correct classifications.

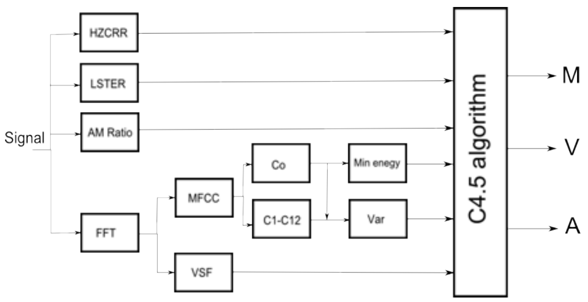


Figure 4: *Block diagram of feature extraction*

After evaluating the performance of each one of these features separately, it seems reasonable to implement a classification method that combines all of these features to improve the results and also to classify the overlapping between speech and music. The C4.5 algorithm consists of two phases: a training or definition of the tree and a test. This has been carried out with cross validation by dividing the database into four subsets randomly. The system used for this purpose can be seen in Figure 4. Thus, the results with the classification tree are presented in Tables 2 and 3.

As can be seen in the results, the proposed classification method greatly improves the results of the individual features

Correctly Classified Frames	297113 → 99.83%
Incorrectly Classified Frames	487 → 0.16%
Total Number of Frames	297600

Table 2: Result of C4.5 tree

	Classification		
Original	Speech	Music	Both
Speech	99.77%	0.06%	0.17%
Music	0.03%	99.94%	0.02%
Both	0.35%	0.09%	99.56%

Table 3: Confusion Matrix

showing that the selected set of features complement each other. It is also interesting to note that the biggest classification error belongs to the class representing the segments in which music and speech overlap. This result could be expected beforehand since this class is less homogeneous than the other two and fewer samples are available for training.

5. Conclusion

In this work a set of acoustic features has been selected and evaluated for unsupervised speech/music classification on a broadcast radio database. Time-domain, frequency-domain and cepstral-domain features have been considered along with a decision tree based classification method. Experimental results show that the proposed approach can be successfully applied to the considered task. The use of the decision tree method outperforms the results obtained with each one of the features individually what highlights the complementarity among them. The proposed combination by using a decision tree classifier obtains an improvement of more than 10% over the most discriminative feature allowing also the introduction of a new class which is the overlapping of the two previously defined classes, music and speech, which occurs very often on broadcast radio.

6. References

[1] Lavner, Y and Ruinskiy, D, “A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation”, EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[2] Saunders, J. , “Real-time discrimination of broadcast speech/music”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP, 1996.

[3] Lu ,L. and Zhang, H.J., “Content analysis for audio classification and segmentation”, IEEE Transactions on Audio, Speech, and Language Processing, 2002.

[4] Scheirer, E. and Slaney, M., “Construction and evaluation of a robust multifeature speech/music discriminator”, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP, 1997.

[5] Alexandre, E. and Rosa, M., “Application of Fisher linear discriminant analysis to speech/music classification”, Proc. of the 120th Audio Engeneering Society Convention AES, 2006.

[6] Quatieri T.F., “Discrete-Time Speech Signal Processing”, Prentice-Hall, Englewood Cliffs, NJ,USA, 2001.

[7] Quinlan, R., “C4.5: Programs for Machine Learning”, in Morgan Kaufmann Publishers Springer Netherlands, 1993.

FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Automatic Metadata Extraction from Spoken Content using Speech and Speaker Recognition Techniques

Héctor Delgado, Javier Serrano, Jordi Carrabina

Software-Hardware Prototypes and Solutions Lab, Autonomous University of Barcelona, Spain
hector.delgado@uab.cat, javier.serrano@uab.cat, jordi.carrabina@uab.cat

Abstract

Today information extraction plays a significant role in management of massive data quantities for different purposes. One of the open challenges in this field is the automatic extraction of information from audio streams. This paper describes a useful metadata extraction system which performs a powerful combination of speech and speaker recognition tasks. The system carries out the speech transcription through a Catalan language recognizer based on Hidden Markov (HMM) tied-state cross-word triphones acoustic models, Mel Frequency Cepstral Coding (MFCC) and N-gram language modeling. In addition, a speaker diarization is performed using HMM based segmentation and Perceptual Linear Prediction (PLP) feature extraction. Both speech-to-text transcription and speaker diarization can be utilized as annotation data for multimedia content. In order to make indexing and retrieval more flexible and efficient, the extracted metadata is stored using the MPEG-7 multimedia content description interface. The system has been successfully tested on the recordings of the plenary sessions of the Catalan Parliament.

Index Terms: Metadata extraction, Automatic speech recognition, Speaker diarization, GMM, HMM, MPEG-7

1. Introduction

Multimedia content volume has increased hugely since storage capacity has become almost unlimited, and information technologies has extended widely. Content management, indexing and retrieval are key challenges that are increasingly becoming more difficult, due to the massive data quantity. Since manual annotating of such massive content is not viable, it forces content to be annotated automatically through information extraction approaches.

Speech inside audio streams is a huge information source where data may be obtained from, in an automatic way. Speech technologies are able to extract different kinds of information from audio. Instead of manual annotation of audio information, speech technologies offer automatic extraction which is significantly efficient in terms of time and accuracy, particularly in huge amount of repositories. Spoken content automatic transcriptions of the audio/video streams are also suitable for automatic on-line or off-line subtitling, word spotting, or as support for either foreign or hearing impaired people. On the other hand, automatic speaker diarization allows direct access to the parts where particular speakers participate.

This paper describes a useful metadata automatic extraction system which focuses on the spoken content. The system performs a powerful combination of speech and speaker recognition tasks. The metadata extraction system carries out a speech transcription by means of a Catalan recognizer based on cross-word tied-state triphone HMMs, MFCC and N-gram language models for Catalan language. In addition, a speaker diarization is performed using HMM based segmentation and PLP feature extraction. For both transcription and speaker diarization, a comparison of performance has been performed with different configurations, in order to improve results. Once the information has been extracted, it is automatically stored using the MPEG-7 content description interface.

The system may be used in different application domains. Concretely, the system has been successfully applied on the recording of the plenary sessions of the Catalan Parliament. Parliament daily generates big amounts of video and audio files. Therefore, it is a domain where content management is a key task, and it could take advantage of the automatic metadata extraction system.

The paper is organized as follows. Section 2 refers to the problem of managing big amounts of audiovisual content and how spoken content can be exploited to extract metadata in order to obtain suitable annotation information. Section 3 describes in detail how the system works, the experiments carried out, the results obtained and a discussion. Finally, some conclusions are given in section 4.

2. Spoken content management and automatic metadata extraction

Multimedia content management is a key task in any big repositories of audiovisual content. Annotation metadata is indispensable for an effective management, but it cannot be obtained in a manual way since it is an inviable and very time-consuming task. For that reason, information extraction must be done in an automatic way. Metadata extraction could be performed at several levels (from video and audio), but one of the most important information source is the spoken content. Numerous approaches have been developed in automatic speech recognition and automatic speaker recognition. Thus it is a good idea to exploit these kind of systems to extract metadata.

Particularly, speech-to-text transcriptions and speaker diarizations are specially useful as annotation data for multimedia content. Therefore an automatic metadata extraction that acts over the spoken content and extract the speech transcription and performs the speaker diarization is proposed.

The speech transcriptions of the spoken content provides a wide range of possibilities. Usually ASR systems output contains information about each recognized word and temporal information about the occurrence. On the management side, indexing and retrieval can be carried out utilizing the spoken content transcription. On the user side, this data can be used to different purposes, like to perform keyword spotting, providing direct access to the desired words inside a given audiovisual stream, or as an complementary resource to the audiovisual content (subtitles).

Regarding the speaker diarization, it has also applications for both managers and users. On one hand, speaker diarization may be utilized as annotation data to be applied on content indexing and retrieval purposes. On the other hand, users can take advantage of the diarizations when browsing through content, allowing direct access to the segments where particular speaker participates.

The extracted metadata must be stored in a convenient way in order to be utilized for indexing and retrieval purposes. Therefore, such metadata is stored using the MPEG-7 multimedia content description interface, centering on the spoken content description schemes. It allows to apply the great variety of indexing and retrieval techniques developed in previous research about spoken document retrieval over MPEG-7 descriptions. The fact of following MPEG-7 also raises content exchange and compatibility between different platforms.

The automatic metadata extraction from spoken content can be addressed to a great variety of domains, where it is necessary to manage great quantities of audiovisual content (broadcast news, meeting, etc). One target domain is the audiovisual material recorded at the Parliament. Everyday the plenary sessions are registered on audio and video. In Catalonia, such material from the Catalan Parliament is in the public domain and is made available to citizens on the Internet. It has 2 important consequences. First, a better accessibility to content would improve the user experience when browsing trough content. Secondly, there are public resources available to the research community to be used to develop speech recognizers (for instance, the transcriptions of the sessions can be used to train language models) and speaker diarization systems.

3. Experiments and results

This section describes the experiments in detail. Both speaker diarization and speech-to-text transcription have been performed using the Hidden Markov Model Toolkit (HTK) [1].

Once the transcription and the speaker diarization are obtained, the results are parsed and the MPEG-7 description is generated.

3.1. Speaker Diarization

The current subsection describes the speaker diarization system set-up in detail, particularly the training and test data, the feature extraction and the speaker modeling and configuration.

3.1.1. Training and test data

The training and test data are recordings from the Catalan Parliament. The original 16 bit, one-channel and 48 kHz audio has been down-sampled to 16 kHz. 4 hours of speech were used for training the models and 1 hour for testing.

3.1.2. Feature extraction configuration

Some techniques has been successfully applied on speaker recognition tasks, such as LPC, LPC-Cepstra, MFCC and PLP. In previous work carried out at CEPHIS about speaker diarization in broadcast news audio [2], PLP features have been empirically proven to be beneficial for this purpose. For that reason, PLP method has been chosen for the current task. Firstly, a speech signal processing is made for each type of feature. A 0.97 coefficient pre-emphasis filter is applied, and a 25 ms Hamming window that scrolls each 10 ms is used to obtain signal frames. Then, a feature vector of 12 PLP coefficients is obtained from each frame. Finally, the energy coefficient, delta- and delta-delta features (time derivatives) are added to the feature vectors. Further detailed explanation of the PLP technique can be found in [3].

3.1.3. Speaker Modeling

After a previous study of the Parliament audiovisual content, a categorization of the participating speakers must be carried out. Generally, one can divide the participants into these categories: the premier, the government ministers, the president of the Parliament and the representatives of the parliamentary groups.

In addition, there are sound events that are not properly speech, such as background noise, murmur or silence. It must be taken into account that politicians remain for at least 4 years, and usually 1/3 of the parlamentarians change. Consequently, it worth to develop models for each member of the parliament.

Having done the categorization, it is necessary to decide how to model the different categories in a useful way that facilitates direct access to each speaker participation. Some participants are especially relevant, such as the premier, the president of the Parliament and the government members. For these speakers it is highly useful to have their particular segments. On the other hand, some speaker like the representatives of the parliamentary groups could be merged into a general ‘other speakers’ category.

After this study, models will be developed for the following cases: single-model for the premier, single-model for the president of the Parliament, one single-model for each government minister, one ‘shared’ model for the representatives of the parliamentary groups and single-model for silence/background noise.

3.1.4. HMM configuration

An HMM for each case listed in subsection 3.1.3 is created in this stage. Each HMM has three states in a left-to-right topology. Only the central state has a GMM as emitting probability density function. Diagonal covariance matrices have been proved to be beneficial, thus they are used here. Particularly, there are three main reasons to use only diagonal covariance instead of full covariance matrices [4]. Firstly, the density modeling of an Mth order full covariance GMM can be achieved using a larger order diagonal covariance GMM. Furthermore, diagonal covariance matrices GMM are computationally more efficient than full covariance matrices GMM. Finally, empirically diagonal matrix GMM have been observed that outperform full matrix. Then, the single-gaussian model is split into 8, 16, 32 and 64 mixture gaussians. The model parameters are iteratively re-estimated using the implementation of the Baum-Welch algorithm in HTK (HERest tool).

Once the model set is obtained, the diarization is performed through the Viterbi algorithm (HVite tool). One general HMM is generated from the individual models creating a model loop.

3.2. Speech-to-text transcription

A description of the speech-to-text system is given next, focusing on the training and test data, acoustic modeling and language modeling.

3.2.1. *Training and test data*

The acoustic models have been trained using the SpeechCon Catalan speech corpus [5]. The corpus has spontaneous and read speech from 550 speakers, recorded with four microphones at different distances. Each utterance is stored in 4 independent (one per microphone) 16 bit, 16 kHz uncompressed audio files.

The test consists of 13 minutes of speech extracted from the recordings of the Catalan Parliament plenary sessions. The original 16 bit, one-channel and 48 kHz audio has been down-sampled to 16 kHz. The nature of the training and test data is very different (clean speech versus noisy, non spontaneous speech). For that reason, an adaptation stage will be necessary in order to improve accuracy.

The audio files are then parametrized into a 39-dimensional feature vector consisting on 12 cepstral coefficients plus the 0th coefficient, deltas and delta-deltas.

3.2.2. *Acoustic modeling*

Firstly, a set of 40 HMMs (39 monophones plus 1 silence model) is obtained. The HMM consists of 3 output states with self-loops, in a left-to-right topology. This set is calculated by means of a flat initialization, and each model is re-estimated using the Baum-Welch algorithm. A short pause model is then created by cloning the central state of the silence model and adding a skipping transition.

The phone-level transcriptions are converted into cross-word transcripts. New triphone models are created by cloning the central state of its corresponding monophone. Every triphones with the same central monophone share their transition matrices. Then the model parameters are re-estimated again.

Since the triphone set do not cover the all the possible triphones in the language, they are synthesized and their states are tied to physical models states. It also contributes adding a more robust set of models. The tying process is carried out through decision tree clustering that uses linguistically motivated questions about a triphone's context. The resulting tied-state triphones are re-estimated.

Finally, the single gaussian models are split and re-estimated consecutively into 2, 4, 8, 16 and 32 gaussian components, obtaining the final set of tied-state triphone models.

In addition, a further improvement of the ML models is carried out. A set of discriminatively trained models is developed from the ML set using MMI, running 4 iterations of the EBW algorithm (HMMIRest tool).

3.2.3. *Language modeling*

The language model is a 64k word based 3-gram LM that has been developed using the transcriptions of the plenary sessions of the Catalan Parliament, consisting of around 24 million words. The completed 167000 word vocabulary was reduced to 64k for two reasons. Firstly, the majority of the words inside the original vocabulary have very few occurrences and they are considered as rare words that are hardly pronounced. Therefore suitable probabilities cannot be calculated for those uncommon words. Secondly, the decoder used for the experiments imposes a restriction on the vocabulary size of 64k words. Thus, the 64k most common words in the corpus were taken. The utilized training tool was also HTK (LBuild).

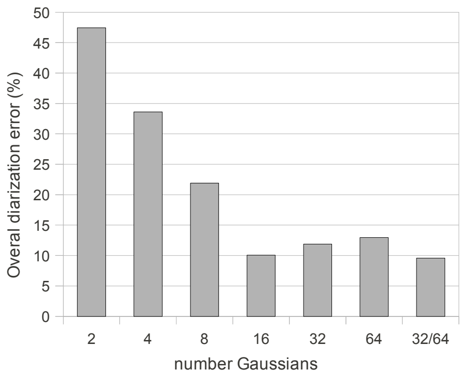


Figure 1: *Speaker diarization results.*

3.3. Results and discussion

3.3.1. *Speaker diarization results*

This section shows the results obtained from the experiments carried out to test the speaker diarization system. The speaker diarization results evaluation has been made following the Spring 2006 Rich Transcription Meeting Recognition Evaluation Plan (NIST) [6]. This plan proposes a "who spoke when" diarization scoring.

The total error percentage is called Overall Diarization Error (ODE). Fig. 1 shows the ODE obtained using 2, 4, 8, 16, 32 and 64 gaussians. Furthermore, a combined model of 32 gaussians for particular speakers and 64 gaussians for 'other speakers' and silence has been tested.

Analyzing the results, it can be observed that the ODE decreases significantly when incrementing the number of gaussians, up to 16 components. However, this trend changes when increasing the number of gaussians at this point. It has been observed that the models for known speakers perform better when using 64 gaussian components, whereas the 'other speaker model' works better with 32 components. Thus, a combined 32 and 64 gaussian model set has been evaluated, decreasing the ODE up to a 11.68% ODE. In any case, the combined model hardly outperform the 16 gaussian models (10.06% ODE), and it may not worth it due to the computational cost.

3.3.2. *Speech-to-text results*

After carrying out the speech-to-text experiments, the obtained results are shown below. Firstly, the Acoustic models have been tested on clean speech, recorded in office environment. Fig. 2 shows the Word Error Rate (WER) for the different trained models: maximum likelihood models (ML) and discriminatively trained models using MMI criterion. The same models have been evaluated performing MLLR speaker adaptation as well.

The same experiments have been done using real noisy speech from the Catalan Parlament, whose figures are depicted in Fig. 3. In this case, the adaptation carried out is not based on particular speaker, but on the general features of the speech in the Parliament.

It can be observed that there is a significant difference of performance when changing the characteristics of the audio. For clean audio recorded in office environment, that is quite similar to the features of the training corpus, WER decreases up to 20,21% and 16,14% using speaker adaptation and discrimi-

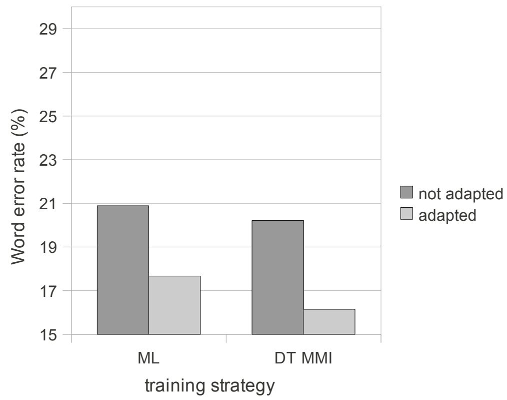


Figure 2: *Word error rate in clean speech experiments.*

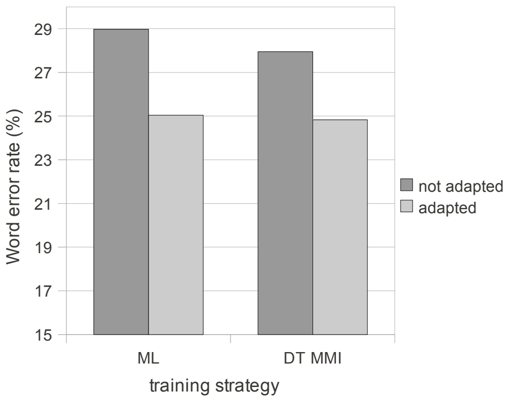


Figure 3: *Word error rate in parliament speech experiments.*

natively trained models. When using Parliament speech, the accuracy decreases to 24,3 %due to the environmental noise and overlapped speech, although an improvement of around a 3% is achieved with MLLR adaptation.

3.4. Metadata representation in MPEG-7

Once the transcription and the speaker diarization have been obtained, it is necessary to put them together to form the MPEG-7 description. For that purpose, a software have been developed in Java. The tool parses the files generated by HDecode (transcription) and HVite (speaker diarization), and generates a MPEG-7 description in XML, using the Spoken Content Description Scheme [7]. In relation to the speech-to-text transcription, the best word output is represented by means of the *SpokenContent* and *Lattice* descriptors.

4. Conclusions

We have proposed a system that agglutinates speech recognition and speaker diarization. The system have been successfully applied on the recording of the plenary sessions of the Catalan Parliament. The speech transcription is performed using a set of tied-state cross-word triphone models and a 3-gram LM. In relation to the speaker diarization, a categorization of the speakers that participate in the Parliament recordings was carried out and a set of HMM was developed using different number of gaussian components. Then, HMM based speaker diarization was done.

As far as the speech-to-text transcription is concerned, the results state the importance of the audio quality in automatic speech recognition. Noisy speech means an important degradation of accuracy. It can be managed by means of the application of robust ASR and noise reduction techniques. However, word error rates using clean audio decreases under 17%. This figure is supposed to get lower using a more robust LM trained with a better textual corpus. In addition, discriminatively trained models have improved error rates around 1%, but in some cases it could not worth it due to discriminative training is a very time-consuming process. In other works, discriminative training has resulted to be beneficial for large vocabulary tasks. Thus, more work is needed in order to get better error rates with discriminatively trained models. Discriminative training using MPE criterion has been proven to be useful for large vocabulary tasks. Therefore, it should be tested as future work. Finally, the use of speaker adaptations improves significantly the accuracy.

Focusing on the speaker diarization system, results reveal that PLP features are adequate to perform the speaker segmentation in our data set. This means that PLP features extract inter-speaker variability correctly. Regarding the GMM, an increment in the components number not always improve error rates. Particularly, increasing mixture components over 16 does not contribute with any improvement.

Finally, the extracted metadata was stored according to the MPEG-7 content description interface automatically through the Java parser. That results very useful because it allows the use of a great variety of content based indexing and retrieval techniques.

5. Acknowledgements

This work has been partly founded by the Spanish Ministerio de Ciencia y Innovacion, Project Reference TEC2008-03835/TEC. This article is supported by the Catalan Government Grant Agency Ref. 2009SGR700

6. References

[1] S. Young and G. Evermann, *The HTK Book*. Cambridge University Engineering Department, March 2009.

[2] H. Delgado, "Segmentación de Vídeo mediante Reconocimiento de Locutores," Master's thesis, Universitat Autónoma de Barcelona, June 2009.

[3] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[5] Asunción Moreno, Albert Febrer and Lluis Márquez, "Generation of Language Resources for the Development of Speech Technologies in Catalan," in *Proceedings of the Language Resources and Evaluation Conference LREC 06*, Genova, Italy, 2006.

[6] NIST, *Spring 2006 (RT-06S) Rich Transcription Meeting*, 2006.

[7] "MPEG-7 Overview," Internet: http://www.chiariglione.org/mpeg/standards/mpeg-7/.

On the Detection and Classification of Frames from European Portuguese Oral and Nasal Vowels

Yosvany Llerena Rodriguez[1], *António Teixeira*[2]

[1]IEETA, Universidade de Aveiro, Portugal
[2]DETI/IEETA, Universidade de Aveiro, Portugal
llerena@ua.pt, ajst@ua.pt

Abstract

Our aim is to perform a comparative evaluation of potential of oral versus nasal sounds in a European Portuguese Speaker Verification system. For that, we report, in this paper, the work on the necessary detection of the relevant segments. Implemented detection and classification consists in a typical cascade of speech framing, feature extraction and the use of classifiers. A total of 31 different features - including a subset of the features used recently by Pruthi and coworkers for American English nasal vowels detection - were extracted from each frame. Taking into account the small dataset restriction, we selected three classifiers: the well known SVM; the, more recent, Naive Credal Classifier 2 (from the Naive Bayes family of classifiers) and a Metaclassifier based on boosting (MultiBoostAB). Results, using a small database, showed as the best classifier the MultiBoostAB. Best results for Recall, Precision and F-measure, of 87.04, 88.0 and 87.5 %, were obtained for this classifier when trained with an equal number of samples of each class and non-including the first 40% of the production of the nasal vowels.

Index Terms: Nasal vowels, Blind Segmentation, Naive Credal Classifier 2, MultiBoost, European Portuguese.

1. Introduction

Segmentation and labeling of speech material according to phonetic or similar linguistic rules is a fundamental task in speech processing. A "Blind speech segmentation procedure allows a speech sample to be segmented into sub-word units without the knowledge of any linguistic information (such as, orthographic or phonetic transcription) [1].

One application of speech segmentation is on Speaker Verification systems, to provide a sub-word level segmentation [1].

Nasal sounds are more speaker-dependent due to the considerable differences among individuals shapes of the resonators involved [2]. These differences on size and shape of the nasal cavities can cause differences in the spectral characteristics of nasal murmurs for different speakers [3]. As Portuguese is a language rich in nasal sounds, the segmentation, detection and classification of these sounds needs attention.

According to [4] a vowel nasalization detector is essential for speech recognition (particularly the recent landmark-based recognition) in languages with phonemic nasalization (as is the case of Portuguese). To best of our knowledge no such detector has been developed yet.

Our goal is to perform a comparative evaluation of oral versus nasal sounds potential in an EP speaker verification system. For that, we report, in this paper, the work on the necessary detection of the relevant segments. As no comparable database as the TIMIT used by, for example, Pruthi [5], for our investigations, the problem of developing the speaker independent segmentation system included the constraint of being developed with only a small amount of data (pre-existent if possible), including noisy recording conditions e less than perfect articulations.

1.1. Related work

Recently the detection/segmentation of nasal vowels was addressed by Pruthi and coworkers. In [5], the authors evaluate, for American English, a set of nine acoustic parameters (APs). Those APs came from previous studies and almost referenced as capable to well describe the vowel and nasalize segments. In addition, they work only with the middle 1/3rd of the frames for oral vowels and the last 1/3rd of the frames for nasalized ones. They used SVMs with linear and Radial Basis Function (RBF) kernels as classification methods and three different databases (StoryDB, TIMIT and WS96/97). The SVM outputs were mapped to pseudo-posteriors histogram to achieve the final decision using a probabilistic measurement. Results for selected APs were compared with other two sets (6 other features and 39 MFCC). Best accuracies were achieved by the RBF Kernel SVM with proposed APs, obtaining accuracies of 96.28%, 77.90% and 69.58%, respectively for the databases StoryDB, TIMIT and W96/97.

Most representative works on Portuguese phoneme segmentation were performed using HMMs, not suitable for small datasets situations, such as the one addressed in this paper.

2. System Overview

The speech waveform is first split into small segments (frames). All subsequent processing is frame based.

First step consists in feature extraction (details in sec. 3) for each of the frames of the input signal. As our target classes are included in the Voiced part of the speech signal, our second step is to classify each frame as Voiced or Unvoiced. The algorithm used for this step is based in [6]. As a third step, classifiers are applied to classify each frame in one of the classes of interest: Oral Vowels, Nasal Vowels, Other. The classifiers considered were chosen by their potential to being trained with a small data set.

3. Features

An extended set of 31 features was used in our experiments. A subset of the features used recently by [5] was implemented, being the basis of our feature set. Our purpose is to investigate if they can be useful for languages other than English, particularly in European Portuguese (EP) - a language with nasal vowels in its phonological inventory. Due to their higher relevance for the present work, these features are described in the following subsections. Other feature groups were also included: frequencies (F0,F1,F2,F3,F4,Instant Frequency), Childer's [7] nasal and vowel detection features (Nasal Rate, Vowel Rate, Volume of Low and High Band frequencies) and others commonly used in speech segmentation (Energy, Energy rate, Energy in low and high bands, ZCR, Entropy, Spectral features (rollOff, centroid, flux, flattening)).

3.1. teF1

Teager energy operator was used by [8] for hypernasality detection, using a pitch synchronous approach. As an alternative [5] proposed the use of the correlation between the Teager energy profiles of lowpass filtered speech and bandpass filtered speech, the correlation between the Teager energy profiles of narrow bandpass filtered speech and wide bandpass filtered speech centered around two different frequency regions be considered. In this case, the frequency regions were centered on the first two formant frequencies, obtained from a formant tracker.

Teager energy profile, $\Psi_d[x(n)]$, for a signal $x(n)$ is calculated as:

$$\Psi_d[x(n)] = x^2(n) - x(n+1)x(n-1) \tag{1}$$

The teF1 feature can be estimated as the correlation between Narrow and Wide band filtered signals:

$$teF1 = \rho(\Psi_d[sNBF1], \Psi_d[sWBF1]) \tag{2}$$

where sNBF1/F2 and sWBF1/F2 are the narrowband and wideband filtered speech signal centered around F1/F2 respectively. The narrowband filter uses a 100 Hz bandwidth and the wideband filter was set to 1000 Hz both with a 200 order filter.

3.2. nPeaks40dB

This feature [5] was designed to capture the large extra poles across spectrum as author mention. Basically count the number of peaks in limits of 40 dB of maximum dB amplitude of complex cepstral spectrum. Including only peaks between $0 - 4000$ Hz.

3.3. F1BW

Pruthi in [5] suggest that even though bandwidths of oral formants may not increase due to the losses in the nasal cavity, the bandwidths of these formants may seems to be wider because of unresolved poles which appear at frequencies very close these oral formants. The feature in question means is obtained from the ESPS formant tracker algorithm, as made available in the Snack ToolKit.

3.4. A1-H1fmt

In [9] author distinguish the difference between the amplitude of the first formant (A1) and the first harmonic (H1), and change in A1-H1 over time as being correlated to the perception of nasality. A reduction in A1-H1 is expected because A1 reduces with nasalization also confirmed by [5].

3.5. std0-1k

Standard deviation (STD) of the local spread of energy around the centre of mass (CM) was found to be a very good measure of nasalization. The feature can be estimated by measuring the second moment of local energy around CM. The term "local" was defined to include all energy within a specified frequency radius of CM [10].

Calculated between two frequency ranges, f_1, and f_2, the CM, $\bar{f}$, is defined as:

$$\bar{f} = \frac{1}{A_1} \sum_{f=f_1}^{f_2} fX(f) \text{ with } A_1 = \sum_{f=f_1}^{f_2} X(f) \tag{3}$$

where $X(f)$, is the value of the Discrete Fourier Transform (DFT) spectra at frequency f. For nasalized vowels the CM must be computed between 0 and 1 kHz, which covers the first formant range of most men and women.

Before STD estimation and proposed by Glass in [10], the DFT spectra was windowed with a trapezoidal windows before the CM computation to reduce the CM function sensitivity to sudden changes at the end points, such as formant passing below 1000 Hz. The windows shape is flat between 100 and 900 Hz, and had 100 Hz tapers at each end. Applying this windows the spectra were not sudden changes in CM caused by the marginal movement in energy across the upper boundary.

In [5] the author also proposes, before CM calculation, to set any amplitude value less than threshold (20 dB below maximum) equal to threshold, and then subtract threshold from all values to set floor to zero.

4. Classifiers

Taking in consideration the small dataset restriction, as classifiers, besides the commonly used SVM, 2 others were selected. One from the Naive Bayes family, the Naive Credal Classifier 2 (NCC2) [11]; the other a Meta classifier using Boosting [12].

4.1. Support Vector Machines (SVM)

SVMs (ex: [12]) are based on the concept of decision planes that separates between a set of objects having different class memberships. Most classification tasks demand complex structures in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). To address this problem, the original objects are mapped using a set of mathematical functions, known as kernels. The mapped objects will be linearly separable and, thus, instead of constructing the complex separation curve, all we have to do is to find an optimal line that can separate the classes. There are number of kernels that can be used in SVM models. These include linear, polynomial, RBF and sigmoid. The RBF is by far the most popular choice of kernel types. This is mainly because of their localized and finite responses across the entire range of the real x-axis. We used libSVM implementation, running under Weka [12].

4.2. Naive Credal Classifier 2 (NCC2)

The NCC2 is an extension of Naive Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with **small or incomplete data sets** [11]. Robustness is achieved by delivering set-valued classifications (that is, returning multiple classes) on the instances for which (i) the learning set is not informative enough to smooth the effect of choice of the prior density or (ii) the uncertainty arising from missing data prevents the reliable indication of a single class. As on small data sets Naive Bayes Classifiers (NBC) may return prior-dependent classifications, leading to fragile predictions, to deal with this problem, NCC2 specifies a set of prior densities, referred to as prior credal set; the credal set is then turned into a set of posteriors via element-wise application of Bayes rule [13].

We used the Java implementation, named JNCC2, released under the GNU GPL license and capable of processing ARFF format files.

4.3. AdaBoost

AdaBoost, a diminutive for Adaptive Boosting [14], is an algorithm for creating a "potent" binary classifier as linear combination of a simple one. Boosting decides the weak classifiers and their weights based on the minimizing of loss function in a two-class problem. Boosting is usually fast and has high performance. As an interesting property of Adaboost, we can mention the potential to reduce bias and variance from, for example, tree based classifiers.

For the present work we used the MultiBoostAB made available in Weka [12].

5. Classification experiment

This experiment evaluated the efficiency of the proposed system. Given a set of features extracted form EP speech frames, the system had to assign a class to each of the frames. The classes considered were: Nasal Vowel, Oral Vowel and Others.

5.1. Database

A small database was created consisting of 2 parts:

First part consisted on 3 minutes of speech from an EP native speaker reading random news. Recording took place in a normal office. Speech signal was recorded using 22050 Hz sample rate at 16 bit mono.

The second part consists on the speech recordings made during ElectroMagnetic Midsagittal Articulography (EMMA) acquisition for EP nasals. It is very rich in nasal sounds, particularly nasal vowels, and an example of noisy speech produced with articulations far from perfect. The database includes recordings from two native EP speakers(one male and one female). Two speaking rate conditions were recorded: normal and fast rate.

First part was manually annotated at segment level for all produced sounds; the second part was not fully annotated: only oral and nasal vowels segments and their context were contemplated.

Considering the frames (20 ms, no overlap), the database has the following distribution: 6836 Oral Vowels, 9763 Nasal Vowels and 43236 Others. If only the final 60% frames from nasal vowels are kept, the number of frames from nasal vowels decreases to 3377.

5.2. Metrics

For evaluation, we used 3 criteria: recall ratio (R), precision ratio (P) and F-ratio (F): $R = \frac{tp}{tp+fn}$, $P = \frac{tp}{tp+fp}$, $F = \frac{2RP}{R+P}$, being tp the number of true positives, fp the false positives, fn the false negatives.

As we are only interested on two of the three classes (Nasal and Oral vowels), values for R, P and F reported are averages of this metrics calculated separately for each of the two classes.

5.3. Results

5.3.1. Classifiers comparison

We started our experiments by comparing the three classifiers in a common setup, 10 fold cross validation. Besides classifier effect, two other factors were considered: the balanced number of examples for each class and the use or not of all the frames from the nasal vowels. This resulted in 4 evaluation scenarios. The inclusion of the second factor (discarding the initial frames of nasal nasal) was motivated by reports claiming that they have an initial oral (or oral like) phase. For NCC2 only determinate classification (only one class selected) is considered. The results are presented in Fig. 1.

From the figure is clear that the best results were obtained with the MultiBoost classifier. The SVM gave, in general, poor results. For all classifiers the best performance was obtained when combining the use for training of a balanced number of samples of each class and the inclusion of only the frames from the final 60% of the nasal vowels productions.

5.3.2. Evaluation of features subsets

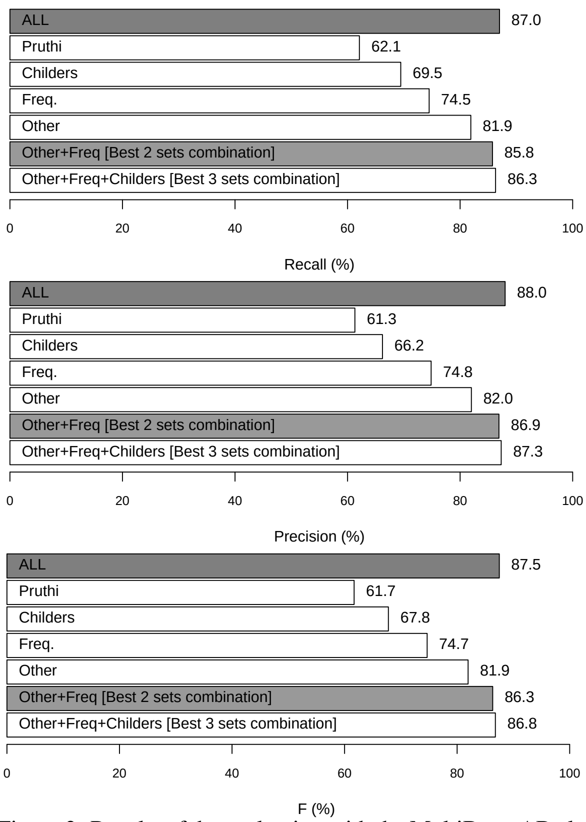


Figure 2: Results of the evaluation with the MultiBoostAB classifier on different subsets, and combinations of subsets, of features. In all cases the training was performed with equal number of examples for all classes and discarding initial part of nasal vowels production.

The results for evaluation on different subsets of the features (maintaining the results for ALL features as reference) are presented in Fig. 2. Results on each of the four subsets mention in 3 are complemented with the best results obtained on two and three subsets combinations.

The best results for a single subset were obtained with the "Others" subset, followed by the "Frequency" subset. Contrary

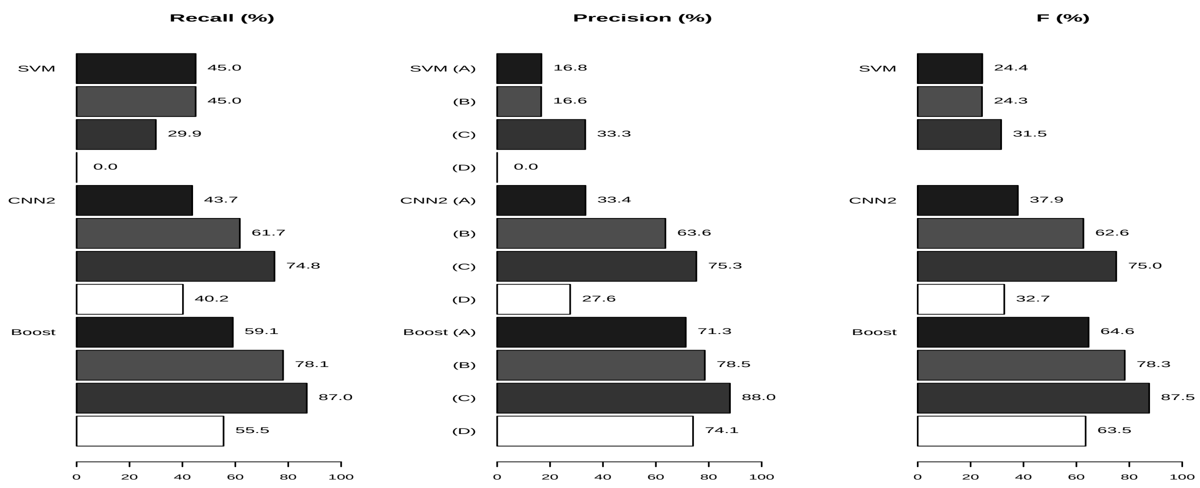


Figure 1: Results for Recall, Precision and F-measure for the 3 classifiers when evaluated with all the features. The following test variants were considered: all frames used (unbalanced number of samples for each class), named A; number of samples equal for the 3 classes (B); balanced with only the final 60 % frames from nasal vowels (C); unbalanced version of the previous (D).

to our expectations, the worst results were for the "Pruthi" subset. As best results with only one subset is more than 5 % lower than using all features, combinations were also evaluated. The two best subsets work well when combined and constitute the best combination of two subsets (Others+Frequency). The best 3 subsets combination is also formed by the classifiers on the first 3 places when used alone. With these subsets the performance is quite close to the one obtained with all features, interestingly not including features of the "Pruthi" subset.

6. Conclusion

In this paper we investigated the usefulness of three classifiers and several subsets features in the detection and classification of speech frames from productions of EP oral and nasal vowels. The classic SVM classifier was compared against the NCC2 and MultiBoost classifiers. A small data set was used for classifiers and features evaluation. Results point to the better performance of MultiBoost in our experimental conditions. Also, the use in the training process of equal number of examples for all the classes and discarding frames from the initial parts of nasals vowels contributes to better performances.

Overall, the best results obtained are inside the interval reported in [5] for American English. But, contrary to [5], we didn't achieve good results with the SVM classifier.

Future research will include: experiments with other classifiers , features selection, adding a 3rd level of classification to individualize the vowels, experiments on classifiers output fusion and, our main goal, use of the detected frames in a phonetic-based speaker verification system for EP.

7. Aknowledgments

First author work was made possible by a scholarship from the Portuguese Science and Technology Foundation (PhD grant SFRH/BD/18230/2004).

8. References

[1] M. Sharma and R. J. Mammone, ""blind" speech segmentation: automatic segmentation of speech without linguistic knowledge," in *ICSLP*, 1996, pp. 1237–1240.

[2] A. Kanae, S. Tostomu, and A. Takayuki, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoust Sci Technol*, vol. 27, no. 4, pp. 233–235, 2006.

[3] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," *Speech Communication*, vol. 43, pp. 225–239, 2004.

[4] T. Pruthi and C. Y. Espy-wilson, "Acoustic parameters for the automatic detection of vowel nasalization," in *INTERSPEECH 2007*, Antwer, Belgium, August 2007.

[5] T. Pruthi, "Analisys, vocal-tract modeling and automatic detection of vowel nasa-lization." Ph.D. dissertation, University of Marylans, UMI, 2007.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 195, ch. 14, pp. 495–516.

[7] D. G. Childers, *Speech processing and synthesis toolboxes*. John Wiley & Sons, Inc., 2000.

[8] D. A. Cairns, J. H. L. Hansen, and J. E. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 35–45, 1996.

[9] M. K. Huffman, "The role of F1 amplitude in producing nasal percepts," *Journal of the Acoustical Society of America*, vol. 88, no. S1, 1990.

[10] J. Glass, "Nasal consonants and nasalized vowels an acoustic study and recognition experiment," Master's thesis, MIT, Cambridge, USA, 1984.

[11] G. Coranin and M. Zaffalon, "Learning reliable classifiers from small or incomplete data sets: The Naive Credal Classifier 2," *Journal of Machine Learning Research*, vol. 9, pp. 581–621, 2008.

[12] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, 2005.

[13] G. Coranin and M. Zaffalon, "JNCC2: The java implementation of naive credal classifier 2," *Journal of Machine Learning Research*, vol. 9, pp. 2695–2698, 2008.

[14] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," 1995.

Dispersion of Vocal-Fold Biomechanical-Parameter Estimates

Pedro Gómez, Roberto Fernández-Baíllo, Víctor Nieto, Victoria Rodellar, Luis M. Mazaira, Cristina Muñoz, Agustín Álvarez, Rafael Martínez

Grupo de Informática Aplicada al Procesado de Señal e Imagen
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
pedro@pino.datsi.fi.upm.es

Abstract

Modeling the vocal fold biomechanical system is relevant for several fields in speech and voice studies, as in voice production, natural speech synthesis, emotion estimation or voice pathology detection. The key stone to good phonation models is the availability of reliable estimates. An indirect method for the estimation of the biomechanical parameters of the voice production model is presented. The availability of previously-normalized voice databases for pathology studies allows the evaluation of the methodology, and the establishment of the distribution profiles of the parameters under study. The paper illustrates the validation process and the descriptive statistics of the biomechanical parameters.

Index Terms: voice production, speech synthesis, speech biomechanics, gender-sensitive parameter-distributions

1. Introduction

During the last seven years interest in the Liljencrants-Fant voice-production source-filter model [1] has been retaken in the sense of establishing better models and estimation methods of the glottal excitation. For long this pattern had been referred as a 1/f power spectrum signal considering it as a useless signal with nothing else to offer. Nevertheless since the pioneering work of P. Alku [2] contributing to better estimate the glottal source by recursive inverse filtering, a growing interest in this specific signal has contributed to a hatching of different studies, as in the measurement of the open-close cycle, or the power spectral profile [3], the correct reconstruction of its causal and anti-causal components [4] or the use of its spectral profile singularities in pathology detection [5]. What is intended in the present work is to derive estimations for the most relevant biomechanical parameters, as dynamic masses and elastic ratios, these being especially important in model building as well as in pathology studies or in voice education and rehabilitation, as well as in natural speech synthesis. The paper is divided into five main sections besides the present one. The next section is devoted to present the biomechanical model used in the study. Section 3 is intended to explain the indirect parameter estimation methodology. Through section 4 the statistical description of the estimates is given for four different groups of 50 speakers each: male and female normophonic and dysphonic, respectively. Section 5 presents and analyzes the results obtained for the study sets. Conclusions are presented in section 6.

2. Source-Filter Production Model

The Voice Production Model is depicted in Figure 1 For phonated speech (voiced) the lungs inject a flow of air through the vocal folds to the pharynx, nose and larynx. The vocal folds (presented in cross-section) vibrate under the action of the forces exerted by the differential pressure between the subglottal and supraglottal sides.

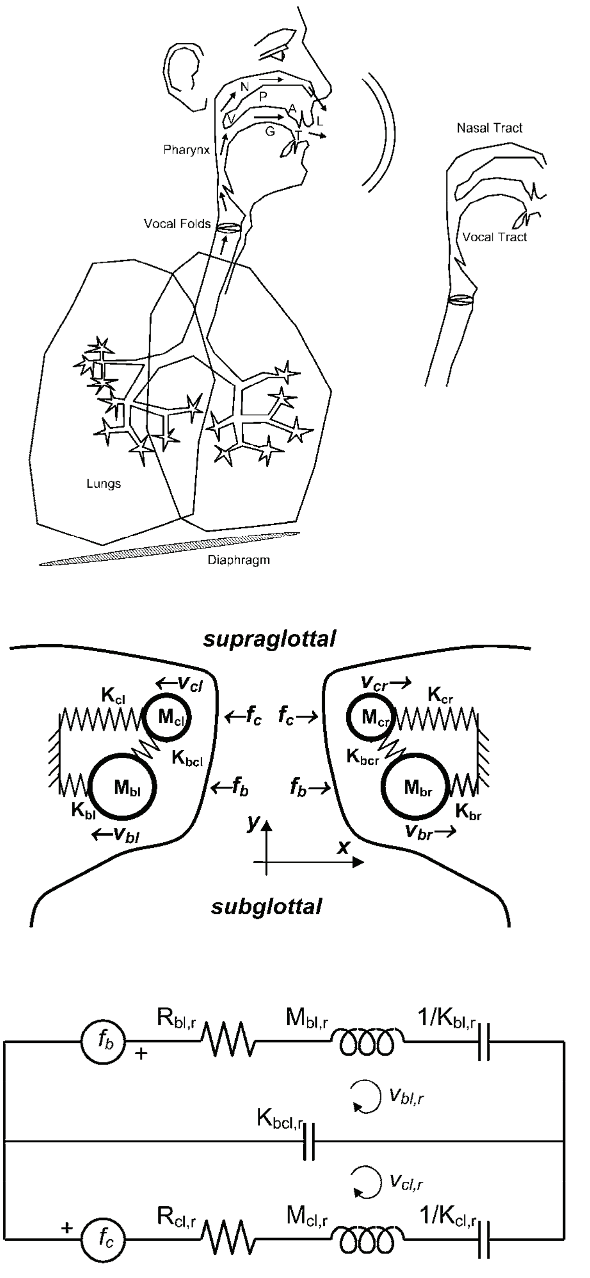


Figure 1 Top: The Voice Production System. Middle: Biomechanical 2-mass model of the vocal folds. Bottom: Electromechanical equivalent.

The system of body (b) and cover (c) masses ($M_{b,cl,r}$) of the left (l) and right (r) vocal folds move against the elastic elements represented by the springs ($K_{b,cl,r}$). The system behaves as the electromechanical equivalent in the bottom section of the figure. This linear model can not cope with nonlinear effects present in the vibration, but is accurate enough to represent the overall behavior of the glottal source (pressure wave exciting the vocal tract as a results of vibration) in the frequency domain (power spectral density). Having in mind that this is an over-simplified model, it must be stressed that its main interest is to be found in the relative feasibility of its inversion by numeric methods for most of the voiced segments of interest in speech studies. The articulation acoustic model is presented in Figure 2 showing the models representing the glottal (voicing) and turbulent (unvoicing) excitations, the articulatory organs (vocal tract) which could be summarized in a transfer function $F_v(z)$ and the radiation effects. is the transfer function of the vocal tract and $g_s(n)$ is the glottal excitation (glottal source) during the voicing segments of speech (unvoiced segments are not considered here as the study is exclusively concentrated in voicing). Classically the glottal source is conceived as a pressure wave resulting from the opening and closing of the vocal folds following the natural vibration of a pair of masses linked elastic tissues to the walls of the larynx. To reconstruct the glottal excitation the influence of the vocal tract has to be removed from voice following the system in Figure 2

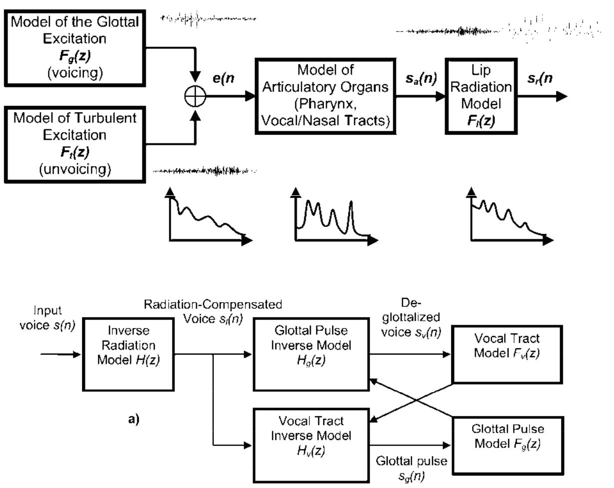


Figure 2 Top: Source-Filter model of Gunnar Fant (see text) widely used in acoustic phonetic studies. The glottal excitation (top trace) enters the chain of tubes modulated by the articulatory organs (vocal tract) and is radiated as phonated speech. If the excitation is not harmonic (bottom trace) the result is unvoiced speech. Bottom: Iterative extraction of the glottal pulse. The voice input is compensated in radiation. A first estimate of the glottal pulse is inverted and used to remove the glottal influence in the input. The resulting de-glottalized voice is modelled to extract the vocal tract model $F_v(z)$, which is used to remove the vocal tract influence on the input voice, giving birth to the glottal source estimation. This signal is modelled to be removed from the input voice by $H_g(z)$, in a cris-crossing iteration which is repeated 2-3 times.

The inversion system is divided in a section to compensate lip-radiation, a second block to model and remove the inverse vocal tract transfer function $H_v(z)$ and a third block to reconstruct the glottal source $g_s(n)$ from the residual signal left at the output of $H_v(z)$: $r_s(n)$. Usually this structure is refurbished as a recursion, as once an estimate of the glottal source is available its inverse sequence may be used to remove the influence of the glottal source in the radiation-compensated speech, thus producing a speech signal which is easier to invert to estimate a good vocal-tract inverse. All these structures: lip-radiation, vocal tract and glottal source cancellers may be implemented as lpc lattices [6]. The main problem now is how to better determine their respective filter orders. Classically lip-radiation cancellers can be first-order lattices. The issue of vocal tract modeling is a little bit more complex, as many possibilities are at hand, the two most popular ones being to select the filter order in the order of the sampling frequency f_s divided by 1,000. This may be enough for the purposes of voice coding, but for an accurate glottal source reconstruction for pathology detection, the adequate order has been established in around twice this order [5].

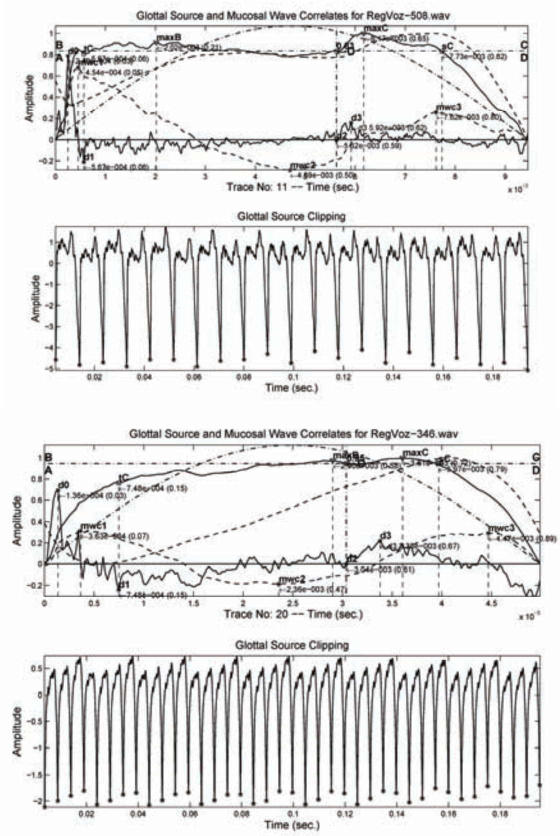


Figure 3 Typical glottal source excitations (thick filled trace) and clipped frames. The spikes mark the instant of closure. Male prototype (top). Female prototype (bottom).

The issue of better modeling the glottal source is a complex one. A first approach would use an order-one or two lattice to reproduce the inverse behavior of the glottal source, but it must be considered that such signal, as presented in Figure 3 has two specific parts: a closed segment which is minimum phase, and an opening segment which is not minimum phase [4]. Therefore Linear Predictive methods will lack precision when reconstructing this last segment. To have this effect into account the order of the lattice modeling the inverse of the glottal source should be at least order-two. Larger orders are not advisable as the zeroes of the glottal source cancellers could lock to the largest poles of the vocal tract. This would result in cross-talk between the estimates of the vocal tract and the glottal source. Reasonable results can be obtained with $K_r=1$, $K_s=2$, $K_v=32$ for a sampling frequency of 16,000 Hz, these being the respective orders of the radiation, source and tract inverse cancellers. The first part of the excitation in Figure 3 is a recovery to the average neutral situation (top horizontal line). The start of the opening phase is marked by the thin vertical line, approx. in 4.7 msec for the male case, and at 2.4 msec for the female one. Relative opening happens earlier in the phonation cycle of normophonic female subjects. The maximum opening is marked by the maximum amplitude in the pulse after which the closing phase starts, which is complete at the end of the cycle. Other reference traces as the average acoustic wave, the mucosal wave, its first derivative, and the glottal flow are represented as well.

3. Biomechanical parameter estimation

The key point to reconstruct the glottal source is to use the glottal residual after the removal of the vocal tract by $H_v(z)$. The resulting signal $g_r(n)$ can be considered as the first derivative of the glottal source $g_s(n)$, therefore the glottal source may be obtained by direct integration of the residual. Once the glottal signal is reconstructed its power spectral density (psd) may be obtained by taking the pitch-synchronous modulus of the glottal source Fourier Transform cycle by cycle. The behavior of this psd can be seen in Figure 4and is that of a general *1/f* decay function with specific peaks and troughs which are more noticeable in the 0-2 kHz interval. It may be shown that these have to see with the resonances and anti-resonances of the electromechanical equivalent of the vocal folds model [7]. Therefore a direct relationship between the main peak in the glottal source psd and its gentle *1/f* decay may be established with respect to the mass and elasticity parameters of a biomechanical model as the one in Figure 1.

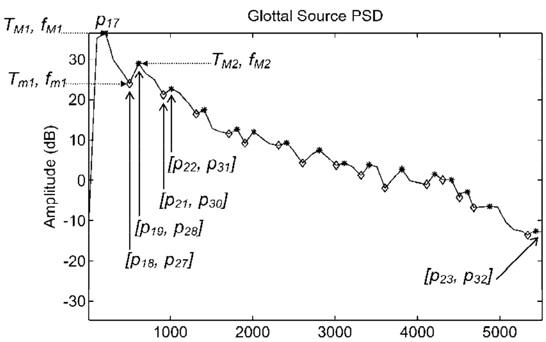


Figure 4 Power spectral density of the glottal source for a phonation cycle. The singularities of the spectral profile are estimated in terms to their amplitude and position relative to the main peak (T_{M1}, f_{M1})

The process of biomechanical parameter estimation from the glottal source power spectral density is covered by the following steps

$$M_{bl,r} = \frac{\omega_2}{\omega_2^2 - \omega_r^2}\left[\frac{T_r - T_2}{T_r T_2}\right]^{1/2} \tag{1}$$

ω_r being the resonance frequency given by

$$\omega_r^2 = \frac{K_{bl,r}}{M_{bl,r}} \tag{2}$$

where the square modulus of the psd is given by

$$T(\omega) = \frac{1}{\left[R_{bl,r}^2 + \varpi^2 M_{bl,r}^2\right]^2} \tag{3}$$

with the frequency relative to the resonance point (maximum)

$$\varpi = \frac{\omega^2 - \omega_r^2}{\omega} \tag{4}$$

and

$$T_r = T(\omega = \omega_r) = \frac{1}{R_{bl,r}^2} \qquad T_2 = T(\omega = 2\omega_r) \tag{5}$$

The estimation procedure must detect the value of pitch, which is used to evaluate ω_r. The determination of T_r and T_2 is carried out on the power spectral density of the glottal source spectral profile. This leads to the determination of the losses from (5) and to the mass (1) and stiffness (2). $M_{bl,r}$ are the equivalent (dynamic average) masses of the vocal fold body, and $K_{bl,r}$ are the (dynamic average) elastic coefficients of the vocal fold (body), estimated pitch-synchronously at the k-th phonation cycle.

4. Validation: materials and methods

For the validation of the methodology a data set of 100 normophonic speakers of both genders have been selected from a previously-tested database used in pathology studies [8]. These speakers have been inspected by endoscopy to discard organic alterations of the vocal folds, their phonation cycle has been inspected using stroboscopic illumination to check its apparent normality, have been GRBAS graded [9] and automatically checked by an inspection tool using acoustic analysis (classical distortion measurements as jitter, shimmer, or harmonics-to-noise ratio: HNR) [5]. As a control a set of 100 dysphonic speakers of both genders have been used for contrast. The inspection methodology is the following:

- 200 msec. of phonations of the vowel /ah/ have been selected from each speaker
- The glottal source is reconstructed for each speaker following Figure 2.
- Phonation cycles are being detected, which for male speakers yield some 20 arches, this being about twice for the typical female speaker.
- The glottal source psd is estimated for each cycle.
- Jitter, shimmer and HNR are evaluated in a per-cycle basis.
- The first, second and third peaks in the glottal source psd profile are detected, as well as the two first troughs.
- The average dynamic mass, elasticity and loss are estimated for each cycle.
- The unbalances between neighbor phonation cycles for each parameter are also evaluated.

For inspection purposes the centroids of the respective four clusters are evaluated (MN: male normophonic; MD: male dysphonic; FN: female normophonic; FD: female dysphonic) and the respective Mahalanobis distances of each sample to the gender-respective normophonic centroids are estimated:

$$D_{ti} = \left[(y_{ti} - \mu_M)^T C_M^{-1} (y_{ti} - \mu_M)\right]^{1/2} \tag{6}$$

Where μ_M and C_M are the centroid and the Covariance matrix of the Model Sets (normophonic subjects). Figure 5 shows the respective histograms of the Mahalanobis distance of each subject respective to its respective gender normophonic cluster.

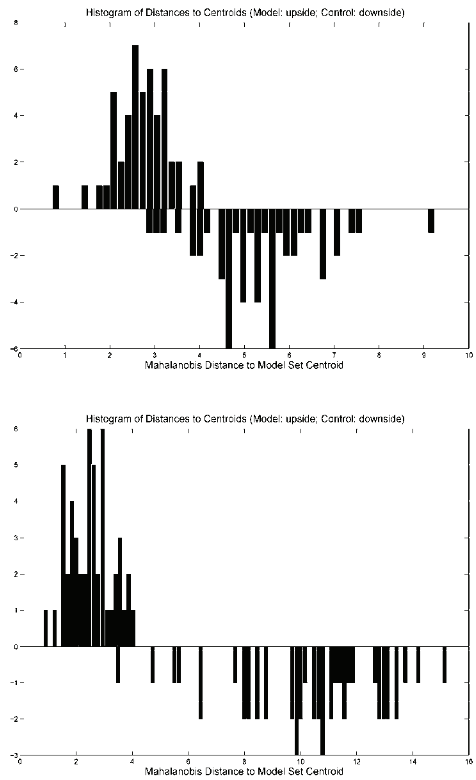


Figure 5 Top: Histograms of Mahalanobis distances of normophonic (upside) and dysphonic (downside) male subjects to normophonic centroid (model set). Bottom: Id. for female subjects.

It may be seen that both genders are clearly differentiated according to the statistical distributions of their parameters.

5. Results and discussion

The statistical description of each parameter distribution is given as well in Table 1 in terms of means and standard deviations.

Table 1. Statistical description of the biomechanical parameter estimates. MN: Male Normophonics; MD: Male Dysphonics; FN: Female Normophonics; FD: Female Dysphonics

Group	Ave. Pitch (Hz)	Std. Pitch (Hz)	Ave. M_b (mg)	Std. M_b (mg)	Ave. K_b (dyn.cm^{-1})	Std. K_b (dyn.cm^{-1})
MN	113.8	11.8	21.7	2.2	11101	1101
MD	113.4	22.8	22.0	4.2	11251	2347
FN	203.4	19.6	12.0	1.2	19682	1897
FD	189.0	40.3	14.0	4.0	19397	8690

The first interesting result is that the average pitch is almost the same for the normophonic and dysphonic male sets, but the dispersion is almost twice for the dysphonic set. Concerning the female sets the pitch is larger for normophonics than for dysphonics, and the dispersion is clearly larger for dysphonic than for normophonic. This result is consistent with classical assumptions. Regarding the fold body dynamic mass, this is slightly larger than 20 mg both for normophonic and dysphonic male subjects, although the dispersion is again larger in this last case. For female subjects the average mass estimates are 12 and 14 mg, which is consistent with what was expected from anatomic expectations. The dispersion in this case is much larger for female than for male dysphonics, which is also consistent with classical assumptions. Regarding elastic parameter estimates, again the difference between male normophonics and dysphonics is not significant as far as averages are concerned, dysphonics being twice more disperse. In the case of females, the elastic parameter is larger, which implies a tighter vocal fold, as expected, the dispersion being also much larger in dysphonics than found in males.

6. Conclusions

The methodology used for the determination of the biomechanical parameters for both genders is consistent with classical assumptions: male speakers present average dynamic body masses almost twice those of the female group, whereas the elastic tension is larger in female than in male in a proportion of around 80%. Dispersion is larger in the female group in general. The conclusion is that the indirect estimations are in good agreement with both gender characteristics. The contrast of indirect measurements with direct evaluation is not possible in living subjects, the possibility of doing this same study with excised larynges being out of the scope of the present research. The results give interesting hints on plausible values to be used in 1-mass biomechanical models, the methodology being extensible to 2-mass models as well. These models are especially interesting for voice synthesis with more realistic glottal excitations in voice synthesis. This is also the case for artificial larynges after radical laryngectomy. Other application field is the biometric description of the speaker.

7. Acknowledgements

This work has been funded by grants TIC2003-08756, TEC2006-12887-C02-01/02 and TEC2009-14123-C04-03 from Plan Nacional de I+D+i, Ministry of Science and Technology, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http.//www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

8. References

[1] Fant, G., et al., "A four-parameter model of glottal flow", STL-QSPR 4 (1985) 1- 13. Reprinted in *Speech Acoustics and Phonetics: Selected Writings*, G. Fant, Kluwer Academic Publishers, Dordrecht, 95-108, 2004.

[2] Alku, P., "An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulseform", Proc. of the ICASSP'92, II/29-32, 1992.

[3] Orr, R. et al., "An investigation of the parameters derived from the inverse filtering of flow and microphone signals", Proc. of VOQUAL'03, Geneva, 35-40, 2003.

[4] Doval, et al., " The voice source as a causal/anticausal linear filter", Proc. of VOQUAL03, 15-19, 2003.

[5] Gómez. P. et al., " Glottal Source Biometrical Signature for Voice Pathology Detection", Speech Communication, (51): 759-781, 2009.

[6] Deller, J. R., et al., Discrete-Time Processing of Speech Signals, Macmillan, NY, 1993.

[7] Gómez, P., et al., " Biomechanical Parameter Fingerprint in the Mucosal Wave Power Spectral Density", Proc. of the ICSLP'04, 2004, pp. 842-845.

[8] Project MAPACI: http://www.mapaci.com.

[9] M. Hirano et al., Acoustic analysis of pathological voice. Some results of clinical application, Acta Otolaryngologica 105 (5-6): 432-438, 1988.

A Multi-lingual TN/ITN Framework for Speech Technology

Hyongsil Cho[1,2], *Daniela Braga*[1,2], *Cristiano Chesi*[1,2], *Daan Baldewijns*[1], *Manuel Ribeiro*[1], *Kaisa Saarinen*[1], *Jeppe Beck*[1], *Silvia Rustullet*[1], *Peter Henriksson*[1], *Miguel Dias*[1,2], *Heiko Rahmel*[3]

[1] Microsoft Language Development Center, Portugal
[2] ISCTE-Lisbon University Institute, Portugal
[3] Microsoft Speech Components Group, Redmond, USA

{v-hych, i-dbraga, v-crches, v-daanb, i-manrib, v-kasaa, v-jeppeb, i-sirust, Miguel.Dias, heikora}@microsoft.com

Abstract

Although the research community pays little attention to (Inverse) Text Normalization (TN and ITN), this is an essential module in Text-to-Speech (TTS) and Speech Recognition (SR) systems. It has a significant development timeline and requires deep linguistic expertise. One of the main issues is ambiguity resolution, which is particularly problematic when handling numerals in different languages, especially those with gender or case variation. In this paper, we present a framework that can deal simultaneously with TN and ITN and which was applied to twelve different languages. The rules were tested and subsequently refined. The overall performance of the system is presented and discussed.

Index Terms: (inverse) text normalization, text-to-speech, automatic speech recognition

1. Introduction

Text normalization (TN) is a core module in any speech synthesis or Text-to-Speech (TTS) system. As a part of text pre-processor, the text normalization module converts a raw text file, which contains non-standard words such as sequences of digital bits or abbreviations, into a well-defined sequence of linguistically-meaningful units. The same module can be also used by speech recognition systems (SR), where the module performs the same task but in the opposite way and the operation is called then Inverse Text Normalization (ITN).

Text pre-processing is an essential part of any NLP system, since the characters, words, and sentences identified at this stage are the fundamental units passed to all further processing stages, from analysis and tagging components, such as morphological analyzers and part-of-speech taggers, through applications, such as information retrieval and machine translation systems [2].

The usual approach for this is to use phonetic lexica, i.e., a list of entries and their respective orthographic expansion or immediate phonetic transcription. However, a list-based approach has a number of important drawbacks.

First, a list-based approach is linear. So, to include all numbers up to a million, the same number of list entries would be required. More complex structures as dates, times, currency units, mathematical expressions and telephone numbers – which often have many possible orthographic representations – are even more problematic to cover exhaustively.

A second issue is the problem of ambiguity for TN. A non rule-based approach hardly ever offers descriptions or solutions for the different cases of ambiguities in this module, and for which disambiguation rules need to be proposed. For example, while a native language speaker of Portuguese language might find it relatively easy to decide whether *SPA* should be read as an acronym and therefore spelled out as *Sociedade Portuguesa de Autores*, or as an acronym and pronounced as a single syllable when it means *Salutem per Acquam*, this is not true of a computer application, which requires context information and rules to make such a decision. The same happens when with Roman numerals, such as *I, V, X, C, D* or *M,* which are written using letters and therefore cause ambiguities with acronyms. For example, a sequence like *D.* occurring in a text written in Portuguese may mean not only the number *500* but also *Dom* or *Dona*, depending on the context

This problem is also related to the last disadvantage of list-based text normalize. A simple list, when performing TN, is unable to resolve agreement between the item to be normalized and its context. In Portuguese, the number 1 can have a masculine or feminine readout, depending on the context; and for highly inflected languages like Finnish which has a system of 15 cases, the chance of having this kind of problem is even bigger.

In this study, we created a rule-based multi-lingual TN/ITN algorithm and tested its performance over a number of languages from different families.

2. Domains and issues of text normalization

2.1. Domains of text normalization in Speech Synthesis and Recognition Systems

The size of a text normalization module varies depending on its applications. A number of studies have defined objects of text normalization and identified characters of each object.

Some previous works investigated the problems of named entity recognition in informally inputted texts and proposed improving the performance of personal name recognition in emails using two machine-learning based methods [6][1]. For email data cleaning, a cascaded approach by employing Support Vector Machines and rules has been proposed in [10]. While some studies pay attention to the case restoration problem [4][3][5], others concentrated to the normalization of non-standard words in texts, including numbers, abbreviations, dates, currency amounts, and acronyms. They propose taxonomy of non-standard words and apply n-gram models, decision trees and weighted finite-state transducers to the normalization [8].

Our approach to the text normalization focuses on the non-standard words normalization. To do this, we first defined the categories of non-standard words to be normalized by our text normalization module. Each of the categories can be described by its example as follow:

- Cardinal (positive/negative numerals)
  - Integers of range 1 - 999: 2, -59
  - Integers of range 1,000 - 10,000: 5000, 3,589 or 3589
  - Large integers : 301,020 or 301020
  - Decimals : 0.1, 0.03, 0.987
- Ordinal: 1st, 2nd
- Roman Numbers: I, II, V
- Date
  - Date expression by digits: 5/10/2000
  - Stand alone year: 1989, '89
  - Day and month: 5/10, 5-10
  - Month and year: 10/2000, 10.2000
  - Decade: 60', 70'
- Time : *time expression with a variety of signs* (:, -, a.m., p.m., etc)
- Fraction: ½, ¼, ¾
- Telephone Number
  - Local numbers: 555-1212
  - Local numbers with area code: (267) 555-1212
  - Local numbers with country code: +1-212-735-0989
  - International dialing: +31 24 323 5647
- Measurement: 20l, 300km, $5m^2$
- Degree: 480°
- Address
  - Street address line 1: *Street number and name*
  - Street address line 2: *Name of town, State (abbreviation)* + *ZIP code*
  - Optional street address line: *any optional information*
  - Post office box: P.O.BOX 1320
- Currency
  - ISO Standard: 200$, $200
  - Subdivision of currencies: 1.75$
- Contact Names
  - (Honorific) Titles: Mr., Mrs., Dr.
- URL/e-mail address: abc@12f3.net
- Range expression for numerals and date: 1~20, 5/10/2000~10/10/2000

2.2. Language specific issues

To normalize correctly all of the categories described above, a number of language specific issues must be considered.
For example, in some West Germanic languages, numerals between 11 and 99 need a special attention for their specific word order. The following example illustrates the case of the Dutch language:

Ex) 23: drieëntwintig (3+and+20 in one word)
44: vierenveertig (4+and+40 in one word)

Moreover, in case of the French language, the expression of some numbers is made in a very particular way:

Ex)10 : dix
60: soixante
70: soixante-dix (60+10)
80 : quatre-vingts (4 times of 20)

Also, as described briefly in the section 1, the gender distinction is very important for Romance languages and a part of Germanic languages.
The forms and the expression of numerals get a bigger diversity when it comes to the telephone number: the number of units in a standard telephone number is different from a country to another and the way of reading a telephone number is also highly varied depending on the country.
For example, a French standard telephone number is composed by 5 times of two unit numbers and all of them are read as tens.

Ex) 01 91 28 64 32
: zéro un quatre-vingt-onze vingt-huit soixante-quatre trente-deux

Whereas a Korean telephone number must be read digit-by-digit.
Ex) 361-2839
: 삼(3)육(6)일(1)에(-) 이(2)팔(8)삼(3)구(9)

In some languages, text normalization modules need to consider the phonetic context of items. For instance, in Italian, the common abbreviation for *Saint* (i.e. S.) should be normalized not only according to the gender of the saint's name:

Ex) S. Marco: San Marco ("Marco" is masculine)
S. Maria: Santa Maria ("Maria" is feminine)

but also keeping into account if the name starts by vowels or not:

Ex) S. Antonio : Sant'Antonio
("Antonio" is masculine and begins with a vowel)
S. Anna: Sant'Anna
("Anna" is feminine and begins with a vowel)

Concerning to the context, another critical issue is that ambiguities in normalization are not always locally solvable, that is, it is not enough to look at the adjacent words to guess the correct normalization. In that sense, the item *S.*, mentioned in the previous examples, can be a hard element to handle.

Ex) S.S. Appia 7: Strada Statale Appia sette

In this example, without a deep understanding of the word which follows *S.S.*, a text normalization module may make the fatal mistake to get *Santissima Appia sette* or *Santo Santa Appia sette* instead of *Strada Statale Appia sette*.
To handle all these issues, we built a rule-based TN/ITN module with possibilities of implementing and controlling the information about context.

3. A Rule-based multi-lingual TN/ITN framework

3.1. Structure and hierarchy of the rules

We built, in 2.1, a list of domains of non-standard words. In our TN/ITN module, these domains will be associated with top-level rules, with each one being basically independent. Although the rules are self-referent and the domains intercross, the fact that they are top-level rules enables them to be independent without being influenced by the behavior of other rules (except for cases where there are identical inputs.

Text normalization rules are structured based on a language called TNML (Text Normalization Mark-up Language), which is an adaptation of SSML (Speech Synthesis Markup Language) for the TN module. SSML has been recommended by the W3C (World Wide Web Consortium) since 2004 and is currently one of the most commonly used mark-up languages.
However, while the language is easy to understand, the sheer number of rules and constant referencing makes the process too intricate to be handled in a text editor. For this reason, we decided to make easier the process of writing down the text normalization rules by creating an internal tool that establishes a bridge between the rules and the language in which they are written. The TNML programming language and testing process presented in this paper are protected by patent (Patent Serial No. 12/361,114).

Using this tool, we can create a TNML file that we call a "map". The structure of a TN map is basically a tree-type scheme, with successive function-based references and positions. Four types of rules have been defined in TNML: terminals, sequence rules, list rules and top level rules. Thus, a map is composed by one or more top level rules and a top level rule possess one or more sequence rules, list rules and/or terminals. The characteristics of each rule are described in the following sections.

3.1.1. *Terminals*

Terminals are the elements placed at the bottom in the rules tree of a text normalization map. Terminals contain the information to be used during TN/ITN process and establish a one-to-one relation between a display form and a spoken form. For example, in case of the Portuguese language (see 1), we put the information that the numeral '1' must be normalized as *um* or *uma* is contained on the terminal level by making two different terminals, one with *um* and the other with *uma*, for one same item '1'.
In French language's case, some items may have a bigger number of terminals.

Display	Spoken
4	Avril
4	quarante
4	quart
4	quatorze
4	quatre

Table 1: *Terminals related to the item '4' in French*

All the other levels of rules are authored based on the terminals. In a rule which normalize numerals 0~9, the terminal 4⇔quatre will be used whereas a rule to normalize the date, especially the month, will need the terminal 4⇔Avril.

3.1.2. *List rules*

List rules represent their references in a vertical way. Their main characteristic is that they allow for only of the different available normalizations in the list. This allows for a grouping of the elements in a way that they can be reused or referenced by other rules where only one of the elements is selected. For example, the list rule shown in the Table 2 regroups a number of terminals of English language which may function in the same way.

List rule name	Composition (Terminals)	
	Display	Spoken
Cardinal 1 to 9	1	one
	2	two
	…	…
	8	eight
	9	nine

Table 2: *A list rule*

A list rule can be a direct constituent of a top level rule, a sequence rule or another list rule.

3.1.3. *Sequence rules*

Differently from the list rules which regroup terminals and rules of a same class, the sequence rules concatenate terminals rules by defining their order within a sequence. Given a terminal 3⇔thirty and the list rule shown in the Table 2, we can create a sequence rule which will normalize all the numerals between 31 and 39.

Sequence rule name	Composition	Example
Cardinal 31 to 39	(3⇔thirty)+(Cardinal 1 to 9)	31⇔ thirty one … 39⇔thirty nine

Table 3: *A sequence rule*

A list rule can be a direct constituent of a top level rule, a list rule or another sequence rule.

3.1.4. *Top level rules*

Top-level rules are the elements placed at the top in the rules tree of a text normalization map. All other rules run towards the top-level rules, which are the entry gate to a TN map. When starting a normalization process, the system always starts off by using a top-level rule. In other words, any terminal, list rule or sequence rule which is not included in a top level rule will not make any effect in the normalization. As such, all rules must directly or indirectly be associated with a top-level rule.

In our TN/ITN module, we built one top level rule per category (defined in 2.1).

3.1.5. *Other components*

In addition to the four types of rules described above, our TNML maps may contain some supplementary information such as:

- Spaces control between the constituents of a sequence rule. They can be toggled on or off.

Name	Display	Spoken
SpcToSpc	<sp>	<sp>
SpcToNo	<sp>	<ns>
NoToSpc	<ns>	<sp>
NoToNo	<ns>	<ns>

Table 4: *Space control*

In fact, in the sequence rule shown in Table 3, NoToSpc is applied between <3⇔thirty> and <Cardinal 1 to 9> and that's how the rule could produce <31⇔ thirty one> and not <31⇔ thirtyone>.

- Priorities: Assuming a map for TN (used in speech synthesis) and ITN (used in speech recognition), priorities allow for assigning values to disambiguate identical expressions, by using a system of weight defined for each constituent of a list rule.

- Agreement: Agreements make it possible for the TN rules to bridge into an annotated lexicon and extract information from it to facilitate disambiguation in specific cases.

3.2. TN/ITN framework applied to 12 languages

The performance of our TN/ITN framework was tested for a number of languages from different language families. The languages are en-GB, de-DE, fr-FR, it-IT, pt-PT, ca-ES, es-ES, nb-NO, da-DK, nl-NL, sv-SE and ko-KR.

To test the performance, a text normalization map was created. During the whole process of building and re-fining the text normalization map, 3 types of tests were carried out:

- Accuracy tests in internal tool
- Performance tests on a large sized text corpus
- Overall tests on a set of pre-selected items

First, the accuracy tests in internal tool were performed during the coding and fine-tuning of rules. These are simple tests (input/output) normally associated with a top level rule. The purpose of these preliminary tests is to assess the functionality of the rules as they are being created, and to fix any small concatenation and referencing errors, such as extra or missing spaces, agreement issues, reference or structure errors, etc.

Although the number might vary depending on the rule, a minimum number of nine cases have been established for each top-level rule. A few other cases were added to sub-rules in order to test their efficiency during the rule-writing process. The intended mark for these preliminary tests is a 100% accuracy rate.

Second, the performance tests on a large sized text corpus assume the existence of a beta version of the text normalization map, with all domains finalized and preliminary tests completed. During these tests, TN/ITN rules are applied on corpus composed by 50,000~100,000 sentences, depending on the language and the nature of corpus, collected from various sources in order to obtain and analyze approximately 20,000 normalized items. For example, to test the performance of European Portuguese map, a corpus of approximately 60,000 random phrases was used.

The goal of these tests is that our map recognizes any non-standard word occurring in the raw text data and normalizes it correctly. After analyzing the results, the rules are reviewed again in order to add unexpected patterns and fix any errors found. By analyzing the errors, we could observe how our TN/ITN map works when it is combined with the other modules of a TTS or SR system.

To finish, overall tests on a set of pre-selected items were made. A text data set composed by approximately 1,000 non-standard words (distributed over the all domains defined in 2.1) was built. Differently from the corpus used for the previous performance tests, the data set built for the overall tests doesn't contain any standard shape of sentence but only non-standard words like 11/5/1989 for date or 378-2684 for telephone number. Each of those items has the (pre-defined) expected spoken form and the goal of the overall test is i) our TN/ITN map of the given language places a given item in the right category, ii) the rules applied to the item generate the correct expected spoken form for the given item and iii) the map works also correctly in the ITN direction (from the spoken form to the display form).

After having run a first normalization, the errors were categorized, analyzed and fixed, and a final hit rate of a 100% has been achieved.

3.3. Result and discussion

For a text normalization module, it is a very hard task to get a 100% accuracy rate on an infinitely large corpus composed from various sources. There will always be word sequences or characters that contain unexpected problems for a phonemic conversion algorithm. Therefore, building and maintaining a text normalization module is inevitably work in progress and will always be a cyclic process.

In this sense, our rule-based multi-lingual TN/ITN framework gave a reasonable result by getting a 100% success rate in the overall tests on preselected items. The result leads us to conclude that the rule-based approach provides solutions to ambiguities and contextual agreement, and allows a much higher level of exhaustiveness. However, it must be realized that reviewing and modifying the maps and the whole framework will remain necessary.

4. Conclusion

We built a rule-based multi lingual TN/ITN framework, in which maps for a number of languages were authored by a group of language experts. The performance tests of each language's map provided satisfactory results.

In the future, improvement of our module and enhanced accuracy are sought by focusing on a specific text domain: for instance medical text, bible texts, historical text, etc. It will be interesting to implement an auto-tuning function to refines the maps depending on the application (TTS, SR, or other application). We will also look for ways of reducing the size of maps without loss of performance or accuracy.

5. References

[1] Carvalho V. R. and W. W. Cohen. 2004. Learning to Extract Signature and Reply Lines from Email, Proc.of CEAS 2004.

[2] David D. Palmer, "Text Pre-processing", in Handbook of Natural Language Processing, Second Edition, Goshen, Connecticut, USA, 2010

[3] Golding A.R.and D. Roth. 1996. Applying Winnow to Context-Sensitive Spelling Correction, Proc. Of ICML'1996.

[4] Lita L. V., A. Ittycheriah, S. Roukos, and N. Kambhatla.2003. tRuEcasIng, Proc. of ACL 2003.

[5] Mikheev, A. 2000. Document Centered Approach to Text Normalization, Proc. SIGIR 2000.

[6] Minkov E., R. C. Wang, and W. W. Cohen. 2005. Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text, Proc. Of EMNLP/HLT-2005.

[7] Palmer D. D. and M. A. Hearst. 1997. Adaptive Multilingual Sentence Boundary Disambiguation, Computational Linguistics, Vol. 23.

[8] Sproat R., A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 1999. Normalization of nonstandard words, WS'99 Final Report. http://www.clsp.jhu.edu/ws99/projects/normal/.

[9] Stone, H.S., "On the uniqueness of the convolution theorem for the Fourier transform", NEC Labs. Amer. Princeton, NJ. Online: http://citeseer.ist.psu.edu/176038.html, accessed on 19 Mar 2008.

[10] Tang J., H. Li, Y. Cao, and Z. Tang. 2005. Email data cleaning, Proc. of SIGKDD'2005.

Automating psycholinguistic statistics computation: Procura-Palavras

João F. Machado[1], *José João Almeida*[1], *Alberto Simões*[1], *Ana Soares*[2]

[1]Departamento de Informática, Universidade do Minho, Portugal
[2]Escola de Psicologia, Universidade do Minho, Portugal
joaoffm@gmail.com, jj@di.uminho.pt, ambs@di.uminho.pt, asoares@psi.uminho.pt

Abstract

This article describes psycholinguistic lexical databases available in various languages, including English, Spanish and Portuguese. These lexical databases are important for researchers in Psycholinguistics and other related areas, providing a pool of experimental materials and allowing for an efficient process of selection of these experimental materials.

The process of gathering statistics is slow, resulting in a small pool of materials in the short-term. The need to find an alternative method to gather limited or yet unavailable statistics for a specific language led us to consider gathering statistics from other languages and to compute their triangulation. Our aim was to automatize the computation of statistics such as Familiarity, Imageability, Age of Acquisition and Written Word Frequency for that specific language.

We will describe the process of preparing this data and triangulating and comparing statistics for some languages in an attempt of finding a relationship between them. The results were analysed considering correlations between each statistic in each pair of languages and by computing the mean of absolute differences between each language's values.

Index Terms: psycholinguistic, lexical databases, psychology, linguistics

1. Introduction

Psycholinguistics is an interdisciplinary area related to various fields, such as psychology, cognitive science and linguistics. It is the study of the process by which the human mind understands language.

For those planning studies on linguistic processing, an important requirement is the undeniable need for lexical databases. These databases are the foundation of most psycholinguistic studies and they can have a great impact in the quantity and quality of those studies, providing a large pool of experimental material and allowing for a strict selection of that material.

This paper's main objective is to analyse available resources, not only for the European Portuguese language, but also for various others languages. This analysis will aid in the development of a tool that will support investigation and will be of crucial importance mainly in the area of Psycholinguistics, but also for other areas such as Cognitive Psychology, Neurosciences or Artificial Intelligence.

Despite being primarily a survey, it also contains experiments aiming to investigate the possibility of automating the computation of Portuguese words' psycholinguistic statistics based on other languages' statistics and determining the reliability of those computations. This research main goal is to report the study of an alternate method of obtaining statistics with some degree of confidence, that will allow a potentially larger pool of estimated statistics available to researchers, when none or few statistics are available.

2. Lexical databases

A key component for conducting a thorough Psycholinguistic investigation involving linguistic stimuli is the availability of comprehensive software applications that enable researchers to compute relevant psycholinguistic statistics based on lexical resources.

2.1. English resources

The English language has an application and bundled database called *N-Watch* [1], a simple tool for obtaining a broad range of lexical statistics. It provides measures of word frequency, orthographic similarity, orthographic and phonological structure, age of acquisition and imageability. The default vocabulary of 30 605 words was obtained from the *CELEX ECT* [2], a corpus of 17.9 million words.

2.2. Spanish resources

An adaptation of the *N-Watch* application was prepared for the Spanish Language, named *BuscaPalabras* (B_PAL) [3], which includes measures of word, syllable, bigram and biphone frequencies, orthographic similarity, orthographic and phonological structure, concreteness, familiarity, imageability, valence, arousal and age of acquisition. It features some important differences from the original English program, such as support for the Spanish orthographic system, statistics related to syllable measures, and lastly it enables user-defined statistics. The default vocabulary of 31 491 words was obtained from *LEXESP* [4], a corpus of approximately 5 million words.

2.3. Portuguese resources

In 2003 a European Portuguese lexical database called PORLEX [5] was made available. It provides a series of psycholinguistic statistics, and although it contains information for a total of 29 238 words, it has several limitations. The lexical frequency value is only available for 5% of those words, and it lacks semantic information and subjective psycholinguistic statistics that recent research [6] has proven to be of great importance.

Another research focuses on rated age of acquisition norms and their relation with other psycholinguistic statistics [6]. It contains a database of 834 nouns that includes age of acquisition information but also imageability, familiarity, written word frequency, concreteness, number of syllables and number of words.

In 2000, the project *Multifunctional computational lexicon* *of contemporary Portuguese* [7] was concluded, with a default vocabulary of 26 443 words and 140 315 lemmas. It added much needed frequency values based on a large and diverse corpus, called CORLEX [8]. This corpus includes 16 210 438 words (95% of written corpora and 5% from oral corpora).

2.4. The project *Procura palavras*

As opposed to other languages, the current databases available for European Portuguese (eg. PORLEX, CORLEX) are outdated, limited or small, especially regarding lexical frequency and subjective psycholinguistic statistics.

In light of these conditions, we are initiating a project called *Procura Palavras* (*P-PAL*)[1], whose main goal is the development of a multi-platform software application that enables researchers to easily and simultaneously compute a broad range of objective and subjective linguistic and psycholinguistic statistics.

3. Triangulating psycholinguistic statistics

From the analysed resources we can argue that the current European Portuguese databases are unsatisfactory regarding psycholinguistic statistics. Although a gathering of psycholinguistic statistics is currently in motion, this process is extensive and time-consuming. In order to prepare a base set of statistics in less time another solution came to light: computing Portuguese statistics using other languages statistics.

For this process to be attainable with some degree of confidence, some crucial steps are required, including:

- normalizing the statistics of each language so they can be comparable (as different languages use different scale ranges),
- importing statistics to a single database for proper querying,
- connecting statistics trough translations from English (EN) to Portuguese (PT), English to Spanish (ES), Spanish to Portuguese and Spanish to English,
- filtering out possible erroneous or low confidence translations,
- determining correlations[2] and mean[3] of absolute differences between each language values,
- computing initial values for statistics with valid correlations.

For this article we selected a few psycholinguistic statistics: Familiarity (FAM), Imageability (IMG), Age of Acquisition (AoA) and Written Word Frequency (WWF):

- *Familiarity* is a statistic measured by asking readers to rate theirs familiarity with a word. A common query is to ask readers to rate how familiar they are with a specific word, measuring this familiarity on a scale of 1 (very unfamiliar) to 7 (very familiar).
- *Imageability* is a statistic measured by asking readers to rate how well they can form an image of that word in their heads, and rating it in a scale of 1 (impossible) to 7 (very easy).
- *Age of Acquisition* is a statistic measured by asking readers to estimate the age at which they think they have learned the real meaning for a word and, for example, estimating it on a 7-point scale (0-2, 3-4, 5-6, 7-8, 9-10, 11-12, and 13 or more years).
- *Written Word Frequency* is the frequency with which a word appears in a written corpus.

3.1. Available statistics

This section shows the available statistics for each language, that are relevant for the research, including the English language with statistics from the *N-Watch* application [1], the Spanish language with statistics from the *BuscaPalabras* application [3], and the Portuguese language with statistics from the paper *Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables* [6]. All this information is summarized in Table 1.

Table 1: Language (Lang), Statistic (Stat), Number of Words (N), Range from (F), and Range to (T) for: Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Written Word Frequency (WWF). AOA range in years (Y).

Lang	Stat	N	F	T
EN	AOA	3 136	100 ($Y \leq 2$)	700 ($Y \geq 13$)
EN	FAM	4 944	100	700
EN	IMG	4 944	100	700
EN	WWF	30 591	0	1.000.000
ES	AOA	139	1 ($Y \leq 1$)	11 ($Y \geq 11$)
ES	FAM	6 223	1	7
ES	IMG	6 096	1	7
ES	WWF	31 491	0	1 000 000
PT	AOA	834	1 ($Y \leq 2$)	7 ($Y \geq 13$), 8
PT	FAM	808	1	5
PT	IMG	249	1	7
PT	WWF	790	0	15 354 243

To make this table easier to understand consider the following explanation: there are 834 words available from the Portuguese Database that have values for at least one of the statistics considered. AoA rates are available for all 834 words, range from 1 (2 years old or less) to 7 (13 years old or more) and include an eight extra point (meaning learned in adulthood). FAM rates are available for 808 words and range from 1 (highly familiar) to 5 (very unfamiliar). IMG rates are provided for 249 words ranging from 1 (smaller imageability) to 7 (greater imageability), and 790 words have WWF measures per 15 million.

3.2. Triangulation Procedure

In order to process these statistics there was the need to import them to a common database, allowing for a proper and simpler querying and providing a meticulous analysis. This was done by developing a Perl script to parse each language database and import their data to a MySQL database.

An additional task consisted in normalizing statistics, as they have different ranges for each language. Table 2 presents the normalization formula used for each language and its resulting normalized range. Without this normalization, values would not be comparable and no analysis could have been performed.

Given the difficulty to compare WWF among languages the Logarithm of Written Word Frequency (LOG-WWF) was computed. LOG-WWF ranged from -2.81 to 13.81. To make it

[1] `http://natura.di.uminho.pt/p-pal/`

[2] the degree to which two or more attributes or measurements on the same group of elements show a tendency to vary together

[3] something having a position, quality, or condition midway between extremes; a medium

Table 2: Language (Lang), Normalization result (N) and Normalization formula (F) for ES and PT databases: Familiarity (FAM), Age of Acquisition (AOA), Imageability (IMG) and Written Word Frequency (WWF)

Lang	Statistic	F	N
ES	FAM	$R \times 100$	100–700
ES	IMG	$R \times 100$	100–700
ES	AOA	$((R * \frac{1}{2}) + \frac{1}{2}) \times 100$	100–700
PT	FAM	$((5 - R) \times \frac{3}{2} + 1) \times 100$	100–700
PT	IMG	$R \times 100$	100–700
PT	AOA	$R \times 100$	100–800
PT	WWF	$R/15$	p/million

easier to compare, we added 3 units to the value, resulting in a positive range of 0.18 to 16.81 that has better legibility.

There was a need to connect each word from one language to another. This connection was performed by translating each English (EN) word to Portuguese (PT) and Spanish (SP), and each Spanish word to Portuguese and English. For this task a Perl Module, `Lingua::Translate` , was used with a backend for Google's translation system .

There are two approaches for word connection, each resulting in different levels of confidence. The first method consists in linking words from each language through its equivalent in Portuguese (translating each word to Portuguese and using that word as the pivot element).

$$\textbf{Method 1 (M1)} \begin{cases} \text{Translate-To-PT(EN Word) = PT Word} \\ \text{Translate-To-PT(SP Word) = PT Word} \end{cases}$$

The second method consists primarily in matching Spanish and English words with a single equivalent word in Portuguese, and also ensuring that English-to-Spanish and Spanish-to-English translations match.

$$\textbf{Method 2 (M2)} \begin{cases} \text{Translate-To-PT(EN Word) = PT Word} \\ \text{Translate-To-PT(SP Word) = PT Word} \\ \text{Translate-To-EN(SP Word) = EN Word} \\ \text{Translate-To-SP(EN Word) = SP Word} \end{cases}$$

This last method will prevent duplicate words and result in more accurate translations, though it will render a smaller intersection set.

The last step consists in computing statistics, including amplitude and mean for each language, and correlations and mean of absolute differences between values for each pair of languages.

4. Results and Discussion

The reliability of each statistic in English-Portuguese, Spanish-Portuguese and English-Spanish triangulation was analyzed in two different stages:

- **Stage 1:** the distance mean for each statistic and language pair was computed, i.e., the mean of the absolute values of the difference between values of a given statistic in a language pair.
- **Stage 2:** correlation (ranging from -1 to 1) was computed for each pair of languages. A value close to 0 shows that there is no relationship within the variables, whereas a value close to $+1$ or -1 indicates that the variables are related.

4.1. English–Portuguese and Spanish–Portuguese with M1

Considering English and Spanish words connected only unidirectionally to Portuguese words, as shown in Table 3, correlations for IMG have high values (0.78 and 0.70), with an average of absolute differences of 15.16% from English and 15.86% from Spanish, thus suggesting a more viable triangulation with an average error of 15-16%. In plain terms, an error of 16% on a 7-point scale corresponds to an error of 1 point, which translates to the minimum possible error, one likely to occur when asked of participants to estimate a word's statistic.

Although AoA also has high correlations (0.62 and 0.80), its smaller sample of only 78 words from Spanish may well be viewed as less reliable. On the other hand, its sample of 307 words from English appears less unreliable suggesting a triangulation with an average error of 11.51%.

Table 3: Differences between EN–PT & ES–PT values: Language Pair (LP), Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 1.

Statistic	LP	N	M	A	C
FAM	EN-PT	457	14.52	0.00–53.83	0.29
FAM	ES-PT	536	15.71	0.00–70.67	0.18
IMG	EN-PT	213	15.16	0.17–58.00	0.78
IMG	ES-PT	248	15.86	0.00–72.83	0.70
AOA	EN-PT	307	11.51	0.17–66.17	0.62
AOA	ES-PT	78	7.86	0.33–28.33	0.80
LOG-WWF	EN-PT	863	8.13	0.00–40.90	0.54
LOG-WWF	ES-PT	1350	13.00	0.00–47.86	0.42

4.2. English–Portuguese and Spanish–Portuguese with M2

When connecting English and Spanish words bidirectionally to Portuguese (see table 4), a slight reduction of connected words occurs. Correlations for IMG have small but significant increases (0.78 to 0.86 and 0.70 to 0.83). For AoA there is a minor increase (0.80 to 0.82) in Spanish to Portuguese, which is meaningful (0.62 to 0.79) in English to Portuguese, ensuing a more confident triangulation with an even better average error of 8.9%. Lastly there is a major increase in LOG-WWF (0.54 to 0.81 and 0.42 to 0.85), adding one more statistic to the list of possible reliable triangulations with smaller average errors of 4.68% and 4.91%.

4.3. English–Spanish with M1

When comparing English and Spanish statistics, connected by their Portuguese word equivalent, a larger pool of words becomes available, along with their corresponding statistics. This may well result in a more confident analysis, reflected in Table 5. After careful examination, IMG once again tends to be a more reliable triangulation with a high correlation of 0.62 and an average error of 12.88%.

4.4. English–Spanish with M2

A decrease in connected words occurs when connecting English to Spanish by the Portuguese translation, and also by ensuring that the English to Spanish translation matches the original Spanish word, and that the Spanish to English translation matches the original English word. Despite this relatively large

Table 4: Differences between EN–PT & ES–PT values: Language Pair (LP), Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 2.

Statistic	LP	N	M	A	C
FAM	EN-PT	340	13.21	0.00–38.50	0.40
FAM	ES-PT	292	14.38	0.00–57.50	0.30
IMG	EN-PT	152	15.60	0.17–58.00	0.86
IMG	ES-PT	138	14.99	0.00–72.83	0.83
AOA	EN-PT	227	8.90	0.17–49.50	0.79
AOA	ES-PT	65	7.88	0.33–28.33	0.82
LOG-WWF	EN-PT	505	4.68	0.00–27.82	0.81
LOG-WWF	ES-PT	445	4.91	0.00–23.54	0.85

Table 5: Differences between EN & ES values: Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 1.

Statistic	N	M	A	C
FAM	4 030	13.17	0.00–72.67	0.35
IMG	3 950	12.88	0.00–75.00	0.62
AOA	86	7.67	0.00–52.33	0.44
LOG-WWF	32900	10.34	0.00–62.54	0.40

decrease, their absolute values are still very high. As shown in Table 6, IMG increases considerably in its correlation (0.62 to 0.73), supporting the hypothesis that this is a reliable statistic for triangulation with a slightly smaller average error of 11.01%. Another increase occurs in LOG-WWF (0.40 to 0.79) once again confirming this statistic's possible triangulation with an even smaller average error of 5.33%.

Table 6: Differences between EN & ES values: Number of Words (N), Means in percentage (M), Amplitude in percentage (A), and Correlation (C) for Age of Acquisition (AOA), Familiarity (FAM), Imageability (IMG), and Logarithm of WWF plus 3 (LOG-WWF). Using method 2.

Statistic	N	M	A	C
FAM	1 720	11.32	0.00–59.17	0.60
IMG	1 684	11.01	0.00–75.00	0.73
AOA	58	5.53	0.17–35.17	0.66
LOG-WWF	7 651	5.33	0.00–36.98	0.79

5. Conclusions

This paper reports a research on *Automating psycholinguistic statistics computation based on other languages' statistics*, including an analysis of the English program called N-Watch, the Spanish program called B_PAL and, lastly, a smaller Portuguese database.The analysis of these tools and of the available Portuguese databases (eg. CORLEX, PORLEX), acknowledges the need to design and implement a similar tool that can incorporate existing databases and promote research in Psycholinguistics for the Portuguese language.

For this paper's research Psycholinguistic statistics were imported from those three databases, including Familiarity, Imageability, Age of Acquisition and Written Word Frequency.

Although higher pools of words are available by connecting words only with their Portuguese translation, these words include erroneous or duplicate translations, which results in less reliable data. To enhance this, a stronger approach was used to enable a more accurate connection between each language's words. Although resulting in a smaller pool of words and statistics, these results appear more reliable allowing for a more confident and accurate inference of possible triangulations.

Results for correlation and mean of absolute differences between each language's values for each statistic seems to indicate that Imageability and WWF (through the analysis of LOG-WWF) may be automatically computed with some reliability from a triangulation from English and Spanish languages. Age of Acquisition yield interesting results, although further analysis with larger samples will be needed to conclude about their reliability.

6. Acknowledgments

This paper stems from project "Procura-Palavras (P-Pal): a software program for deriving objective and subjective psycholinguistic indices for European Portuguese words" (PTDC/PSI-PCO/104679/2008), funded by FCT (Foundation for Science and Technology), by NSRF (National Strategic Reference Framework), and by COMPETE (Operational Agenda for Competitiveness Factors), a program created by the EU as part of the European Regional Development Fund (ERDF).

7. References

[1] C. J. Davis, "N-Watch: a program for deriving neighborhood size and other psycholinguistic statistics," *Behavior Research Methods*, vol. 37, no. 1, pp. 65–70, 2005.

[2] R. H. Baayen, R. Piepenbrock, and L. Gulikers, "The CELEX lexical database (release 2)[cd-rom]," *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]*, 1995.

[3] C. J. Davis and M. Perea, "BuscaPalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in spanish," *Behavior Research Methods*, vol. 37, no. 4, pp. 665–671, 2005.

[4] N. Sebastián-Gallés, M. A. Mart, M. Carreiras, and F. Cuetos, "LEXESP: léxico informatizado del espanol," *Barcelona: Edicions de la Universitat de Barcelona*, 2000.

[5] I. Gomes and S. L. Castro, "Porlex, a lexical database in european portuguese," *Psychologica*, vol. 32, p. 91–108, 2003.

[6] J. Marques, F. Fonseca, A. Morais, and I. Pinto, "Estimated age of acquisition norms for 834 portuguese nouns and their relation with other psycholinguistic variables," *Behavior Research Methods*, vol. 39, no. 3, pp. 439–444, Aug. 2007.

[7] M. Nascimento, J. Casteleiro, M. Marques, F. Barreto, and R. Amaro, "Léxico multifuncional computorizado do português contemporâneo [Multifunctional computational lexicon of contemporary Portuguese](data file)," *Available from Centro de Linguística da Universidade de Lisboa Web site: http://www.clul.ul.pt*, 2000.

[8] M. F. B. do Nascimento, J. M. Casteleiro, M. L. G. Marques, F. Barreto, R. Amaro, and R. Veloso, "Corlex: Léxico de frequências do português [Base lexical]," *Consultado em http://www.clul.ul.pt (30 de Outubro de 2007)*, 2007.

A statistical study of the WPT05 crawl of the Portuguese Web

David Batista, Mário J. Silva

LaSIGE, Faculty of Sciences, Lisbon, University of Lisbon, Portugal
dsbatista@xldb.di.fc.ul.pt

Abstract

This article presents a statistical study of WPT05, a text corpus derived from a crawl of the Portuguese Web performed in 2005. This corpus is a valuable resource for researchers in Natural Language Processing (NLP). As one of the biggest publicly available collections of European Portuguese texts, we provide statistical analyses of the contents, covering the languages identified, the representativity of the top-level domains crawled and terms frequency and size. An analysis of an n-grams collection extracted from the Portuguese documents in the corpus is also presented. We analyze the occurrence of first names, surnames and geographic names in the corpus. Since some toponyms are named after personal names, we show the overlap of Portuguese names with geographic entities corresponding to places in Portugal.

Index Terms: web corpus, resources, Portuguese

1. Introduction

This study presents a statistical analysis of the textual contents of WPT05, a 2005 crawl of the Portuguese Web. WPT05 is the successor to WPT03 [1], a crawl from 2003 released earlier. WBR-99, a crawl from 1999 of the Brazilian Web, is another large collection of 6 million documents [2].

The Web pages that are part of the WPT05 Collection were retrieved by the crawler of the Tumba! search engine [3]. This crawl targeted documents written in Portuguese, hosted in a .PT domain, or hosted in the .COM, .NET, .TV, .INFO, .BIZ, .TK, .CC and .FM domains and referenced by a hyperlink from, at least, one page hosted in a .PT domain. In addition to these domains, a set of individual sites considered relevant by the developers of the crawler as well.

The content of WPT05 is available in 3 formats: as raw data, as text only documents with the metadata associated and as an n-grams collection.

The raw format includes the documents as they were crawled, without any sort of post-processing, such as filtering of some document types, elimination of duplicates, or text encoding normalization. We adopted the Internet Archive file format (ARC), designed for the specific purpose of preserving web pages as they were crawled [4].

The text-only format of the collection contains metadata associated with each document. Its production is described in the Master dissertation of David Cruz [5]. This format uses the Resource Description Framework (RDF) technology and the Open Archives Initiative Object Reuse and Exchange (OAI-ORE) specification [6]. It allows the preservation of the hierarchy of pages within each domain and the flagging of duplicate documents, for wich we mark the additional URL where the same contents were found in the case of duplicates instead of including a replica. We provide, for each document, the hierarchy of domains and duplicate information along with the identified language and crawling metadata, such as the IP address, the HTTP server running and the date and time when the document was fetched. This format contains text-rich documents only, namely, documents of the following MIME types:

- application/pdf
- application/postscript
- application/vnd.ms-office
- text/html, text/plain
- text/rtf

All the extracted text is encoded in the UTF-8 format and each file of the distributed collection is a valid XML file, enabling its handling by the tools commonly available for RDF and XML processing.

A third format of the collections is an n-grams dataset, which is described in detail in the next section. The n-grams collection was extracted from the collected documents whose identified language was Portuguese. We extracted word n-grams up to the fifht order (5-grams) using the Ngram Statistics Package [7]. A set of regular expressions to tokenize the text were applied. These regular expressions are part of the Lingua-PT-PLNbase-0.21 [8], a Perl extension for NLP of Portuguese which include a tokenizer available from Linguateca [9]. After the extraction, all n-grams with tokens having more than 32 characters were discarded. N-grams with frequencies below 5 were discarded as well. The n-grams collection is available as a set of UTF-8 encoded files, containing the n-grams and their frequencies.

2. Statistics

We analyzed the languages present in each document and the top-level domains from where documents were crawled and obtained several statistics concerning the number of unique extracted terms and the frequency and length of each, the top n-grams and also the amount of geographic information present in the WPT05 crawl. All the statistical data presented in this section was obtained from the RDF format of the collection or from the extracted n-grams. The RDF version of the collection has a total of 12,523,110 URLs, of these 9,483,489 with unique textual content.

2.1. URLs

Table 1 shows the percentage of the most targeted top-level domains (TLD) from which documents were crawled. Almost 70% of the crawl comes from .PT followed by .COM and .NET.

2.2. Language

The language for each document was detected with a popular n-gram analysis algorithm [10]. NGramJ [11], a software tool

TLD	Percentage of URLs
.pt	69.92%
.com	23.26%
.net	3.76%
.vu	1.73%
.org	1.13%
.others	0.21%

Table 1: *Top Level Domains of URLs crawled.*

implementing the algorithm was used to perform language detection in the extracted text from each document. NGramJ contains profiles for about 70 languages using up to 4-grams for identification. Only documents with more than 200 bytes in size were considered, which totals 8,877,430 documents. Documents classified as unknown correspond to harvested pages that despite presenting rich-text, the contents only contain URLs, email addresses, web server directory listings or similar contents.

Language	N^{o} Documents	MBytes	Percentage
Portuguese	7 412 778	24 707	83.50 %
English	941 711	3 423	10.61 %
Spanish	206 732	800	2.33 %
Others	210 014	720	2.37 %
Unknown	106 195	308	1.20 %

Table 2: *Language Distribution over documents.*

The languages per document distribution presented in Table 2 has a similar pattern as that of the crawl of the Portuguese Web from 2003 [12], although in this study the percentage of Portuguese documents was higher. The amount of Portuguese text in the collection is around 25 Gbytes, from almost 7.5 millions documents. There is no distinction between the different variations of Portuguese, such as European, Brazilian or African.

2.3. Terms

An n-gram is a subsequence of n items from a given sequence. The items can be, for example, words from a sentence, characters from a word or phonemes from a sound, depending on the application. We extracted up to 5 word n-grams from the Portuguese documents. Table 3 lists the number of unique identified n-grams from unigrams up to pentagrams, as well as the size of each set. We used the extracted unigrams to calculate the number and frequency of individual terms. As the n-grams were extracted only from documents identified as Portuguese, most of the terms have a high likelihood of being used in Portuguese.

Table 4 shows the average, median, standard deviation and mode for the frequency of terms and size of terms. Regarding the frequency, the median of 16 and the mode of 5 show that most of the identified terms have the cut-off frequency (5) of the collection. Half of all the identified terms have a frequency of 16 or less. This suggest that the term frequency, as in the crawl of 2003 [1], and other web crawls, follows a Zipf law [13].

2.4. Top N-Grams

We present in Table 5 the top 25 most frequent unigrams and bigrams. Only n-grams with tokens that do not contain any punctuation mark are included. These n-grams are potential candidates to a Portuguese stop-words list. Table 6 lists the top 25 trigrams. Some of the n-grams contain terms which are not Portuguese. This happens because a large portion of documents identified as Portuguese also contain English terms. These terms, such as *Blog* or *Next Blog*, are most likely part of English interfaces of content publishing systems, such as blogs.

N-Grams	Count	Size
Unigrams	2 111 004	25 Mb
Bigrams	27 674 092	432 Mb
Trigrams	71 307 404	1 400 Mb
Tetragrams	89 668 947	2 100 Mb
Pentagrams	84 378 473	2 300 Mb

Table 3: *Statistics of the WPT05 Portuguese N-Grams collection*

3. Personal Names and Toponyms prevalence

We analyzed the occurrence of personal names, surnames and toponyms in the extracted n-grams. We were interested in discovering the overlap between person names and toponyms, as traditionally many geographic references, such as streets, are named after a personality's name, and many people have a placename as their surname in Portuguese.

3.1. Geographic Entities

The corpus was analyzed for the presence of geographic references. We did a search with base on Geo-Net-PT02 [14] [15] a public geographic ontology of Portugal, that contains the geographic administrative and physical data about districts, municipalities and streets, rivers, beaches, among others. We looked up in the n-grams collections for occurrences of names which correspond to geographic concepts in the geographic ontology.

Each geographic concept in Geo-Net-PT02 is associated to a name. The name is represented by 3 different variations: capitalized, non-capitalized, and simple ASCII. Table 7 shows an example of the representations. Geo-Net-PT02 contains 51,292 unique names for different geographic concepts.

We searched in n-grams for occurrences of the three different representations, 97.82% of the geographic concept names were found in WPT05 in a capitalized representation. This evidences the use of capitalization to refer to geographic place names.

Table 8 shows the coverage of geographic names in WPT05, that is, the percentage of geographic concept names found for each representation, as well as the average number of occurrences, the median, standard deviation and mode.

This approach is naive, as the occurrences of these names in WPT05 might be references to other concepts rather than only

Measure	Term Frequency	Term Size
Average	2.14	8.29
Median	16	11
Standard deviation	2 421 778	3.18
Mode	5	7

Table 4: *Term size and term occurrences statistical characterization*

Unigram	Count	Bigram	Count
de	151 331 293	para o	3 654 827
a	80 751 534	o que	3 588 803
e	78 057 840	que o	3 510 621
o	59 632 368	para a	3 450 908
que	58 002 495	e a	3 353 043
do	48 119 636	com a	3 156 764
da	39 445 585	com o	3 131 003
em	31 807 331	de um	3 122 238
para	30 871 814	que se	2 930 294
com	29 709 820	que a	2 763 518
um	24 032 617	e o	2 714 772
se	23 482 819	Todos os	2 603 578
os	21 718 820	que não	2 510 046
não	19 841 653	a sua	2 412 501
é	19 392 183	de uma	2 408 046
por	19 273 135	todos os	2 090 440
no	18 954 414	o seu	2 089 973
A	17 753 909	Powered by	1 813 626
uma	17 575 533	Responder com	1 763 910
O	17 201 084	Enviar Mensagem	1 729 514
na	15 501 678	Ver o	1 649 408
as	14 618 221	com Citação	1 636 297
dos	14 265 211	é o	1627 771
mais	13 425 740	os direitos	1 568 047
ao	11 609 150	em que	1 543 058

Table 5: *Top 25 occurring unigrams and bigrams in WPT05 corpus. N-grams with punctuation marks were removed*

geographic locations, such as personal names, organizations. However, it still provides a relevant measure of the prevalence of geographic names in WPT05.

3.2. Personal Names and Surnames

We gathered Portuguese personal names and surnames from a public list and looked for its occurrences in the WPT05 unigrams. Our list consists of 1,786 unique personal names and surnames. These were collected from the public lists of placed secondary teacher names in the 2009 recruitment, available from the Portuguese Ministry of Education website. Table 9 lists the top twenty most frequent first names and surnames. Here is also important to note that some surnames might have other semantic meanings, for instance a reference to a month.

3.3. Overlap of Personal Names, Surnames and Toponyms

Typically many first names and surnames are used as toponyms. We looked for the overlap between Portuguese names and toponyms, based on the occurrences in WPT05. From the 1,786 names, 1,030 where found to have a correspondent geographic name in Geo-Net-PT02, around 57%. Table 10 shows the top 20 most frequent Portuguese names in WPT05 that also represent a geographic concept names, and the number of geographic concepts having that name. This information could be useful for word-sense-disambiguation systems on words that can represent both a geographic concept and a person's name.

Trigrams	Count
Responder com Citação	1 630 843
Ver o perfil	1 516 648
o perfil de	1 503 227
os direitos reservados	1 460 648
Enviar Mensagem Privada	1 414 293
Todos os direitos	1 366 069
perfil de utilizadores	1 196 436
de utilizadores Enviar	1 176 949
utilizadores Enviar Mensagem	1 174 793
Get your own	939 337
your own blog	934 967
Next blog BlogThis	934 480
Blogger Get your	934 450
own blog Next	915 500
blog Next blog	915 500
Voltar acima Ver	911 284
Índice do Fórum	763 560
de Julho de	759 366
Não há mensagens	756 468
há mensagens novas	731 423
a um amigo	700 625
Julho de 2005	675 378
Powered by Blogger	650 165
a última mensagem	560 605
mensagem Não há	489 854

Table 6: *Top 25 trigrams in WPT05 corpus. Trigrams with punctuation characters were removed*

Capitalized	Non-Capitalized	Simple ASCII
Alcácer do Sal	alcácer do sal	alcacer do sal
Dão-Lafões	dão-lafões	dao-lafoes
Lisboa	lisboa	lisboa

Table 7: *Different representations of a geographic concept's name*

4. Conclusions

This was a first statistic study over the text extracted from the WPT05 collection. The raw format of WPT05 collection was produced by the XLDB Node of Linguateca in 2005. The RDF/XML was produced in 2008 and the n-grams collection was extracted in 2010. By the size of the collection and being the most part of the contents crawled from the .PT top-level domain, this is currently one of the biggest available collections in European Portuguese. The provenance of the extracted textual contents are diverse websites, spreading from personal blogs to newspapers and institutional organizations or forums. This gives a diverse and rich genera of texts, capturing different lin-

Measure	Capitalized	Non-Capitalized	ASCII
Coverage	97.8%	43.6%	42.0%
Average	5.4	3.0	5.4
Median	21	0	0
Standard deviation	62.9	58.4	199.6
Mode	1	0	0

Table 8: *Statistical characterization of occurrences of geographic names in WPT05*

Names	# Occurrences
Portugal	4 340 513
Porto	2 074 629
João	1 886 903
São	1 701 404
Pedro	1 643 292
Paulo	1 587 559
José	1 580 473
Maio	1 512 650
Janeiro	1 403 262
Novo	1 329 434
Maria	1 278 973
Silva	1 178 842
Dias	1 061 872
Bem	1 045 555
Nuno	1 034 905
Miguel	1 003 402
Carlos	971 723
Rui	969 096
Jorge	961 599
Nova	923 395
Rio	913 218
Deus	913 098
António	901 979
Santos	845 191
Manuel	834 351

Table 9: *Top 25 occurring Portuguese first names and surnames in WPT05*

Names	# Occurrences
Portugal	4 340 513
Porto	2 074 629
Pedro	1 643 292
Paulo	1 587 559
Maio	1 512 650
Janeiro	1 403 262
Novo	1 329 434
Maria	1 278 973
Silva	1 178 842
Dias	1 061 872
Miguel	1 003 402
Carlos	971 723
Jorge	961 599
Nova	923 395
Rio	913 218
Deus	913 098
Santos	845 191
Saúde	832 797
Costa	770 628
Rua	769 114
Ferreira	748 912
Luís	717 840
Ana	707 308
Tiago	692 283
Pereira	674 330

Table 10: *Top 25 overlapping Portuguese names with Portuguese geographic place names*

guistic styles.

All the three forms of the web crawl are available upon request through the Linguateca [1] and XLDB [2] websites. WPT05 is made available exclusively for research purposes.

5. Acknowledgements

We wish to thank Daniel Gomes for harvesting the documents in WPT05 and also to David Cruz for generating the RDF/XML format of the collection. This work was supported by FCT (Portugal), through the project PTDC/EIA/73614/2006 (GREASE-II) and the Multiannual Funding Programme.

6. References

[1] B. Martins and M. J. Silva, "A Statistical Study of the Tumba! Corpus," in *Advances in Natural Language Processing, 4th International Conference, EsTAL 2004, Alicante, Spain, October 20-22, 2004, Proceedings*, 2004, pp. 384–394, also available as University of Lisbon, Faculty of Sciences, Technical Report DI/FCUL TR 4-4.

[2] Pável Calado, "The WBR-99 collection: Data-structures and file formats," Department of Computer Science, Federal University of Minas Gerais, Tech. Rep., 1999. [Online]. Available: http://www.linguateca.pt/Repositorio/WBR-99/wbr99.pdf

[3] Daniel Gomes and Mário J. Silva, "The Viúva Negra crawler: an experience report," *Software: Practice and Experience (SPE)*, vol. 38, no. 2, pp. 161–168, February 2008. [Online]. Available: http://dx.doi.org/10.1002/spe.825

[4] "Internet Archive ARC File Format," http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml.

[5] D. Cruz, "Sidra5: A search system with geographic signatures," Master's thesis, University of Lisbon, Faculty of Sciences, November 2007.

[6] "Open Archives Initiative Object Reuse and Exchange," http://www.openarchives.org/ore/.

[7] "Ngram Statistics Package (NSP)," http://ngram.sourceforge.net/.

[8] "Lingua::PT::PLNbase - Perl extension for NLP of the Portuguese," http://search.cpan.org/~ambs/Lingua-PT-PLNbase-0.21/.

[9] D. Santos, "Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva," *Linguamática*, vol. 1, no. 1, pp. 25–58, May 2009. [Online]. Available: http://linguamatica.com/index.php/linguamatica/article/view/20/9

[10] William B. Cavnar and John M. Trenkle, "N-Gram-Based Text Categorization," in *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175.

[11] "NGramJ, Smart Scanning for Document Properties," http://ngramj.sourceforge.net/.

[12] B. Martins and M. J. Silva, "Language Identification in Web Pages," in *ACM-SAC-DE, 20th ACM Symposium on Applied Computing, Document Engeneering Track*, April 2005, pp. 764–768. [Online]. Available: http://doi.acm.org/10.1145/1066677.1066852

[13] G. Zipf, *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA), 1949.

[14] F. J. Lopez-Pellicer, M. Chaves, C. Rodrigues, and M. J. Silva, "Geographic ontologies production in grease-ii," University of Lisbon, Faculty of Sciences, LASIGE, Tech. Rep. TR 09-18, November 2009. [Online]. Available: http://hdl.handle.net/10455/3256

[15] M. S. Chaves, "Uma metodologia para construção de geo-ontologias," Ph.D. dissertation, Faculty of Sciences, University of Lisbon, September 2009. [Online]. Available: http://www.linguateca.pt/documentos/TeseDoutMarcirioChaves2009.pdf

[1] `http://www.linguateca.pt`

[2] `http://xldb.fc.ul.pt/wiki/WPT_05_in_English`

Keynote Talks

Heiga Zen. Toshiba Research Europe Ltd. (UK)

Fundamentals and recent advances in HMM-based speech synthesis

Abstract

Statistical parametric speech synthesis based on HMMs has grown in popularity over the last years. In this talk, its system architecture is outlined, and then basic techniques used in the system, including algorithms for speech parameter generation from HMM, are described with simple examples. Relation to the unit selection approach and recent improvements are summarized. Techniques developed for increasing the flexibility and improving the speech quality are also reviewed.

Speaker Bio

Heiga Zen received the Dr.Eng. degree in computer science and engineering from Nagoya Institute of Technology in 2006. He is currently a Research Engineer in the Speech Technology Group of Toshiba Research Europe Ltd. Cambridge Research Laboratory. He was an intern researcher at the ATR Spoken Language Translation Research Laboratories in 2003 and an intern/co-op researcher at the IBM T. J. Watson Research Center from 2004 to 2005. From April 2006 to July 2008, he was a postdoctoral research associate at the Nagoya Institute of Technology. He has been working on HMM-based speech synthesis for 9 years after joining Prof. Tokuda's research group in 2000. He was also the main developer and maintainer of HTS, one of the main developers of the Festival Speech Synthesis System, one of the main developers of SPTK, and one of the active contributors to the hidden Markov model toolkit (HTK). He published over 10 journal papers and over 40 conference papers, and received 5 paper awards.

Alex Acero
Microsoft Research (USA).

New Machine Learning approaches to Speech Recognition

Abstract

In this talk I will describe some new approaches to speech recognition that leverage large amounts of data using techniques from information retrieval and machine learning.

Speaker Bio

(picture and more at http://research.microsoft.com/en-us/people/alexac/) Alex Acero received a M.S. degree from the Polytechnic University of Madrid, Madrid, Spain, in 1985, a M.S. degree from Rice University, Houston, TX, in 1987, and a Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, in 1990, all in Electrical Engineering. Dr. Acero worked in Apple Computer's Advanced Technology Group in 1990-1991. In 1992, he joined Telefonica I+D, Madrid, Spain, as Manager of the speech technology group. Since 1994 he has been with Microsoft Research, Redmond, WA, where he is presently a Research Area Manager directing an organization with 70 engineers conducting research in audio, speech, multimedia, communication, natural language, and information retrieval. He is also an affiliate Professor of Electrical Engineering at the University of Washington, Seattle.

Dr. Acero is author of the books "Acoustical and Environmental Robustness in Automatic Speech Recognition" (Kluwer, 1993) and "Spoken Language Processing" (Prentice Hall, 2001), has written invited chapters in 4 edited books and 200 technical papers. He holds 78 US patents. Dr. Acero is a Fellow of IEEE. He has served the IEEE Signal Processing Society as Vice President Technical Directions (2007-2009), Director Industrial Relations (2009-2011), 2006 Distinguished Lecturer, member of the Board of Governors (2004-2005), Associate Editor for IEEE Signal Processing Letters (2003-2005) and IEEE Transactions of Audio, Speech and Language Processing (2005-2007), and member of the editorial board of IEEE Journal of Selected Topics in Signal Processing (2006-2008) and IEEE Signal Processing Magazine (2008-2010). He also served as member (1996–2000) and Chair (2000-2002) of the Speech Technical Committee of the IEEE Signal Processing Society. He was Publications Chair of ICASSP98, Sponsorship Chair of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, and General Co-Chair of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding. Since 2004, Dr. Acero, along with co-authors Drs. Huang and Hon, has been using proceeds from their textbook "Spoken Language Processing" to fund the "IEEE Spoken Language Processing Student Travel Grant" for the best ICASSP student papers in the speech area. Dr. Acero served as member of the editorial board of Computer Speech and Language and and member of Carnegie Mellon University Dean's Leadership Council for College of Engineering.

Bill Byrne
University of Cambridge (UK)

Hierarchical phrase-based statistical machine translation with weighted finite state transducers

Abstract

I will present a introduction and review of recent developments in statistical machine translation which exploit weighted finite state transducers to implement a variety of search and estimation algorithms. The presentation will describe work done at the University of Cambridge and the University of Vigo by Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, and Jamie Brunning. The focus will be mainly on translation but the approaches described are general and are also applicable to other problems in speech and language processing.

Speaker Bio

Bill Byrne is a Reader in Information Engineering in the Department of Engineering, University of Cambridge. His research is in statistical modelling techniques for speech and language processing, and he has worked on a variety of search and estimation algorithms for speech recognition, speech synthesis, and statistical machine translation. Current research interests include cross-lingual acoustic modelling for speech synthesis, weighted finite state transducers for hierarchical and syntactic phrase-based translation, and the use of natural language generation in statistical machine translation. He has published more than 100 refereed journal articles and conference papers and has an extensive history of editorial and professional service. He has received research funding from NSF(USA), DARPA(USA), Microsoft, and Google, and he is currently coordinator of the ICT-FP7 project FAUST (faust-fp7.eu) on interactive statistical machine translation. He came to Cambridge in 2004 as a Lecturer in Speech Processing, having been Research Associate Professor at the Johns Hopkins University Center for Language and Speech Processing (USA). He received his Ph.D. in Electrical Engineering from the University of Maryland, College Park, and he is a Fellow of Clare College, Cambridge.

Thesis/Project/Demo Session

Estimation and Uncertainty Processing Techniques for Signal Transmission and Recognition

Antonio M. Peinado, José L. Pérez-Córdoba, Victoria E. Sánchez,
Ángel M. Gómez, José L. Carmona, José A. González López Juan A. Morales Cordovilla

Departamento de Teoría de la Señal, Telemática y Comunicaciones
Universidad de Granada, Granada (Spain)
amp@ugr.es

Abstract

This paper presents the research project TEC2010-18009/TCM proposed for its funding by the Ministerio de Ciencia e Innovacin (MICINN). The main goal of this research project is the development of two groups of techniques for the processing of noisy, damaged or lost information: estimation and uncertainty processing. We will consider two different applications with a clear parallelism: speech recognition in adverse environments and error concealment for robust multimedia transmission (speech and video).

1. Introduction

During the last years, our research team has been developing its work on remote speech recognition (RSR). An RSR system employs client/server architecture for transmitting speech signals or parameters from a thin client (typically mobile devices such as smart-phones, VoIP phones or PDAs) to a powerful remote server where recognition is performed. Through a series of R+D projects, we have considered several issues related to RSR. Thus, we have dealt with robustness against adverse acoustic environments and against degraded transmission channels (mobile networks, IP, WAN,). In order to tackle with this last problem, our research team has employed, among others, two approaches especially attractive due to their statistical nature. First, we have considered several estimation techniques (mainly Bayesian) and, specially, the minimum square error (MMSE) criterion. Also, it must be taken into account that the transmitted speech features will not be, even after estimation at the decoder, reliable. The uncertainty processing techniques try to obtain some kind of reliability measures which can be incorporated to the speech recognizer in order to improve the system performance. In this proposal, the research team wants to exploit this previous experience and the results obtained in RSR about estimation/uncertainty techniques with the aim of deepening in them with a more general perspective, extending them to new problems and applications. In order to do that, we must undertake two general issues:

1. Estimation techniques. The most important point for their application is to obtain an statistical model (suitable for every application and signal) for the information (signal/parameters) generation and transmission process which can be integrated in a Bayesian estimation framework. Also, it is important to do a suitable use of the available data (distorted by noise or errors).

2. Uncertainty processing techniques. Basically, we consider here the noise-robust speech recognition problem, since state-of-the-art speech recognizers usually adopt a statistical approach, which allows a natural management of uncertainty. Two issues must be considered here: which reliability measures are to be used and how the speech recognizer is modified to incorporate these measures.

Estimation and uncertainty are related issues since, as mentioned, the estimation process involves uncertainty. Thus, although the reliability measures may be obtained from different criteria, it is also possible extracting them from the probability distributions employed for estimation.

This work has been supported the project TEC2007-66600.

2. Goals, Concepts and State of the Art

Taking into account the discussion in the previous Section, the goal of this proposal can be summarized as follows (see Fig. 1).

• The research team seeks a continuation of its traditional research, automatic speech recognition, as main application. This way, this project continues previous projects developed by the team on RSR, although focusing now on the acoustic noise problem. This is considered the main issue in order to obtain a ubiquitous and pervasive human-machine interaction. In particular, we intend to develop both estimation and uncertainty processing techniques under a joint conception and with the aim of allowing the collaboration between them. As we will see later in this section, missing data techniques provide a suitable framework for this goal.

• We also intend to extend estimation techniques to other applications and, in particular, to multimedia signal transmission. Thus, our team seeks the translation of its previous experience to a new field whith important applications such as VoIP telephony, audio/image/video streaming or DVB-H television. As mentioned, our group has already work in this field and, in particular, on robust speech transmission for RSR. This way, we will consider first the general problem of speech transmission but, also, we intend to exploit synergies and explore new

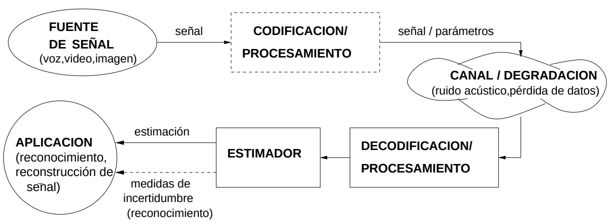


Figure 1: *Estimation and uncertainty processing for different applications.*

applications such as images and, mainly, video.

• Finally, another objective will be the development of techniques which may complement estimation/uncertainty. For example, in transmission applications, source and channel coding techniques not only can be combined with estimation, but also they may be required by the estimator in order to provide useful a posteriori information. Regarding speech recognition, we must also consider complementary techniques (preprocessing, noise detection and modeling) which can support and collaborate with estimation/uncertainty.

The following subsections are devoted to the development of the concepts mentioned above, as well as a review for every considered application.

2.1. Robust Speech Recognition

During the last years, automatic speech recognition (ASR) has shown its utility in different applications, although it is specially promising for human-computer interaction in portable and mobile devices (mobile phones, PDAs, e-books). This fact is reflected in the strategic research agendas of several technology platforms such as eMobility or eMOV. Mobility requires ASR system which must work in noisy environments. Then, robustness against acoustic noise is a crucial issue which is researched by many gropus devoted to ASR. This proposal intends to research this topic employing estimation and uncertainty techniques. In the field of robust speech recognition, the estimation techniques try to recover the speech features when acoustic noise contaminates the speech signal. MMSE estimation has been very successful when combined with stereo data for training, that is, when the same speech signals are available with and without noise, since this allows a suitable modeling of noise. One of the first approaches where SNR-dependent cepstral normalization (SDCN) and codeword-dependent cepstral normalization (CDCN). Since then, other related techniques as multivariate Gaussian cepstral normalization (RATZ), stereo-data linear compensation of environments (SLICE), multienvironment model-based linear normalization (MEMLIN), uncertainty decoding or stereo-data stochastic mapping (SSM) have been developed. All these techniques share the idea of assuming a parametric data pdf (normally Gaussian mixture models, GMM). In a recent work, the research team proposes a first approach to robust ASR based on MMSE and VQ modeling (instead of GMMs) and cepstral normalization. Another purpose of this project is the application of uncertainty processing. As mentioned, the statistical nature of speech recognizers, usually based on HMMs, facilitates the integration of this kind of processing. Uncertainty processing techniques are based on the fact that not all the speech features are equally reliable due to random nature of noise. Therefore, if we can measure their level of reliability, we can also modify the recognition engine to deal with this fact. This way, it is possible to propose an heuristic exponential feature weighting of the HMM observation probabilities according to the input feture reliability. This approach has found also application in RSR systems for channel error mitigation. An alternative is that of considering that every input speech feature is a random variable instead of a deterministic value. Thus, the probability distributions at the recognizer can be modified according to this uncertainty point of view. We must also mention a group of estimation/uncertainty techniques usually gathered under the name of missing data techniques (MD). The stating point of these techniques consists of identifying, usually in the spectrogram, which regions are reliable (speech prevails over noise) and not reliable (noise prevails over speech). As a result, we obtain masks which can be binary (0=not-reliable, 1=reliable) or continuous, where every mask value provides the reliability level of that part of the spectrogram. Main procedures to compute these masks are those based on instantaneous SNR, those based on speech harmonic properties, or those based on intensity level difference in stereo recording. More recently, speech fragment decoding techniques, based on auditory scene analysis have been proposed. Once reliable and not reliable regions are obtained, there are two possibilities for ASR with MD: imputation and marginalization. Imputation is the same as estimation when reliability masks are employed to reconstruct non reliable regions from reliable ones and a priori speech models. These models can be the same as those employed for recognition or simplified models employed only doe feature extraction. Once the clean speech features are estimated, speech recognition is performed as usual. On the other hand, the marginalization techniques are a type of uncertainty processing techniques which manage uncertain data modifying the recognition engine. This way, reliable parts are employed without modifying the recognizer, while the non reliable parts are marginalized by the acoustic models up to the received energy level (under the assumption of speech plus noise energy additivity). Imputation and marginalization have several advantages and drawbacks. Thus, marginalization techniques provide an optimal classification under the assumption of missing data. These techniques can be only applied to speech models trained with spectral features (Mel-filterbank,Gammatone). However, the performance obtained with these features is lower than that obtained with cepstrum. This fact justified the success of imputation versus marginalization, since imputation still allows the use of cepstral features. In spite of this clear advantage of the imputation techniques, they have the problem of a higher sensitivity to mask estimation errors. The study and development of new imputation/marginalization hybrid schemes is one of the objectives of this project. These hybrid techniques would have the advantage of allowing cepstral features along with uncertainty processing at the recognizer.

2.2. Robust Multimedia Transmission Systems

In multimedia transmission systems, robust techniques against channel errors (bit change, packet loss) are usually classified into three types: interactive error control, techniques based on channel coding and error concealment techniques. In this project, we pay particular attention to the latter type, error concealment (EC), which tries to alleviate the possible changes of the transmitted information caused by the transmission channel. The EC techniques employ different approaches, although all of them share the principle of exploiting the multimedia signal redundancies in order to reconstruct damaged or lost data. Repetition, interpolation/extrapolation and estimation are examples of the most important EC techniques. Estimation, similar to interpolation, can employ data correctly received before and after the error interval, although estimation uses explicitly a statistic model of the source and channel in order to reconstruct, which provides a clear advantage. In this proposal, we focus our research on EC by estimation. Also, we consider channel coding techniques that can combine with EC techniques, such as FEC (forward error correction) codes and interleaving, in which the research team has experience. In the following, we review the state of art in EC techniques applied to different multimedia signals, although we also include those channel coding techniques that could cooperate with the first ones.

2.2.1. Speech/Audio Error Concealment

Traditionally, EC techniques applied to coded speech have been oriented to avoid annoying sounds in the synthesized signal. Thus, most of these techniques are based on the substitution of the lost frames and progressive muting. This trend has been also implemented in popular speech coding standards (EFR, AMR, IS-641, G.723.1, G.729). During last years, VoIP telephony has given rise to new techniques designed for increasing the transmission robustness. One of the main problems to face has been the predictive behavior of the existing speech codecs. This implies that the errors can be propagated forward during a period considerably longer than the duration of the errors themselves. Thus, new coding schemes have been proposed in order to limit the error propagation by removing, totally or partially, the prediction. In special, we must remark the success of the iLBC codec that removes the possibility of error propagation by means of an intra-frame coding, although with a considerable increase of bit rate. In other cases, the propagation is limited by means of periodic intra-frames that act as firewalls, emulating the MPEG standard of video. Nevertheless, the most extended codecs are predictive and they are based on the CELP (coded excitation linerar prediction) paradigm. Thus, during the last years, some EC techniques have been proposed for this type of codecs. A first approach is based on the use of frames received longer than the delay imposed by the anti-jittering butter. Although these frames are not usable directly, they can be used in order to resynchronize the excitation and, therefore, remove the error propagation. In some works the backward excitation energy is bounded since it is the main reason of the error propagation. This loss of energy is compensated by means of the innovation codebook (algebraic codebook) or a glottal-pulse codebook is provided in order to encode the first subframe after a loss for voiced sounds. Also, some works employ the Bayesian estimation as EC technique. As aforementioned, the Bayesian philosophy is based on combining the evidence contained in the signal with the a priori knowledge of the probability distribution of the source process. This is achieved by means of the conditional probability of the possible transmitted information given the received one. In our case, the a priori knowledge is will be given by the present redundancies in the signal. Several EC techniques based on Bayesian methodology can be found in the literature. The MMSE estimator (based on obtaining the expected value of the damaged or lost data given the available information) has been extensively used in speech transmission. In particular, the MMSE formulation using hidden Markov models (HMM) as source model has provided excellent results in speech transmission and parameter transmission for remote speech recognition. Other possible Bayesian estimator is the one based on the maximum a posteriori (MAP) that, unlike MMSE, uses the mode of the a posteriori distribution as estimate. Depending on the type of probability distribution, the MAP estimator can be advantageous. In audio signals coded using MDCT is important to stress the use of the maximum likelihood (ML) for the reconstruction of the lost spectral information. One of the most important aspects in our proposal is the combination of FEC codec with EC techniques. This topic has been slightly discussed in audio/speech, possibly, due to the hard delay restrictions imposed by the telephony applications (FEC codes introduce variable delays), although these restrictions are less exigent in streaming applications. Thus, the MMSE estimation of the LSP coefficients is combined with media-specific FEC codes (previous LSPs encoded using a secondary codebook) and introducing a maximum delay of 2 frames. Our group presents previous experience in the use of this type of FEC codes and their combination with EC techniques based on MMSE, as well as the use of suitable interleavers for a Bayesian EC.

2.2.2. Image/Video Error Concealment

First, we review the error concealment in images, since EC techniques for video are often based on those ones corresponding for images. In the field of image transmission, it is common the use of techniques based on interpolation. Thus, in coding schemes based on transform, several techniques carry out an interpolation of the adjacent transformed coefficients to those lost/corrupted ones, or work in an alternative domain such as the wavelet transform. In the spatial domain, it is proposed a pixel interpolation using weights derived from the frontier pixels of the adjacent macroblocks. In the field of Bayesian estimation, the MMSE estimation of lost data (with DPCM coding) has been proposed using the quantized indices received before and after the lost ones. An improvement to the previous techniques is carried out in by means of the use of soft-bits (an error probability is assigned to each bit) that can be complemented with the use of MRF. In a combination of the technique presented with FEC codes is proposed in order to carry out an iterative decoding scheme. The problem of error concealment in video sequences is a generalization of the case for images. Thus, these EC techniques use intra-frame information (image problem) and temporal correlations (inter-frame information). These EC techniques can be classified into spatial methods, temporal methods and mixed ones. Regarding spatial methods, a weighted interpolation for H.26l is proposed, also a scheme based on directional decision and intra-frame prediction is employed for H.264/AVC, and a combination of interpolation with analysis/synthesis of textures is proposed in order to obtain a perceptual optimum result. Also, Bayesian approaches based on MAP estimation and MRF, similar to those ones presented for images. Regarding temporal methods, a simple and effective way of reconstruct the corrupted/lost areas of a image is their substitution by the corresponding areas of the previous frame. However, this approach does not present good results in the case of fast movement or sudden scene changes. Other possibility is the use of boundary matching algorithms for recovering lost/damaged movement vectors. The mixed methods present the advantage of employing all the available correlations in order to provide the lost information. The combination of channel coding and EC techniques is also frequent in the field of Video error concealment. Thus, some works apply a flexible macroblok reordering in order to make easier the concealment task. Intra-coding techniques try to mitigate the error propagation of predictive codecs by means of removing the prediction in some macroblocks. The question here is to determine when intra-encoded macroblocks must be inserted taking into account the EC technique included in the decoder. Other alternative is the use of additional information to help the EC method. Thus, in [Frossard01] the introduction of additional data is proposed in order to resynchronize a video sequence when the loss of MPEG-2 packets reduces the video quality (after applying EC) under a determined threshold. In [Kim01], the EC technique is combined with a coding scheme based on two movement vectors (applied to two different macroblocks) in order to prevent the loss of these references. In [Zhu09] a method of using the redundant frames (media-specific FEC) is proposed for H.264/AVC as EC. Finally, data-hiding techniques offer a way to hide some parameters in the bit-stream in order to help the

EC techniques and, thereby, to achieve a better quality.

3. Project Objectives

The reasons why we consider relevant this proposal and the starting hypothesis that support the project objectives can be summarized in next list:

• Bayesian estimation techniques have provided excellent results in speech transmission and recognition applications. However, we think that that the potential of these techniques has not been fully exploited in the field of robust speech recognition in noisy environments. Thus, current techniques have been more focused on modeling the noise effect over the speech features, but not on providing a whole model of source and channel (degradation by noise). We also consider that the possibilities of the Bayesian techniques must be increased with complementary procedures as noise detection and characterization.

• It can be considered as universal principle that any measuring process is inherently uncertain. Thus, during the last years, several methods accepting this principle have arisen. Robust speech recognition is especially suitable in this framework, given the random nature of the noise, as well as the statistical fundaments of current speech recognizers (usually based on hidden Markov modeling). Therefore, we consider that uncertainty processing techniques deserve more analysis and development as it is proposed in the present project proposal.

• Through previous projects, our research team has acquired a wide experience on robust transmission for remote speech recognition (RSR) systems by means of estimation techniques. We think the translation of this experience to a more general field, as robust multimedia transmission, may be very fruitful. Robust speech transmission is an immediate extension of this experience. Also, we consider that it is particularly interesting to extend also our experience to other fields such as image and, especially, video transmission. We also think that statistical mitigation techniques must be combined with certain channel coding techniques specially adapted to the statistical ones.

3.1. Background And Previous Results

In Section 2, we already summarized the main work which supports the research proposed in this Project. Anyway, we extract below the most relevant contributions with respect to the main points mentioned in Section 3:

Estimation for robust speech recognition. There are a number of recent contributions [1] which show that a suitable selection of the statistical models employed for estimation can provide important performance improvements. The introduction of an additional for the source process has been only considered in recent publication, although has been widely studied by our research team for RSR systems [2, 3, 4].

Uncertainty processing for robust speech recognition. Softdata techniques provide a smart framework where the estimates to be processed by the recognizer are considered random variables instead of deterministic values. However, this previous work only considers Gaussian distributions, what is, in general, false, and leaves space for future improvements. Also, exponential feature weighting, even being a heuristic technique, can provide a better performance than the soft-data approach [4]. Therefore, we think that exponential weighting deserves a more in-depth analysis, applying a more formal view, in order to obtain suitable reliability measures. Finally, missing data (MD) techniques provide a joint framework for estimation and uncertainty processing based on the search of reliable regions over the speech spectrograms. Recent work points out that this point of view has a perceptual fundament which makes the MD approach especially interesting for research over the next years.

Robust multimedia signal transmission. In the case of robust speech transmission, our research team [2, 5, 4], as well as other groups, has already successfully applied estimation techniques for error concealment in RSR systems. Thus, we think that the extension and generalization of these techniques to new speech application (VoIP, videoconference, streaming) is quite timely. In this sense, the combination of estimation with media-specific FEC or interleaving looks very convenient [3]. Finally, although estimation techniques have already been applied to image/video, they have not been applied extensively. A possible reason for it is the computational complexity involved by 2D and 3D statistical models [2]. This issue has already been researched by our research team. Again, the combination of concealment techniques with cannel coding is particularly promising.

3.2. Specific Objectives

Our objectives can be summarized as follows:

1. Robust speech recognition against acoustic noise by means of Bayesian estimation techniques (especially MMSE), focusing on new statistical models for the whole process (speech source, and noisy and clean feature spaces) and on the computational efficiency of the resulting estimators. We include here new techniques for complementing and supporting the Bayesian ones (preprocessing, noise detection and modeling).

2. Development of uncertainty processing techniques for robust speech recognition which may be combined with estimation. We consider here new ways for uncertainty measuring and their incorporation to the recognition engine. We also consider new variants of missing data techniques, seeking the collaboration between imputation and marginalization, as well as the definition of new reliability masks.

3. Development of robust multimedia signal transmission techniques for speech and images/video. Again, we consider very especially estimation techniques for damaged/lost data reconstruction, focusing on the statistical modeling of source and channel. As in objective 1, we include complementary techniques (media-specific FEC, interleaving) which may collaborate with estimation.

3.3. Additional Information

For more detailed information about the project and a complete list of bibliographical references, please visit project web page: http://ceres.ugr.es/tsc/tetitrs/

4. References

[1] J.A. González, A. Peinado, A. Gómez, J.L. Carmona, J.A. Morales: "Efficient VQ-based MMSE for robust speech recognition", in Proc. ICASSP, 2010.

[2] A.M. Peinado, V. Sánchez, J. Pérez-Córdoba, A. Rubio: "Efficient MMSE-based channel error mitigation techniques. Application to distributed speech recognition over wireless channels". IEEE Trans. Wireless Commun., vol. 4 no. 1, 2005.

[3] A.M. Gómez; A.M. Peinado; V. Sánchez; A.J. Rubio; "Combining Media-Specific FEC and Error Concealment for Robust Distributed Speech Recognition Over Loss-Prone Packet Channels", IEEE Trans. Multimedia, Vol. 8, No. 6, 2006.

[4] J.L. Carmona; A.M. Peinado; J.L. Pérez-Córdoba; A.M. Gómez; "MMSE-Based Packet Loss Concealment for CELP-Coded Speech Recognition", IEEE Trans. Audio Speech Lang. Process., Vol. 18, No. 6, 2010.

[5] A. M. Gómez, A.M. Peinado, V. Sánchez, A.J. Rubio: "Recognition of Coded Speech Transmitted Over Wireless Channels". IEEE Trans. on Wireless Communications, vol. 5, pp. 2555-2562, September 2006.

TACOMA: On-line Transcription of Audiovisual Material

Antonio Cardenal, Pablo Peso, Martín Bueno, Alberto Espiña,[1]
Daniel A. Rodríguez-Silva, Lilian Adkinson-Orellana, Alexandre Pellitero-Rivero[2]

[1]Grupo de Tecnoloxías Multimedia, University of Vigo, Spain
{cardenal,ppeso,mbueno,aespina}@gts.tsc.uvigo.es
[2]GRADIANT, ETSI Telecomunicación, Campus Vigo, 36310, Spain
{darguez,ladkinson,apellitero}@gradiant.org

Abstract

TACOMA is a project devoted to the on-line transcription of educational courses which comprises two main systems: a fully automatic and generic subtitling system intended for being applied to any audiovisual material ; and a computer-assisted transcription system intended to facilitate and speed up as much as possible the process of manual subtitling using ASR technologies.

Index Terms: speech recognition, computer-assisted subtitling.

1. Introduction

The project TACOMA [1] *Transcripción on-line de Material Audiovisual* is being developed by the GTM group of the University of Vigo and Krasis Consulting S.L. company , with the participation of Gradiant (Galician Techological Center of Telecommunications).

The main goal of the project is the automatic subtitling of the on-line courses offered by Krasis, (which is a Spanish company devoted to e-learning and e-mail direct marketing), but with the additional objective of developing and make public an on-line universal subtitling application. For these purposes the project involves the developing of two main modules:

- A universal fully-automated subtitling system.
- A computer-assisted transcriber tool.

We are aware of the fact that the objective of achieving an universal subtitling system is very ambitious and quite unrealistic nowadays. Continuous speech recognition systems still offer poor performance when there are mismatches between the statistical models and the recording conditions. To obtain good results in every reasonable condition, it is necessary to use a set of acoustic and linguistic models, covering an extensive range of conditions, which means to capture and classify a large amount of acoustic and text databases. This objective is out of the scope of the project. Here, the objective of universal subtitling may be translated in the provision of mechanisms for acoustic and linguistic model unsupervised adaptation to achieve the best results given the available models.

The computer-assisted transcriber tool is provided to be used in situations in which the automatic system offers poor performance. The idea is to give the user the possibility of taking advantage of automatic speech recognition technology in such scenario, simplifying and speeding up the otherwise expensive and slow process of manual transcription. To achieve this goal, the system uses an interactive feedback strategy, in which the information provided by the user, (basically the corrected subtitles), is employed to improve progressively the acoustic and linguistic models, eventually reaching an ideal point in which no more corrections would be needed.

It should be noted that although our ambition is to develop an universal application, some of the decisions adopted throughout this project have been dictated by the special characteristics of the on-line courses employed. These courses are mainly related with computer programming and web management, so they have a technical, very specialized vocabulary. Also, the courses were recorded while the speaker is interacting with the computer, hence the speech is almost spontaneous, with frequent repetitions, filler words and false starts.

The remainder of the paper is organized as follows. In Section 2 the overall architecture of the system is described. In Section 3 the graphic interface is presented. In Section 4, we describe in deep the recognition module. Finally, in Section 5 we present some conclusions and further work.

2. Description of the architecture

2.1. Service-Oriented architecture

The system is based on a Service-Oriented architecture [1], composed of the transcription module and the application server, both providing a transcription service as shown in figure 1. A web client was implemented to access the transcription services with a simple and powerful interface [2] to manage the generated subtitles.

At the client side, two different parts can be identified: the list of current transcriptions and a player for visualizing the media, with an integrated subtitle editor. The Web Service uses PHP as server-side script language, and includes a database implemented with MySQL. The communication between the client and the application server is performed using REST [3], providing a simple and efficient method for exchanging data. The application server is also in charge of converting the audio and video sent by the client into the appropriate formats: the transcription module needs the audio in a raw file, while the player works with .flv for videos and .mp3 for audio files.

2.2. Commumication between Web Service and transcription module

The transcription module, which will be described more in detail in Section 4, is basically formed by a server containing the recognition system and an internal database. Since this transcription module and the application server are different ma-

[1]This project has been supported by Spanish MITYC program AVANZA I+D under the project TSI-020110-2009-242

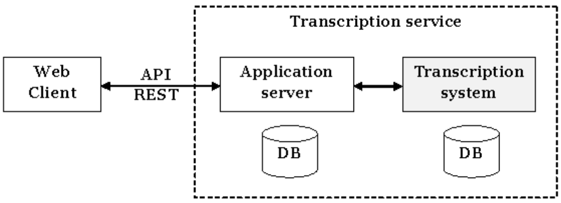


Figure 1: *System architecture overview.*

chines, the communication between them is performed using SSH protocol. This allows exchanging information efficiently in a secure way. Thus, the transcription server has available a set of scripts to manage each transcription and consult its current state.

3. Graphic interface

3.1. List of transcriptions

After logging in, the user can access to a web page where a list of his previous sessions is presented. The user may start a new session or cancel a current process. Two kind of sessions may be selected:

- Automatic, without human intervention.
- Assisted, in which the user may correct the automatic transcription provided by the system.

New transcriptions are started from a web page, in which the user may upload the audio or video file (or indicate an URL) to be transcribed. Additionally, the user can select at this point a trade-off between speed and recognition accuracy. This information will be used by the recognition module to set the number of recognition passes performed, and other minor parameters.

The interface also gives to the user the possibility of including extra information that would be useful in the recognition process. This information includes a text file, an URL pointing to a page with similar topics or a set of keywords. This information is later used by the transcription system to adapt and improve the statistical language and acoustic models.

3.2. Subtitles visualization and edition

After uploading the necessary information to start the transcription, the user will be redirected to a new page containing a player to visualize the audio or video, with a subtitle editor (see figure 2). This player allows the users to see the available subtitles synchronized with the audio or video in order to check the results. Every word is shown following a colour code, based on its correctness confidence, dark colour indicating higher confidence. If the word is suspected to be erroneous it will appear in red, so users can easily identify the wrong words and see if the transcription is being reliable.

If transcription module provides several alternatives for a word transcription, it will appear underlined on the subtitle editor. Making click on this word, a list will appear with the alternatives ordered from higher to lower confidence, so the user can choose the most appropriate. If the correct transcription does not appear in the list, the user has the possibility for typing it. In the case of assisted transcription, if the user considers that the subtitles received at certain stage have achieved enough quality, he can decide to change to automatic mode.

Finally, when the process finishes, the user may download the subtitles in several formats including .srt, TimedText, Subviewer 2.0, Encore, Google Video or JW FLV till 3.8. As we will explain in following sections, the corrections made by the user will be sent to the transcription server to adapt the statistical models to improve the accuracy of the following transcriptions.

4. Speech Recognition module

The recognition module is based on the recognition engine developed by the GTM group [4] in previous projects. This decoder is based on two stages: (1) a Viterbi algorithm which works in a synchronous way with a beam search; and (2) an A^* algorithm. This recognizer was developed for large vocabulary continuous speech recognition applications.

4.1. System initialization

Once the audio file is received, the recognition module performs the following tasks:

- The selection of acoustic and linguistic models.
- The segmentation of the audio file using a simple voice activity detector.
- The audio file parameterization, using standard MFCC_E_D_A parameters with cepstral mean extraction.

The acoustic model selection is very simple. A phonetic recognizer is applied to the first frames of the file using each candidate model. The model which provides best acoustic score is then selected. Our experiments show that this simple mechanism provide good enough results. However, this strategy is valid if the audio file contains recordings of a single speaker, which is true for the on-line courses employed, but false in general. A speaker segmentation module will be incorporated in the future.

The VAD uses a simple algorithm based on energy thresholds and a small state machine. The mean length of the VAD segments may be selected by software, but our experience shows that a length of approximately one minute provides good trade-off between speed and recognition accuracy. This initial segmentation will be used throughout the remainder of the session for presentation and recognition purposes.

4.2. Speech recognition

Once segmented, each VAD section is recognized using a multipass strategy. In each pass an unsupervised acoustic adaptation is performed for each recognized segment. The procedure for the first pass is as follows (Figure 3:

1. Using the word-level transcription of the segment, a model-level alignment is obtained.
2. An MLLR+MAP adaptation of the HMMs is then performed, using the transcription of all segments processed until now.
3. The new models are used to recognize the next segment.

With this algorithm, the first segment is transcribed using the unadapted models, the second segment is transcribed using the models adapted with the the first segment, and so on.

Figure 2: *Subtitle player with editor.*

For the second and successive passes, the procedure is similar. The main difference is that at point 2, the acoustic adaptation is performed using the alignments of pass 1 for all segments, so the first segment is recognized using acoustic models adapted with the transcription of the whole audio segment.

The main motivation of this procedure is to provide enough feedback to the user, avoiding annoying waiting periods with no response of the system, while maintaining the best possible performance.

Our experiments show that this strategy is very effective when there is a severe acoustic mismatch. As an example, in a test with an initial 63% of WER, the algorithm provides a 48.3 % of WER in the first pass, and 43.38 % if a second pass is performed. As a comparison, a 48.0% of WER is obtained in the same experiment if the audio file is recognized with an HMMs adapted using the transcription of all segments.

4.3. Computer-assisted module

The workflow of the computer-assisted transcription module is very similar to the described in the previous section. The strategy of progressive model adaptation is also employed here. The main difference is that the user-corrected transcriptions are now used when available, substituting the automatic recognized word sequence. As can be expected, the adaptation is better, providing an increase of performance.

However, the availability of the correct transcription allows also the use of language model adaptation mechanisms but also produce several new problems that should be addressed.

The main problem to address is related with the phonetic transcription of the new words. The acoustic adaptation process requires the use of a model-level alignment of the audio segment. To perform this task, a forced alignment is also needed. This forced recognition may easily fail if the manual transcription is not accurate enough. The most common cause is the use of foreigner (non-Spanish) which are phonetically transcribed using Spanish rules . This problem is very common in the online courses employed in the project, which are mainly related with software programming and web management. A simple example could be the word ”windows”, which using the Spanish phonetic rules would be transcribed as ”vindovs”. Two main mechanisms are used to overcome this problem: the use of lists of word exceptions, and an automatic detector of foreign words.

The use of a list of exceptions, including several alternative pronunciations of the most common foreigner (mostly English) words used in Spanish, is straightforward, but obviously very limited. To identify foreign words we use the simple algorithm of detecting some consonant groups not used in Spanish (*th, wr, ch*, etc). If a word is detected as non-Spanish and is not in the exception file, the user is prompted to provide an approximate Spanish pronunciation. The word is then added to the exception list to be used in following sessions.

4.4. Language model adaptation

The second use of the user-corrected transcriptions is to improve the language modeling, for example performing topic adaptation, identifying out of vocabulary words (OOvs), etc. At the time of writing this article the mechanisms for LM adaptation and OOv identification have not been included in the prototype, and are still under development. However, the idea is to use mixture language models and topic-based language models [6]. The LM adaptation will be used in two stages of the recognition process: when the transcription session starts, using the optional information provided by the user (http links, text files or keywords) and when an user-corrected transcription is available. In both cases the tasks to perform are very similar:

- The vocabulary of the new material must be extracted

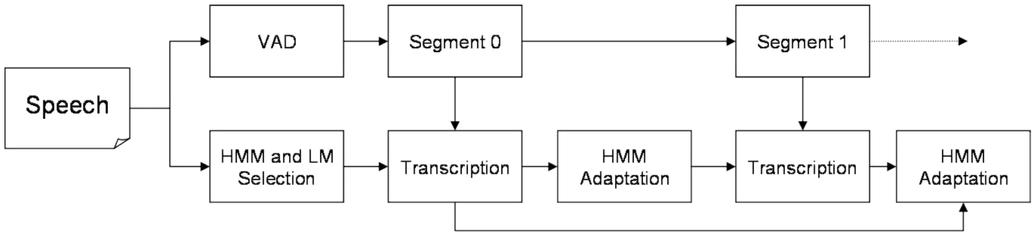


Figure 3: *Unsupervised acoustic adaptation in TACOMA.*

and if necessary, included on the LM. To perform this task, is essential that the new words appear on the text database. Otherwise the training of the new LM is not possible. We don't know any strategy to overcome this particular problem.

- The new text is used as objective for mixture-based LM adaptation, or
- The new text is used to select a topic-based LM.

The first two mechanisms are under development and are expected to be included in the system soon. In an open-task application such this one, the third mechanism requires an unrealistic amount of data, although is been considered, does not seem very promising.

In any case, the LM adaptation mechanisms require the reconstruction of the model for each new transcription received. Unlike the acoustic model adaptation, which is a reasonably fast procedure, the LM adaptation is computationally expensive and time consuming. This is not a problem with the automatic subtitling system, in which the LM adaptation is applied only once at the beginning of the session, but it may be an issue in the computer-assisted system, in which the adaptation should be performed for each corrected transcription received.

4.5. Confidence measures and N-Best lists

Confidence measures are an important part of the system, specially in the computer-assisted system since are used to warn the user from potential transcription errors. We believe that this kind of mechanisms are very important for speeding up the manual transcription task.

The confidence measures employed in this project are based on word posteriori probabilities extracted using word lattices [5]. Unlike most applications we need a continuous confidence measure for employing the colour code explained in Section 3.

The confidence measures are not only computed for the main transcription, but also for the N-Best lists. This is, however, a common mechanism when using word lattices. This information is employed, as was explained in Section 3, to implement an ordered drop-down list on the client side, allowing the user to easily select the right word.

4.6. Conclusions and further work

In this paper we have presented a prototype of the subtitling system developed in the project TACOMA. It is composed of two modules: an automatic subtitling system and a computer-assisted transcriber. The most novel part of the project is the computer assisted transcriber, which is conceived to speed up the usually time-consuming task of manual transcription. The key points of the system are the following:

- The audio file is segmented using a voice activity detector.
- Each audio segment is automatically transcribed, using the result to perform an adaptation of the acoustic models.
- The initial transcription is presented to the user with the confidence of each word indicated by a colour code.
- The user may access an ordered list of alternative transcriptions of each word for correcting purposes.
- The corrections of the user are sent back to the recognition system for acoustic and LM adaptation, improving the transcription of the next segments.

Eventually a point should be reached in which the transcription would be accurate enough to perform an automatic subtitling of the remainder data with reasonable confidence.

The project TACOMA is still under development. In the near future we plan to incorporate some mechanisms of LM adaptation which are been tested at this time. The next improvement will be the incorporation of a speaker diarization and classification module, for dealing with multi-speaker recordings. This module is also under development in our research group [7].

5. References

[1] R. Perrey and M. Lycett, "Service-Oriented Architecture", Proc. 2003 Symp. Applications and the Internet (SAINT 03), IEEE CS Press, 2003, p. 116.

[2] Accessible Rich Internet Applications (WAI-ARIA) 1.0, W3C Working Draft 16 September 2010. Online: http://www.w3.org/TR/wai-aria/, accessed on Oct 7, 2010

[3] Pautasso, Cesare; Zimmermann, Olaf; Leymann, Frank (2008-04), "RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision", 17th International World Wide Web Conference (WWW2008) (Beijing, China)

[4] A. Cardenal-Lopez and F.J. Dieguez-Tirado and C. Garcia-Mateo, "Fast LM look-ahead for large vocabulary continuous speech recognition using perfect hashing". Proc. ICASSP 2002

[5] F. Wessel and R. Schlter and K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition",In IEEE Signal Processing Letters, pages 288-298, num. 3, vol. 9, 2001.

[6] J Dieguez Tirado, C. García Mateo, A. Cardenal, "Effective topic-tree based language model adaptation", Proc. Interspeech 2005.

[7] Laura Docío Fernández, Paula López Otero, Carmen García Mateo. "An adaptive Threshold Computation for Unsupervised Speaker Segmentation". Proc. Interspeech 2005.

A prototype of a spoken dialog system based on statistical models

E.Segarra, L.Hurtado, J.A.Gómez, F.García, J.Planells, J.Pastor, L.Ortega, M.Calvo, E.Sanchis

Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain
{esegarra,lhurtado,jon,fgarcia,xplanells,jpastor,lortega,mcalvo,esanchis}@dsic.upv.es

Abstract

We present in this paper a prototype of a spoken dialog system. One of the characteristics of this system is that most of the modules (speech recognition, understanding and dialog manager) are based on statistical models. The system has the possibility of easily change the task or the language by means of interchanging the different modules. We present in this case a dialog task consisting of booking of sports facilities in the University.

Index Terms: spoken dialog system, statistical model

1. Introduction

The development of spoken dialog systems is a complex process that involves the design, implementation and evaluation of a set of modules that deal with different knowledge sources.

Currently, some of the most successful approaches are based on statistical models estimated using training corpus. Statistical models have been widely used in speech recognition, language modeling and speech understanding. Although in the case of Dialog Managers most of the approaches are based on the manual design of the system behavior, over the last few years, approaches that use statistical models to represent the dialog manager have also been developed ([1], [2], [3], [4]).

In this field, we have recently developed an approach to manage the dialog using statistical models that can be learnt from a data corpus. This work has been applied within the EDECAN-SPORT [5] domain that consist of a task for the booking of sports facilities in the University. In order to increase the reliability of the system, some features were included in the system:

- Unlike classical slot filling tasks, our Dialog Manager takes its decisions not only based on the previous user turns but also considering the information supplied by the Application Manager and it can perform both actions: to provide information and to modify the application data (i.e. after booking or cancelling a court).
- In some cases the system must give many data to the user that are difficult to provide by speech in a concise way. That is the case of offering some courts in some days of the week. In order to convert this in a more friendly communication process a multimodal input/output is available in the system. Therefore in some cases the system gives the information by using a graphical interface, or the user can use the voice or the touchscreen
- In order to take into account some information about the preferences of the user, previous dialogs are saved and its information is used as a priori knowledge.

The prototype has been installed in a kiosk to be accessible to the users in some areas of the University. At the moment it is installed in our laboratory to evaluate its behavior.

2. The spoken dialog system

In figure 1 a scheme of the dialog system developed is presented. The system has been implemented using the architecture defined in the SD-TEAM project [6]. This architecture allows the integration, substitution and collaboration of the modules even if they are located in different computers. The system contains the habitual modules of a Dialog system, ASR, Language Understanding, Dialog Manager, Answer Generator and TTS, besides the specific modules related to the Application Manager and the multimodal user interface. The Application Manager controls the access to the database, not only to provide information but also to modify it when booking or cancellation must be done. In the case of multimodality, we have include two possibilities: the user has a touchscreen to select an item, and the system can give some informations in term of tables. It must be noted that both modules speech understanding and dialog manager are based on statistical models learnt from training samples, as it is described in next sections.

3. The training corpus

In order to design the system, we firstly analyzed human-human dialogs provided by the sports area of our university, which have the same domain that the defined for the EDECAN-SPORT task. From these dialogs we defined the semantics of the task in terms of dialog acts for both the user utterances and system prompts and we subsequently labeled these dialogs. Thus, we had a very small initial corpus for the EDECAN-SPORT task. From this small corpus we learned a preliminary version of the dialog manager. Then we acquired a training corpus by means of a Wizard of Oz technique [7], as it is shown in figure 2. The special characteristic of this acquisition is that we used the preliminary dialog manager learnt from the human-human corpus, and two Wizard were used: one of them for the understanding process and the other to supervise the dialog manager. The reason of this approach is to better simulate a human-machine interaction.

Using this approach a set of 240 dialogs was acquired for our task(a total of 18 different speakers from different Spanish regions). The languages involved in the acquisition were Spanish, Catalan and Basque. A set of 15 types of scenarios were defined in order to cover all the possible use cases of the task. The information available for each dialog consisted of four audio channels, the transcription of the user utterances (with an average of 5.1 user turns per dialog and 6.7 words per user turn) and the semantic labeling of the user and system turns.

Once the corpus was acquired a semi-automatic annotation process was performed. For the user turns (the set of user dialog acts) we defined four task-dependent concepts (*Availability*, *Booking*, *Booked*, *Cancellation*), three task-independent concepts (*Affirmation*, *Negation*, and *Not-Understood*) and six attributes (*Sport*, *Hour*, *Date*, *Court-Type*, *Court-Number*, and

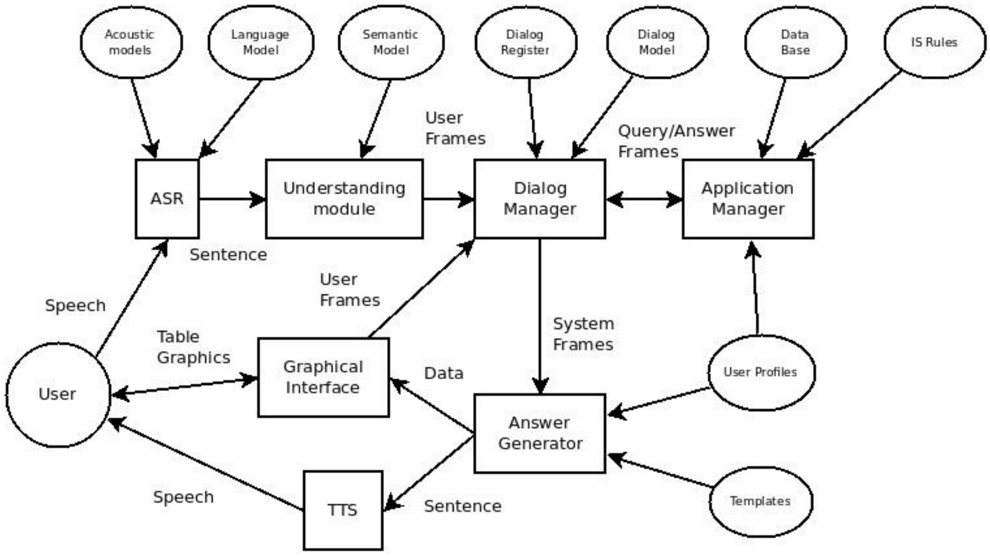


Figure 1: *The EDECAN architecture.*

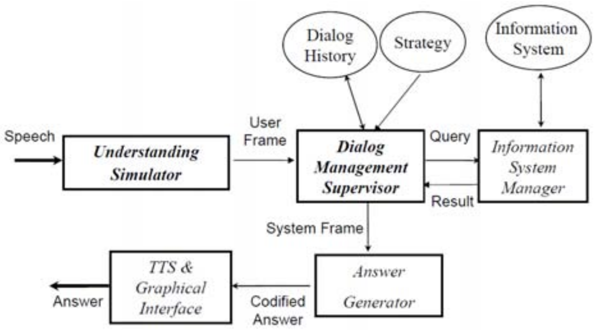


Figure 2: *Acquisition schema using a Wizard of Oz technique.*

Order-Number).

An example of the semantic interpretation of an input sentence is shown below:

I want to book a basketball court for tomorrow.

Semantic Representation:

(Booking)

Sport: basketball

Date: tomorrow

The labeling of the system turns in terms of system dialog acts is similar to the one defined for the user turns. A total of 21 concepts were defined: Task-independent concepts (*Opening* and *Closing*); concepts used to inform the user about the result of a specific query (*Availability*, *Booking*, *Booked*, and *Cancellation*), concepts defined to ask for the attributes that are necessary for a specific query (*Sport*, *Date*, *Hour*, *Court-Number*, and *Court-Type*), concepts used for the confirmation of concepts (*Confirmation-Availability*, *Confirmation-Booking*, *Confirmation-Booked* and *Confirmation-Cancellation*), and attributes (*Confirmation-Sport*, *Confirmation-Date*, *Confirmation-Hour* and *Confirmation-CourtType*).

An example of the labeling of a system turn is shown below:

To play basketball tomorrow, there are two courts: court number 3 at 10:00 and court number 1 at 16:00. Please choose one.

Semantic Representation:

(Booking-Choice)

Sport: basketball

Date: tomorrow

Hour: 10:00 16:00

Court-Number: 3 1

This annotated corpus was used to learn the language models, semantic models and Dialog Manager models.

4. The speech recognition module

The SD-TEAM architecture allows the integration of multiple ASR modules. We are using, in an interchangeable way, both Loquendo ASR and the one developed in our laboratory.

The latter is HMM based, uses a standard speech preprocessor, a n-gram language model and a Viterbi-based search. The speech signal is pre-emphasized by means of a high-pass FIR filter $H(z) = 1 - 0.95z^{-1}$ and then pre-processed to obtain a sequence of frames or acoustic vectors. A 20 ms Hamming window is applied every 10 ms to obtain each frame, which contains 39 parameters: Energy, the first 12 MFCC, and their first and second derivatives. The HMM were trained by means of HTK from the Albayzin Spanish corpus. Albayzin is a phonetically balanced corpus consisting of six hours of speech [8].

5. The understanding module

We propose an understanding process [9] that works in two phases (see figure 3).

The first phase consists of a transduction of the input sentence in terms of an intermediate semantic language. In the second phase, a set of rules transduces this intermediate representation in terms of frames. As the intermediate language is close to the frame representation, this phase only requires a small set of rules to construct the frame. This second phase consists of the following: the deletion of irrelevant segments of the input sentence, the reordering of the relevant concepts and attributes that appeared in the user sentence following an order which has been defined a priori, the automatic instantiation of certain task-dependent values, etc. This last action consists of

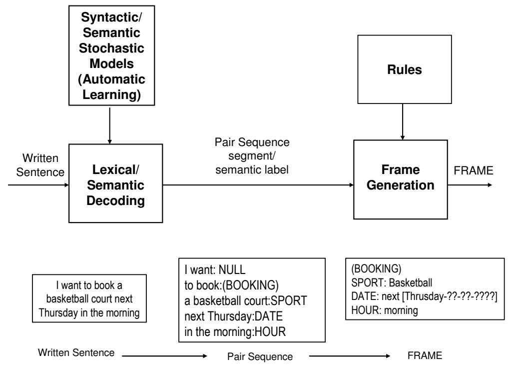


Figure 3: *Understanding module diagram.*

the conversion of dates and hours into their canonical values. For example, “on September the 15th” into “[2010-15-09]”.

For the intermediate language 14 labels were defined, corresponding to the so-called semantic units: *NULL*, *(AVAILABILITY)*, *SPORT*, *(REJECTION)*, *(BOOKING)*, *(ACCEPTANCE)*, *DATE*, *HOUR*, *COURT-NUM*, *(BOOKED)*, *(CANCELLATION)*, *ORDER-NUM*, *COURT-TYPE*, and *NOT*. The goal of the first phase is to find the best sequence of semantic units given the input sentence and a two-level statistical semantic modelization is used in this phase. Figure 4 shows the statistical semantic model. The meaning of each sentence is represented as a sequence of semantic units, and it is associated a segmentation of the sentence in terms of the corresponding semantic units. From an annotated training corpus we learn two kind of models: one of them represents the concatenations of semantic units, and the other represents the lexical realization of each semantic unit (that is, the model of segments of words associated to each semantic units). In both cases the models used for this task are bigrams, that is, both bigrams of semantic units and of words into each semantic unit. The decoding process consists of a Viterbi search over the integrated network, that supplies not only the best sequence of semantic units but also the segmentation of the input sentence associated to it. This segmentation is used in the second phase of the semantic module to associate the values to the attributes (after a normalization process, if necessary).

In other words, given the input sentence $w = w_1 w_2 \cdots w_n \in W$, the process consists of finding the sequence of semantic units $v = v_1 v_2 \cdots v_k \in V$ which maximizes the probability:

$$\hat{v} = \operatorname*{argmax}_{v} P(w|v)P(v)$$

The term $P(w|v)$ is the probability of the sequence of word w given the sequence of semantic units v. We estimate this probability (following the Viterbi algorithm) as the maximum for all segmentations of w in $|v|$ segments.

$$P(w|v) = \max_{\forall l_1, l_2, \ldots, l_{t-1}} \Big\{ P(w_1, \ldots, w_{l_1}|v_1) P(w_{l_1}) \cdot$$
$$P(w_{l_1+1}, \cdots, w_{l_2}|v_2) \cdot \cdots \cdot P(w_{l_{k_1}+1}, \ldots, w_n|v_k) \Big\}$$

If bigram models are used, the probability of each segment given the associated semantic unit is:

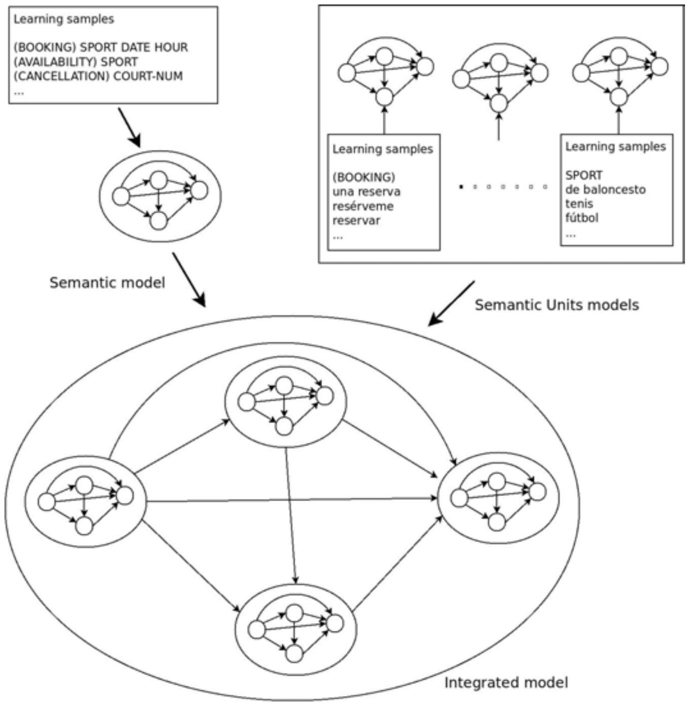


Figure 4: *Understanding model.*

$$P(w_i, \ldots, w_j|v_s) = \prod_{k=i}^{j} P(w_k|w_{k-1}, v_s)$$

and term $P(v)$ is the bigram probability of the sequence v.

$$P(v) = \prod_{i-1}^{k} P(v_i|v_{i-1})$$

6. Dialog Manager

The dialog model proposed in [10] is based on the transduction concept and on the use of Stochastic Finite-State Transducers. In other words, given a state of the system and a user turn, a system turn is generated and a transition to a new state is done. Therefore, dialog management is based on the modelization of the sequences of system and user dialog turns pairs. Thus, a dialog describes a path in the transducer model from its initial state to a final one.

In a dialog system, the Dialog Manager (DM) is the module devoted to choose the best system answer according to its dialog model during the dialog sequence. We consider a dialog as a sequence of pairs $(u_i, a_i), i = 1 \ldots n$, where u_i is the user utterance at time i and a_i is the answer of the system to this utterance. The system answer a_i is selected taking into account not only u_i but also all the information provided by the user throughout all the dialog sequence. All this information (in terms of concepts and attributes) is stored in a data structure we call Dialog Register (DR).

We have developed a statistical DM based on the use of a Stochastic Finite-State Transducer (SFST). A SFST is defined formally by a 6-tuple $(Q, \Sigma, \Delta, q_0, p, f)$. In our approach the input alphabet represents all the allowed user utterances and the output alphabet includes all the system answers defined for

Figure 5: *The kiosk used for the testing.*

Figure 6: *Screenshot of the visual information for the user.*

the task, that is, the set of system dialog acts. $p(q, u, a, q_0) = Pr(u, a, q_0|q)$ is the transition probability from q to q_0 by observing u and emitting a. In our proposed approach, the selection of the best next system answer at time i (a_i) is made by means of the following local maximization:

$$\hat{a_i} = \operatorname*{argmax}_{a_i \in \Delta} p(q_{i-1}, u_i, a_i, q_i) = \operatorname*{argmax}_{a_i \in \Delta} Pr(u_i, a_i | q_{i-1})$$

The dialog ends when a final state q_f , is reached. From this point of view, a dialog can be seen as a path in the transducer from the initial state q_0 to the final state q_f .

7. Multimodal interface

The way to generate the system answers is template-based. That is, the Dialog Manager generates an answer frame that is supplied to the Answer Generator module. This module, using some predefined templates generates the sentence for the TTS. Also, as said before, some informations are presented in a table on the screen (see figures 5 and 6). This process is controlled by the Graphical Interface module that also manages the touch screen that can be used as used input for the Dialog Manager.

8. Conclusions

In this paper we have presented a dialog system based on statistical models that has been applied to a task of booking sport facilities in the University. The system has been implemented using a flexible architecture defined for the SD-TEAM project. This prototype shows that it is possible to build complete dialog systems based on statistical models at different levels. One advantage of this modelization is that the statistical models can be dinamically trained when real users interact with the system. We have also explored the possibility of combining multimodality in the input as well as in the output.

9. Acknowledgements

Work partially supported by the Spanish MEC and FEDER under contract TIN2008-06856-C05-02.

10. References

[1] J. Williams and S. Young. 2007. "Partially Observable Markov Decision Processes for Spoken Dialog Systems". In Computer Speech and Language 21(2), pages 393-422.

[2] F. Torres, L. F. Hurtado, F. Garca, E. Sanchis, E. Segarra. 2005. "Error handling in a stochastic dialog system through confidence measures". SPEECH COMMUNICATION (ISSN 0167-6393). Volume 45. Pages 211-229.

[3] D. Griol, L. F. Hurtado, E. Segarra, and E. Sanchis. 2008. A statistical approach to spoken dialog systems design and evaluation. In Speech Communication, volume 50, pages 666-682.

[4] David Griol, Llus F. Hurtado, Encarna Segarra, Emilio Sanchis. 2008. "A Statistical approach to spoken dialog systems design and evaluation". SPEECH COMMUNICATION. Volume 22. Pages 666-682.

[5] E. Lleida, E. Segarra, M. I. Torres, and J. Macías-Guarasa. 2006. EDECAN: sistEma de Diálogo multidominio con adaptación al contExto aCústico y de AplicacióN. In IV Jornadas en Tecnologia del Habla, pages 291-296, Zaragoza, Spain.

[6] R. Justo, M.I. Torres, E. Lleida, E. Sanchis, R. de Cordoba, J. Macias-Guarasa. 2010 "SD-TEAM: Tecnologas de aprendizaje interactivo, autoevaluacin y multimodalidad en sistemas de diálogo hablado multidominio". Procesamiento de Lenguaje Natural n.45 pp. 331-332, Sept, 2010.

[7] F. Garcia, L.F. Hurtado, D. Griol, M. Castro, E. Segarra, and E. Sanchis. 2007. Recognition and Understanding Simulation for a Spoken Dialog Corpus Acquisition. In TSD 2007, volume 4629 of LNAI, pages 574-581. Springer.

[8] Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J. B. and Nadeu, C. 1993. "Albayzin Speech Database: Design of the Phonetic Corpus", in Proceedings of Eurospeech, 1993, volume 1, pages 653–656. Berlin (Germany).

[9] L. F. Hurtado, E. Segarra, F. Garcia, E. Sanchis. "Language Understanding using n-multigram Models". 4th Internacional Conference EsTAL. Lecture Notes in Computer Science LNAI 3230 (ISSN 0302-9743).

[10] L.F. Hurtado, J. Planells, E. Segarra, E. Sanchis, D. Griol. "A Stochastic Finite-State Transducer Approach to Spoken Dialog Management". In Proc of Interspeech'2010, pag. 3002-3005, 2010.

Introducing Non-Standard Luso-African Varieties into the Digital Domain

Simone Ashby, Sílvia Barbosa, and José Pedro Ferreira

Instituto de Linguística Teórica e Computacional (ILTEC), Lisbon, Portugal
{simone, silvia, zpferreira}@iltec.pt

Abstract

In this paper, we describe the work of the LUPo project to integrate non-standard Luso-African varieties from Cape Verde and Mozambique into a rule system for generating accent-specific phonetic transcriptions for these and other spoken variants of Portuguese. Here, we present a description of LUPo's functions for online use, LUPo's architecture, and the manner of linguistic data collection and analysis that support this application. Implications for a subsequent text-to-speech (TTS) module are also presented in terms of yielding high-quality pronunciation lexica for regional variants of Portuguese, and towards facilitating the entry of hitherto untreated speech varieties into the digital domain.

Index Terms: pronunciation lexicon, speech synthesis, Luso-African speech varieties, digitally endangered speech varieties

1. Introduction

This is a demonstration of the LUPo online prototype, as well as a presentation of some preliminary results for observing non-standard Luso-African pronunciation varieties from Cape Verde and Mozambique. This work marks the first phase of a three-year research project dedicated to the creation of an accent-independent lexicon and rule system for generating accent-specific pronunciations for regional variants of Portuguese. More in-depth information about LUPo and the English Unisyn Lexicon upon which it is based can be found in [1] and [2], respectively.

Here, we present a seminal effort towards developing systematized, multiple, complete phonetic models for non-standard varieties of Portuguese, as it is actually spoken in different parts of the world. We developed broad phonetic segmental models[1] based on six individual speakers, or *idiolects*: three from different locations on the island of Santiago (the largest in the Cape Verdean archipelago), and three from different locations in Mozambique.

The presentation provides a window into these segmental models, as contrasted with those for the São Paulo and standard Lisbon dialects. A selection of phonological rules is also presented, along with a description of how one of LUPo's key components, the regional accent hierarchy, enables the sharing of rules across pluridimensional dialectal and sociolectal varieties. Finally, we demonstrate the LUPo system as it currently exists, and provide some preliminary results for observing sub-national and national varieties of Cape Verdean Portuguese.

The motivation for this research is based around the development of high-quality pronunciation lexica for a pan Lusophone TTS system. As speech technologies become an increasing part of our everyday lives, the users of these technologies represent an ever widening speaker base. Thus, adapting such technologies to a wider number of speakers – and *topolects* – and representing countries and regions for whom such development concerns have been largely overlooked carries significant economic and political weight in narrowing the global digital divide, and promoting further research among lesser studied varieties.

Through the establishment of a linguistically derived rule system for the explicit treatment of allophones within and across regional varieties, LUPo circumvents the cost of producing high-quality phonetic transcriptions by hand, while attracting a wider pan Lusophone audience to the lexical database in which it resides, and providing the research community with a vast resource of Portuguese accent data for evaluating speech applications and testing diachronic, phonological and sociolinguistic theories.

2. Background

Portuguese is a pluricentric language spoken by one-fifth of the world's population, and with regional variants spanning Africa, Asia, Europe, and South America. In addition to Brazil and Portugal, Portuguese is a recognized official language in Angola, Cape Verde, East Timor, Equatorial Guinea, Guinea-Bissau, Macau, Mozambique, and São Tomé and Principe.

According to Alan Baxter, one of the leading scholars in the study of Portuguese-based creoles, there is a growing sense of identity among Luso-African speakers concerned with elevating the level of prestige associated with local varieties of Portuguese. Baxter also cites evidence to suggest that BP may have increasing role in affecting some of the Portuguese varieties spoken in Africa [3]. Unfortunately, and by Baxter's own admission, a dearth in the literature concerning varieties of Portuguese spoken in Africa makes these claims difficult to assess.

In the interest of promoting pluridimensional studies of Portuguese, this work aims to make a meaningful contribution to the understanding of the different sound systems featured among Luso-African varieties.

2.1. Cape Verdean Portuguese

Cape Verdean Portuguese is distinct from Mozambique in that its speakers have a Portuguese based creole, Kabuverdianu, as their mother tongue. The latter is used in everyday communication, while Portuguese is used in the media, education, and official communications. Roughly one-third of the population of Cape Verde is considered fluent in Portuguese; more than half possess a basic understanding of the language [4].

As an archipelago consisting of 10 islands[2], Cape Verde is host to a number of Kabuverdianu and Portuguese accents and dialects. Typically, these varieties correspond to a division of the Windward islands in the north and the Leeward islands in the south. However, a sociolinguistic shift dividing the western islands from the eastern half appears to be on the rise [5].

[1]We also intend to treat cross-word phenomena, such as external sandhi, as part of the LUPo project. Acoustic modeling and suprasegmental feature descriptions will be undertaken in the follow-up TTS project.

[2]All but one island are inhabited.

2.2. Mozambican Portuguese

While Portuguese is the official language of Mozambique, it generally spoken as a second language, after one of a variety of local languages and creoles. Portuguese is spoken as a lingua franca by more than one-third of the population. A very small minority of Mozambicans speak Portuguese at home.

Speakers commonly refer to European Portuguese (EP) as the norm. Nevertheless, the influence of indigenous languages, and the prestige many Mozambicans associate with Brazilian Portuguese (BP) varieties, along with a steady supply of Brazilian television shows, have resulted in the emergence of distinctively Mozambican ways of speaking Portuguese [6].

3. Multidialectal speech technologies

3.1. Portuguese systems

As Portuguese speech technologies continue to advance in the treatment of “standard” BP and EP varieties, efforts to develop computational models of *non-standard* varieties from these and other countries in the world where Portuguese is spoken are conspicuously absent.

One of the more relevant studies to date concerns an accent verification system for Portuguese, based on acoustic, phonotactic, and prosodic cues, for recognizing BP, EP, and what the authors refer to as African Portuguese (AP) varieties [7]. Sub-national differences are not evident due the selection of corpora used in this and related experiments. Two of the experiments, however, examine the ability of the language/accent verifier to distinguish Portuguese varieties from Angola, Cape Verde, Guinea-Bissau, Mozambique, and São Tomé and Principe[3].

The authors of [7] report a limited ability on the part of their system to cope with different accents of Portuguese, with some – the African varieties, in particular – resulting in the degraded performance of their speech recognizer. A separate experiment in which African varieties were bundled into one category revealed a correct identification rate of just 42% for this class. Misidentification of the AP class was also found among the study's human participants, representing Angola, Brazil, Cape Verde, Mozambique, and Portugal.

What these findings show, although not explicitly stated by the authors, is that the training of speech recognizers on BP or EP corpora is insufficient for handling non-standard input. Further, the results in [7] show that bundling Luso-African varieties together not only leads to an extremely low accent verification rate, it is a step in the wrong direction towards creating speech technologies that can cope with these and other regional variants.

3.2. Other models

For a handful other languages, such as Arabic [8], English [2], German [9], Irish [10], Mandarin [11], Romani [12], and Spanish [13], non-standard lectal variants are slowly attracting the attention of speech technologists. However, *slowly* is the operative word, considering the potential value posed by dialectal and sociolectal models for improving spoken dialog systems, enhancing the training of speech recognizers, and creating more “natural” sounding synthetic speech.

[3]It is important to note that the corpora used in these experiments have been drawn exclusively from RTP Africa television news media, for which it can be assumed that journalists and presenters use a marked EP variety, despite that EP speakers were eliminated from this portion of the variety verification corpus.

4. LUPo

By taking an integrative approach and focusing not only on the recognized centers of Brazil and Portugal, but also on the sub-national dialectal and sociolectal varieties from locations around the globe, LUPo aims to create pronunciation models for as many regional variants of Portuguese as the project's time and resources allow.

In section 4.3, we describe the future accessibility of LUPo via the existing free, online lexical knowledge base, the *Portal da Língua Portuguesa*, hereafter referred to as the 'Portal'. A separate project deliverable will be the subsequent release of a free, searchable, online database containing all of LUPo's data and rules.

4.1. Architecture

LUPo's core components include: an exceptions dictionary, an accent-independent master lexicon of underspecified pronunciations (including part of speech and frequency information), a regional accent hierarchy, and the application (through Perl scripts) of morpho-phonological rules that transform the master lexicon pronunciation into the the target output.

4.1.1. Regional accent hierarchy

The model for the regional accent hierarchy is based on that of the original English Unisyn Lexicon [2]. It is made up a system of files containing variant specifications and rule scores. Applying an example from Mozambique (see figure 1), the first set of lines is an entry in the file 'lupo_towns', with 'map' representing the capital city of Maputo, and the next set of abbreviations representing a system of levels that correspond to COUNTRY, REGION, TOWN, and PERSON.

The next set of lines in figure 1 is taken from a file called 'lupo_scores', wherein a general rule is attributed at the country 'CNY' level for the deletion of word-final /r/ across Brazilian and Mozambican varieties.

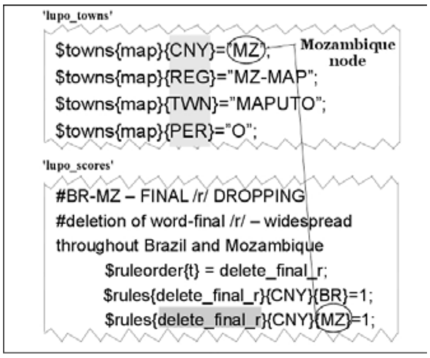


Figure 1: Sample of regional accent hierarchy.

What is interesting about this hierarchical system is the inheritance by each node of features from the previous node, provided the inheritance is not broken by the introduction of a competing feature (or features) at a lower level. As the lowest level in the hierarchy, rules attributed at the person 'PER' level override competing specifications from all the higher levels. Thus, by adding a relatively small number of rules to 'lupo_scores', we can model the unique segmental features that characterize a mesolectal variety of young urban speakers, or even that of an individual – say Mozambique's current president Armando Emílio Guebuza – while implicitly treating the remaining set of allophones as inherited features from the upper nodes TOWN, REGION, and COUNTRY.

In terms of the specific rules presented in figure 1, it is worth noting that word-final /r/ undergoes elision by the Mozambican informants, a phenomenon that is widespread throughout Brazil. Thus, we see possible evidence for the influence of BP on Mozambican varieties.

4.1.2. Rule system

The system stores allophonic rule sets that exploit morphological boundaries to express different accent-specific rules, most of which are post-lexical. Encoding morphology in the pronunciation rules enables the system to identify the correct pronunciation in opaque orthographic contexts, such as the assignment of EP vowel height in the lexically related word pairs 'm[o]lho' and 'm[u]lhada', and 'm[ɔ]lho' and 'm[ɔ]lhada'. Perl scripts are then used to apply rules to the master lexicon and generate accent-specific output. A closer look at the rules is presented in section 4.3.

4.2. Data collection and analysis

The collection and modeling of accent data involves using multiple means – from published studies, to the use of linguistically trained informants, to the collection and analysis of new speech data – to construct complete segmental models for spoken variants of Portuguese.

For each accent treated, a complete segmental model consists of: a list of morphophonological contexts (especially those which are most vulnerable to change) and their corresponding phonetic realizations, i.e. a set of morphophonological post-lexical rules; conditions for the ordering of rules; and a list of lexical exceptions.

For the Luso-African and Luso-Asian varieties in particular, we have initiated a long-term effort aimed at recording Portuguese speakers from capital cities and smaller towns alike. Materials for the elicitation of read speech are based on those developed as part of [14]. The elicitation of spontaneous data is conducted in the form of an oral questionnaire for obtaining general speaker information and attitudinal data[4].

For corpus-based accent models, the assessment of segmental data is performed by trained phoneticians, who use Praat [15] to identify and label target segments.

4.3. How it works

General users will soon be able to access LUPo via the Portal (http://www.portaldalinguaportuguesa.org) to select from a list of available topolects and generate accent-specific pronunciations[5]. With LUPo's online interface, users can select from one of the 11 accents we have modeled so far and search for a given word, as illustrated in figure 2.

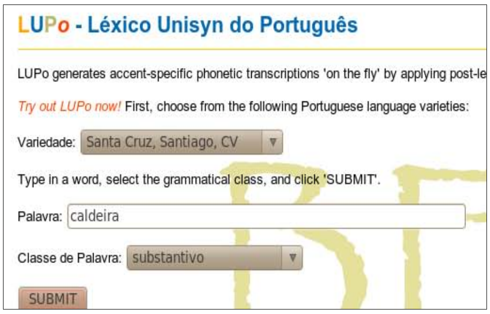


Figure 2: LUPo online prototype.

[4]General speaker information and attitudinal data will be made available through the searchable, online database.

[5]This capability is currently restricted to lemmas. Ultimately, LUPo will be extended to handle word forms and multi-word texts.

In figure 3, the result is displayed for the municipality of Santa Cruz, located in the eastern part of Santiago (Cape Verde) for the noun 'caldeira' (kettle). Here, we see the monothongization of /ej/ as /e/, along with dentalization of the lateral approximant /l/ in coda position.

Figure 3: Pronunciation of 'caldeira' in Santa Cruz.

A quick comparison with our speaker from Praia, Cape Verde's capital city, shows that this speaker maintains the diphthong /ej/, while producing a velarized lateral approximant, such as is common in EP (see figure 4).

Figure 4: Pronunciation of 'caldeira' in Praia.

In figures 5 and 6, respective results are shown for the same word as spoken in the cities of Lisbon (PT) and São Paulo (BR). Here, we see that for speakers of the standard Lisbon variety, the tonic vowel /ej/ is realized as the raised diphthong [ɐj]. Alternatively, for the same phonemic context, there is free variation in the production of both [ej] and [e] among São Paulo speakers, similar to that which was observed for the Santa Cruz (CV) informant.

Figure 5: Pronunciation of 'caldeira' in Lisbon.

Figure 6: Pronunciation of 'caldeira' in São Paulo.

Figure 5 also reveals a velarized lateral approximant [ɫ] typical of EP, as was observed for the Praia (CV) informant, while figure 6 presents the realization of this sound as the labio-velar approximant [w], common throughout Brazil.

The specific rules applied to generate LUPo's accent-specific transcription output are printed in the lower half of the results page (see figure 7). These are not phonological rules in the strict sense, but rather the transformations the master lexicon entry had to undergo to become the sort of output displayed in figures 3, 4, 5, and 6. At the bottom of the page, we include a description of the rules in plain language to make them easier to understand.

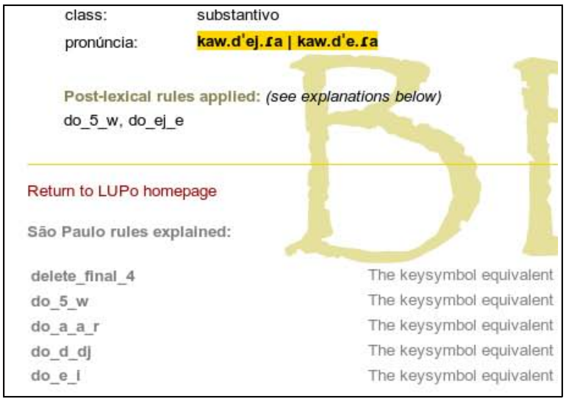


Figure 7: LUPo results page and rule descriptions.

Ultimately, LUPo's online interface (and searchable database) will allow users to observe more than one variant at a time. The map perspective in figure 8 offers one such glimpse at how LUPo is poised to provide linguists with a huge list of varying points and bundled phenomena – along with tangible data links – for testing notions of linguistic similarity and distance, and evaluating the pulling effect of different linguistic centers.

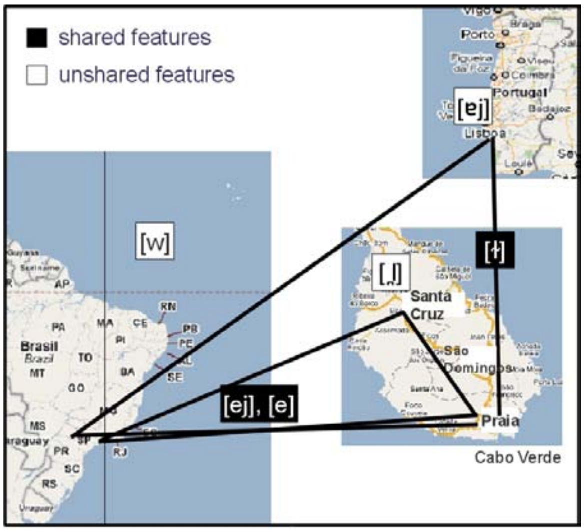


Figure 8: Tangible data links.

5. Conclusions

We have briefly introduced our work on the development of an accent-independent lexicon and rule system for generating phonetic transcriptions for regional accents of Portuguese. We have presented an initial prototype of the online LUPo system, along with a window into the phonetic segmental modeling of Luso-African idiolectal varieties from Cape Verde and Mozambique.

It has been shown that LUPo is designed to handle variability at the national and sub-national levels. This is achieved economically, through the sharing of rules across pluridimensional varieties (as demonstrated in the description of LUPo's regional accent hierarchy), while acknowledging those salient segmental features that are essential in distinguishing one variety from another, and resulting in more "natural" sounding synthetic speech.

In this vein, we seek to contribute to the improvement of Portuguese language speech technologies by providing high-quality pronunciation lexica, derived from linguistic rules, and covering as many topolectal variants as possible. We further anticipate that our work will have a positive impact on non-standard, "digitally endangered" [12] varieties of Portuguese, their enhanced prestige, and as varieties worthy of study in their own right.

6. Acknowledgements

The authors gratefully acknowledge the support of the Fundação para a Ciência e a Tecnología (PTDC/CLE-LIN/100335/2008), and the cooperation of Dr. Susan Fitt, whose development of the original English Unisyn Lexicon is the inspiration for this work.

7. References

[1] Ashby, S., Ferreira, J. P., and Barbosa, S., "Adapting the Unisyn Lexicon to Portuguese: Preliminary Issues in the Development of LUPo", in Proc. Iberian SLTech, 2009.

[2] Fitt, S., "Documentation and User Guide to UNISYN Lexicon and Post-Lexical Rules", technical report, Centre for Speech Technology Research, University of Edinburgh, 2000.

[3] Baxter, A. N., "Portuguese as a Pluricentric Language", in Clyne, M. G. (ed.) Pluricentric Languages: Differing Norms in Different Languages, Mouton de Gruyter, Berlin, 11-44, 1992.

[4] Évora-Sagna, M., Gray, V., Minges, M., "A Internet num PMA Lusófono: Estudo de Casos de Cabo Verde", International Case Studies report, International Telecommunications Union, 2002.

[5] Neves, A. C., "Young Learners and Phonological Variation in the Portuguese Mesolect of Cape Verde", in Linguistica ONLINE 11, 2009.

[6] Albarrán Alves de Carvalho, M. J., "Aspectos Sintáctico-Semânticos dos Verbos Locativos no Português Oral de Maputo", Dissertação de Mestrado, Universidade de Lisboa, 1987.

[7] Rouas, J.-L., Trancosco, I., Viana, C., and Abreu, M., "Language and Variety Verification on Broadcast News for Portuguese", in Speech Communication,50(11-12):965-979, 2008.

[8] Kirchhoff, K. and Vergyri, D. "Cross-Dialectal Data Sharing for Acoustic Modeling in Arabic Speech Recognition", in Speech Communication 46(1):37-51, 2005.

[9] Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F., and Strom, V., "Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-Based Speech Synthesis", in Speech Communication 52(2):164-179, 2010.

[10] Mac Lochlainn, M., "Sintéiseoir 1.0: a Multidialectal TTS Application for Irish", in ReCALL 22(2):152-171, 2010.

[11] Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., Jurafsky, D., Starr, R., and Yoon, S.-Y., "Accent Detection and Speech Recognition for Shanghai-Accented Mandarin", in Proc. Interspeech, 2005.

[12] Rusko, M., Darjaa, S., Trnka, M., Zeman, V., and Glovna, J., "Making Speech Technologies Available in (Serviko) Romani Language", in Sojka, P. et al. (eds.) Lecture Notes in Artificial Intelligence 5246:501-508, Springer-Verlag Berlin Heidelberg, 2008.

[13] Huang, R. and Hansen, J. H., "Gaussian Mixture Selection and Data Selection for Unsupervised Spanish Dialect Classification", in Proc. Interspeech, 2006.

[14] Rodrigues, M. C. M., "Lisboa e Braga: Fonologia e Variação", Fundação Calouste Gulbenkian, Lisboa, 2003.

[15] Boersma, P. and Weenink, D., "Praat: Doing Phonetics by Computer", computer program, vs. 5.1.43, retrieved from http://www.praat.org/ 4 August 2010.

A System for User Location by Means of RFID Devices for a Dialogue System Designed to Interact in an Ambient Intelligence Environment

Gonzalo Espejo[1], *Nieves Ábalos*[1], *Ramón López-Cózar*[1], *Zoraida Callejas*[1], *David Griol*[2]

[1] Dep. of Languages and Computer Systems. CITIC-UGR. University of Granada
[2] Computer Science Department. Carlos III University of Madrid
{gonzaep,nayade}@correo.ugr.es, {rlopezc,zoraida}@ugr.es, dgriol@inf.uc3m.es

Abstract

This paper presents our current work in locating users through RFID devices for a multimodal dialogue system called *Mayordomo*. The paper explains how RFID works and describes the system architecture to connect hardware devices (RFID) and software applications (*Mayordomo)*. Finally several examples are shown to explain how this implementation improves the dialogue between the system and users.

Index Terms: speech recognition, multimodal dialogue system, ambient intelligence, RFID.

1. Introduction

Ambient Intelligence (AmI) is a research area that has attracted a lot of efforts by the scientific community in the last years [1, 2, 3, 4, 5, 6]. The aim of AmI is to create environments in which users are able to interact in a natural and transparent way with systems that help them carrying out their daily leisure and work activities.

AmI has been developed for some different environment, such as medical fields [7], home [8], public spaces [9] and learning spaces, as museums [10].

One of the most important characteristics in Ambient Intelligence systems is proactivity. This means that the system should predict the behaviour and preferences of the users to anticipate them.

In this paper we propose the automatic location of users for a multimodal dialogue system in a home environment through RFID devices. This way, the system is able to adapt its behaviour in several situations once it knows where the user is. For example, in spoken interfaces like ours, the system can vary the dialogue with the user so that he does not have to explicitly provide information about his location.

RFIDs are considered one of the main ways for physical browsing, which is an interaction paradigm that associates digital information with physical objects [1]. RFID devices have been used in several research projects concerned with user localization. For example, [11] used RFID devices to locate students in an educational environment, whereas [2] employed these devices in a home environment. Also, in [12] RFID technology is used in a medical environment.

The paper is organized as follows. In section 2, we describe some features of our *Mayordomo* multimodal dialogue system, focusing on the speech-based interaction module of automatic speech recognition. Section 3 explains some features about RFID devices and how they work. Next, we describe the overall system architecture in order to explain how hardware and software are connected and work together. Finally, conclusions and future work are presented.

2. *Mayordomo* Multimodal Dialogue System

2.1. General Overview

Mayordomo is a multimodal dialogue system developed in our laboratory, to be integrated in an Ambient Intelligence environment. The goal of *Mayordomo* is to centralize the control of the appliances located in a home. The interaction with the appliances can be carried out through spontaneous speech or with the GUI interface shown in Figure 1.

In addition to handling appliances, the system supports different types of users, depending on the administrator level and the experience with *Mayordomo*. The system administrator has privileges to perform special actions, for example, installing and uninstalling appliances Also, restrictions are allowed to some users. For instance, parents can forbid that their children watch TV after 10 p.m. Moreover, the system creates a log of all actions carried out within the environment by any user which is only accessible to the administrator user.

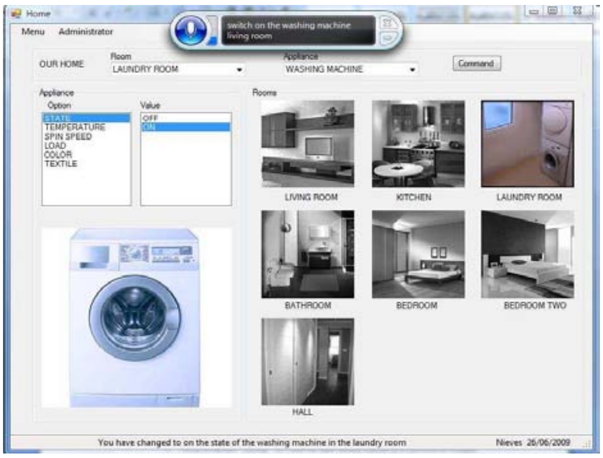


Figure 1: *GUI interface of* Mayordomo.

Scalability is one of the main characteristics of *Mayordomo*. Different appliances can be installed dynamically, and a reboot of the system is not necessary. Scalability is possible due to configuration and grammar files.

2.2. Speech-based Interaction

We use Windows Vista Speech Recognition (WVSR) to implement the speech-based interaction. This package includes both the engine for automatic speech recognition (ASR) and the engine for text-to-speech synthesis (TTS).

Figure 2, shows the detailed architecture of the speech-based interface, focusing on the dialogue manager we have implemented.

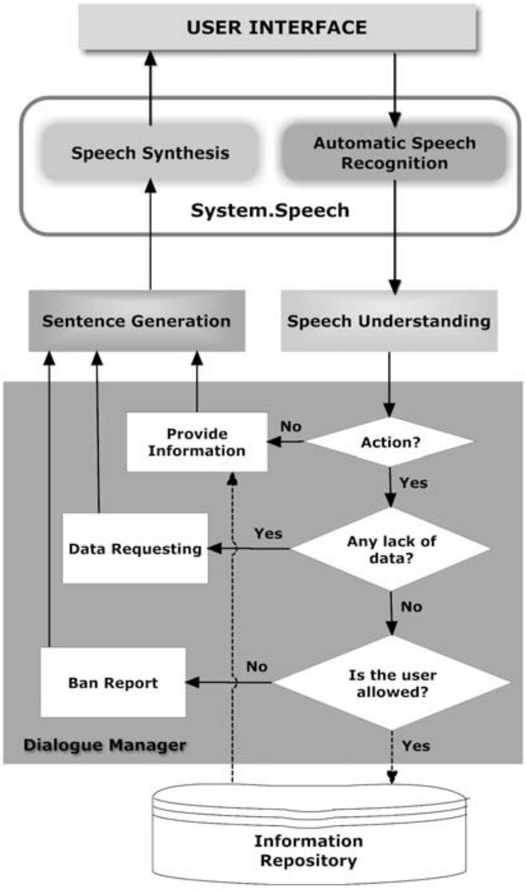


Figure 2. *Speech-based interaction in Mayordomo*.

As mentioned above, each appliance has an associated configuration file that allows the user to control it orally. This file contains a specific grammar for interacting with the appliance that is used for ASR. This grammar is specified in SRGS (Speech Recognition Grammar Specification) format[1].

When specifying grammars for the different home appliances of a home, we can consider three strategies: (1) allowing keyword recognition using the specific subrule *keywords*, (2) allowing keyword recognition without using this subrule, (3) not allowing keyword recognition.

Using the first strategy (which is the one used by the system) and the second one, grammars allow the recognition of keywords. The main difference between these two strategies lies in the way that the recognition is carried out. Using the first one (see Figure 3), the initial rule of the grammar contains four subrules: *order*, *sentence*, *request* and *keywords*. The subrule for *keywords* includes all the words related to the control a particular appliance. For example, keywords related to the T.V. may be concerned with the place where this appliance is (e.g. *living room* or *kitchen*), the attribute or characteristic the user wants to change (e.g. *volume* or *channel*) and the action to be performed with the appliance (e.g. *switch on* or *turn off*).

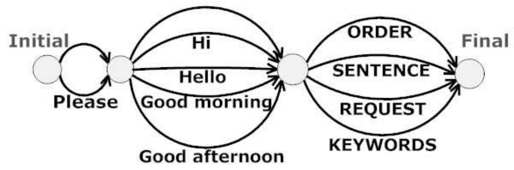


Figure 3. *Initial rule of ASR grammar including a subrule for recognition of keywords*.

This strategy is the most suitable for users who do not provide all the data required to perform an action with an appliance. If any data is missing to perform the action, the dialogue manager of the system prompts the user for the missing information. In this case, the user can provide just the missing data, that is, he does not need to utter the entire order. The advantage of this strategy is that the interaction is more comfortable for the user, particularly if the orders are complex and long. For example, if when processing the order:

"*Set the temperature of the washing machine of the laundry room to thirty degrees*"

the system does not understand the room where the appliance is located, *Mayordomo* may prompt "*Where?*" and the user may answer "*In the laundry room"*.

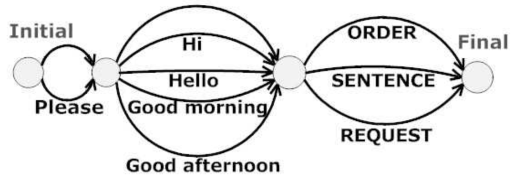


Figure 4. *Initial rule of ASR grammar without subrule for keywords.*

Using the second strategy for ASR (see Figure 4), grammars also allow the recognition of keywords. However, in this case, they do not use the specific subrule for *keywords* and all the elements of the subrules *order*, *sentence* and *request* are optional. An optional element in a rule means that it can be provided or not by the user, thus, providing it is not necessary to trigger the rule. For example, the subrule *order* has three elements: verb prep (subrule with a list of elements such as *"switch on"*), object (i.e. *"the light"*) and where (subrule with the rooms, for instance, *"of the kitchen"*). In this rule, if the element *where* is optional, the phrase *"Switch on the light"* will trigger the rule and will be recognized by the ASR. Therefore, this strategy permits all kinds of combination of words, resulting in a greater number of ASR errors as it allows incorrect combinations of words, for example, *"Hello washing off hall"*.

Using the third strategy for ASR, the initial rule of the grammar is the same as in the second strategy. However, it does not allow optional elements in the subrules. Thus, this strategy is not advisable when the sentences are complex and long.

Grammars are loaded into the system's memory at the beginning of interaction, and remain there throughout the dialogue so that the recognition engine can use them at any time during the interaction.

[1] http://www.w3.org/TR/speech-grammar/

3. RFID Devices

User location is known by the multimodal dialogue system through the information obtained from RFID devices, installed throughout the environment.

The interaction takes place using mobile terminals, such as mobile phones or cards that have a tag associated with a reader device. Each tag contains a universal resource identifier (URI, or Web address). The readers are distributed all over the environment. Each RFID reader has a serial or identification number. When one of the tagged objects is scanned by the reader, the tag and the identification number of the reader are sent to the dialogue system. The distance between the physical object and the RFID reader can range from a centimetre to four meters. Usually, in indoor and small environments this distance is not very large.

To be useful for our purposes, the dialogue system must know which user carries each tag. Taking into account this information, the system can find out which was the last RFID reader used by the user, and thus deduce his location.

Figure 5 shows how the different elements interact with each other. The smaller arrows represent the connection established between the RFID device and the server which manages the information. The big arrows show the way that the information flows from the user to the server. Firstly the user's card is scanned by a RFID device, and then the server receives the information related to the user from the device.

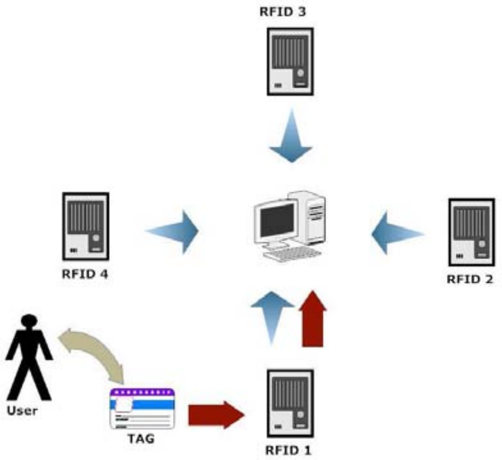


Figure 5. *Elements used in location though RFID*

These RFIDs devices used in *Mayordomo* are Phidgets[2]. We have selected them because the easy way to use and the possibility of extend the architecture to other devices in a simple way.

4. System Architecture

In this section we present the global system architecture. As it shown in Figure 6 the system in ready to use several hardware devices (RFIDs), and these devices are managed by some software application.

A middleware layer has been developed in order to connect hardware devices and software applications. This middleware layer consists in a repository in which information about devices, rooms and users is stored. Moreover, this information can be modified by hardware devices that are triggered when an event happens, for instance, when a device RFID detects a user's tag. In the same way, software applications are allowed to access the information stored with the aim of obtaining information about users or events occurring in the environment.

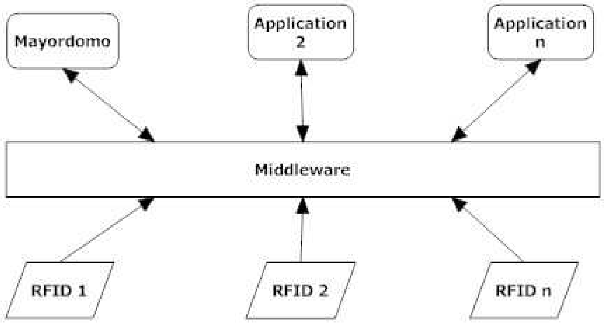


Figure 6. *System Architecture*

The information needed and stored in the middleware layer implemented in *Mayordomo* are a list of pairs users-tag, a list of pairs RFID devices-rooms and a list of RFID devices-tag triggered.

Using this information, the dialogue manager is able to detect who is the user and in which room is interacting with the system.

5. Automatic Location

The main goal of using RFID devices with a dialogue system is to reduce the information that the user must provide in order to provide shorter dialogues in an easy and natural way.

Knowing the room where the user is the dialogue becomes simpler. Comparing dialogue 1 and dialogue 2 with dialogue 3 it is clear how in the second one it is not necessary to explicitly provide to the system the room in which the user is located.

Starting with dialogue 1, the user asks the system to switch on the light, in a natural way, but the system does not know his location yet, and thus some additional turns are necessary for the user to specify his exact location.

User: Please, switch on the light.
System: Where?
User: In the living room.
System: You have changed to on the state of the light in the living room.
(The user goes to the hall)
User: Please, switch on the light.
System: Where?
User: In the hall.
System: You have changed to on the state of the light in the hall.

Dialogue 1.

In dialogue 2 the user includes in his order the room in which he wants to switch on the light, which makes the sentence longer and less natural than in dialogue 1.

[2] http://www.phidgets.com

User: Please, switch on the light in the living room.
System: You have changed to on the state of the light in the living room.
(The user goes to the hall)
User: Please, switch on the light in the hall.
System: You have changed to on the state of the light in the hall.

Dialogue 2.

Finally, in dialogue 3 the system includes RFIDs and these devices have been used by the user in order to identify where he is. This way of interaction is the best for the users: the system requires less information during the dialogue making it more friendly and natural.

(The user goes to the living room and uses the tag with the RFID device)
User: Please, switch on the light.
System: You have changed to on the state of the light in the living room.
(Now, the user goes to the hall and uses the tag with the RFID device)
User: Please, switch on the light.
System: You have changed to on the state of the light in the hall.

Dialogue 3.

6. Conclusions and future work

In this paper we have presented the current status of the design and implementation of location through RFID in our multimodal dialogue system.

We have described how RFID works, the global architecture and the main advantages using RFID devices. The architecture described has been designed to be used, not only with RFID, but with any other type of devices such as touch sensors or temperature sensors.

As it has been shown, using RFID devices less information is required in dialogues between users and our system. This feature makes the dialogues more natural as the user does not have to provide information which is implicit in the context of the interaction.

As future work we plan to include more devices in *Mayordomo*. In addition, we plan to include users' profiles describing their preferences and usual activities in order to make the system proactive. In this way it would be possible, for instance, to detect which is the favourite channel of any user in our environment.

7. Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology, under project TIN2007-64718 Adaptive Hypermedia for Attention to Different User Types in Ambient Intelligence Environments.

8. References

[1] Aghajan, H., López-Cózar, R, Augusto, J. C. "Human-centric Interfaces for Ambient Intelligence". Academic Press. 2010

[2] Haya, P. A., Montoro, G., Alamán, X. "A Prototype of a Context-Based Architecture for Intelligent Home Environments". CoopIS/DOA/ODBASE (1): 477-491. 2004

[3] Augusto, J. C. y McCullagh, P. "Ambient Intelligence: Concepts and Applications" Int'l J. Computer Science and Information Systems, vol. 4, no. 1, pp. 1–28.5

[4] Augusto, J. C. "Ambient Intelligence: Basic Concepts and Applications." Series: Communications in Computer and Information Science, Vol. 10. Springer Verlag, pp 14-24. 2008

[5] Ramos, C., Augusto, J. C., Shapiro, D. "Ambient Intelligence – The next step for Artificial Intelligence". IEEE Intelligent Systems
23(2) (March/April 2008) pp 15 – 18. 2008

[6] Remagnino, P. y Foresti, G. L. "Ambient Intelligence: A New Multidisciplinary Paradigm". IEEE Transactions on Systems, Man and Cybernetics, Vol. 35(1), pp. 1-6, Jan. 2005.

[7] Corchado, J. M.; Bajo, J.; de Paz, Y.; Tapia, D. I.: Intelligent environment for monitoring Alzheimer patients, agent technology for health care. Decision Support Systems, 44, pp. 382–396, 2008

[8] Pérez-Castrejón, E.; Andrés-Gutiérrez, J. J.: AAL and the Mainstream of Digital Home. Lecture Notes in Computer Science 5517, pp. 1070–1082, 2009

[9] Corchado, J. M., Pavón, J., Corchado, E., Castillo, L. F.: Development of CBR-BDI Agents: A Tourist Guide Application. Lecture Notes in Artificial Intelligence 3155, pp. 547-559, 2004

[10] Wakkary, R.; D. Evernden, Museum As Ecology: A Case Study Analysis Of An Ambient Intelligent Museum Guide, in J. Trant and D. Bearman (eds.). Museums and the Web 2005: Proceedings, Toronto: Archives & Museum Informatics, published March 31, 2005

[11] López-Cózar, R., Callejas, Z., Montoro, G., Haya, P. " DS-UCAT: Sistema de Diálogo Multimodal y Multilingüe Para un Entorno Educativo". Proc. IV Jornadas en Tecnología del Habla, pp. 135-140. . 2006

[12] Collins G. R. "Usable Mobile Ambient Intelligent Solutions for Hospitality Customers" Journal of Information Technology Impact, Vol 10, No 1. pp. 45-54. 2010

CHARLATAN: A Task-Independent Dialog Platform

Francisco Torres Goterris

Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València / Universidad Politécnica de Valencia, Spain
ftgoterr@dsic.upv.es

Abstract

In this paper, we describe a task-independent dialog platform which is called CHARLATAN. It is a Java application that can work in stand-alone mode and as an applet added in a website. This dialog platform integrates dialog management, natural language generation, user simulation, and a parameterization of the task-dependent issues. Currently, it allows us to develop dialogs in two domains: a train services information system, and a sport booking system. However, due to the task-independence of CHARLATAN, new domains can be easily incorporated. Dialogs are automatically generated by means of a user simulation technique, or interacting with real users through the graphical interface.

Index Terms: stochastic dialog management, user simulation, task independence, synthetic acquisition

1. Introduction

Advances in developing of spoken dialog systems are usually hard-worked because of different problems: low confidence of the information interchanged among the dialog modules, high cost of the corpora acquisition, subjectivity in real users evaluation, or high task-dependence in the design of the modules. In the development of the CHARLATAN platform, we have focused on solving some of these problems: the acquisition of the corpora (by the user simulation technique) and the independence between the modules and the domains (by isolating the task parameters in configuration files).

The adaptation to new semantic domains is one of the aims in the EDECAN project [1]. In this research frame, the CHARLATAN dialog system has been developed to attend different semantic-restricted tasks: the BASURDE and DIHANA tasks [2], which access a train information system; and the EDECAN-SPORT task, which provides access to a sport courts booking system. In CHARLATAN, the information that is related to the task has been encapsulated into the models, the scenarios, and other configuration files. Thus, the data-structures are initialized reading these files, and the methods have been appropriately parameterized.

In this paper, we present the CHARLATAN platform that is working appropriately in both tasks, and very soon it will be attending other different tasks, because of its parameterization of the task-dependent issues.

2. The dialog platform

The CHARLATAN dialog platform integrates the user and system dialog managers, the user and system natural language generators, and the database manager. It can also integrate understanding modules. Figure 1 shows its block diagram.

The dialog manager [3] is based on a stochastic dialog model, which is a bigram model (BM) of dialog acts, and includes a historic register (HR), which stores all the data provided in previous turns. This module (SDM in Figure 1) can attend different tasks, just reading their corresponding domain parameters (DP) from configuration files.

The user simulator [4] selects states of the same BM, and applies a set of target planning rules (TPR) that implement a collaborative strategy. These rules are task-independent, and they serve to acquire consistent dialogs in any task. This module (UDM in Figure 1) allows us to acquire synthetic dialogs, learn dialog models, and evaluate the dialog system.

Both managers, SDM and UDM, receive and generate frames, which are semantic representations of the natural language sentences. The natural language generators (ULG and SLG in Figure 1) translate the frames into sentences in natural language (currently, in Spanish and English). Both language generators work using a set of templates and rules for instantiating the templates.

According to this design, we have developed a JAVA dialog platform [5]. Using it, we can acquire dialogs for any task that would be appropriately defined by means of a set of CHARLATAN configuration files. In the interactive mode, real users can provide the frames that correspond to their dialog intentions through a graphical interface, and they can read the system answers, carrying out whole dialogs. In the simulation mode, dialogs are completely done by CHARLATAN, allowing us to simulate dialogs turn by turn, or whole dialogs, or series of any number of dialogs, and to specify which scenarios are simulated. In addition, the user frames can be modified by including errors (deletions, insertions, substitutions) in the attributes whose values are critical to the success of the dialog.

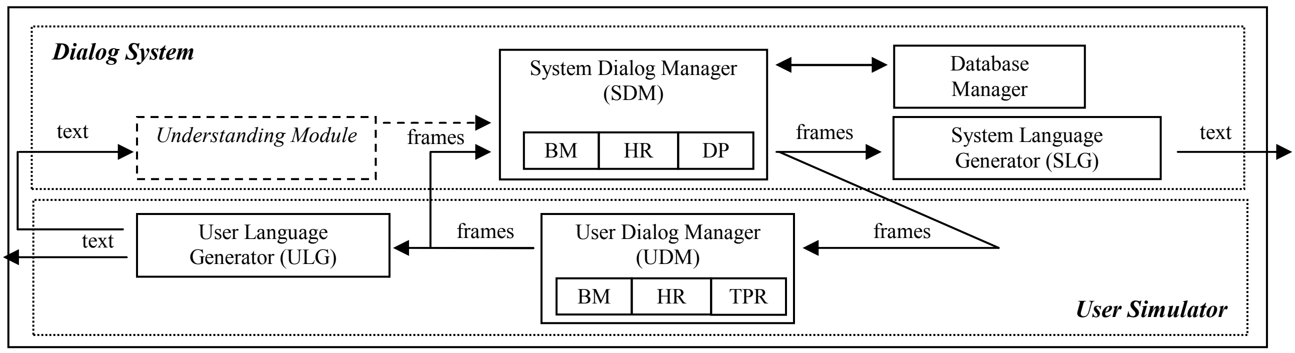


Figure 1. *Dialog system block diagram*

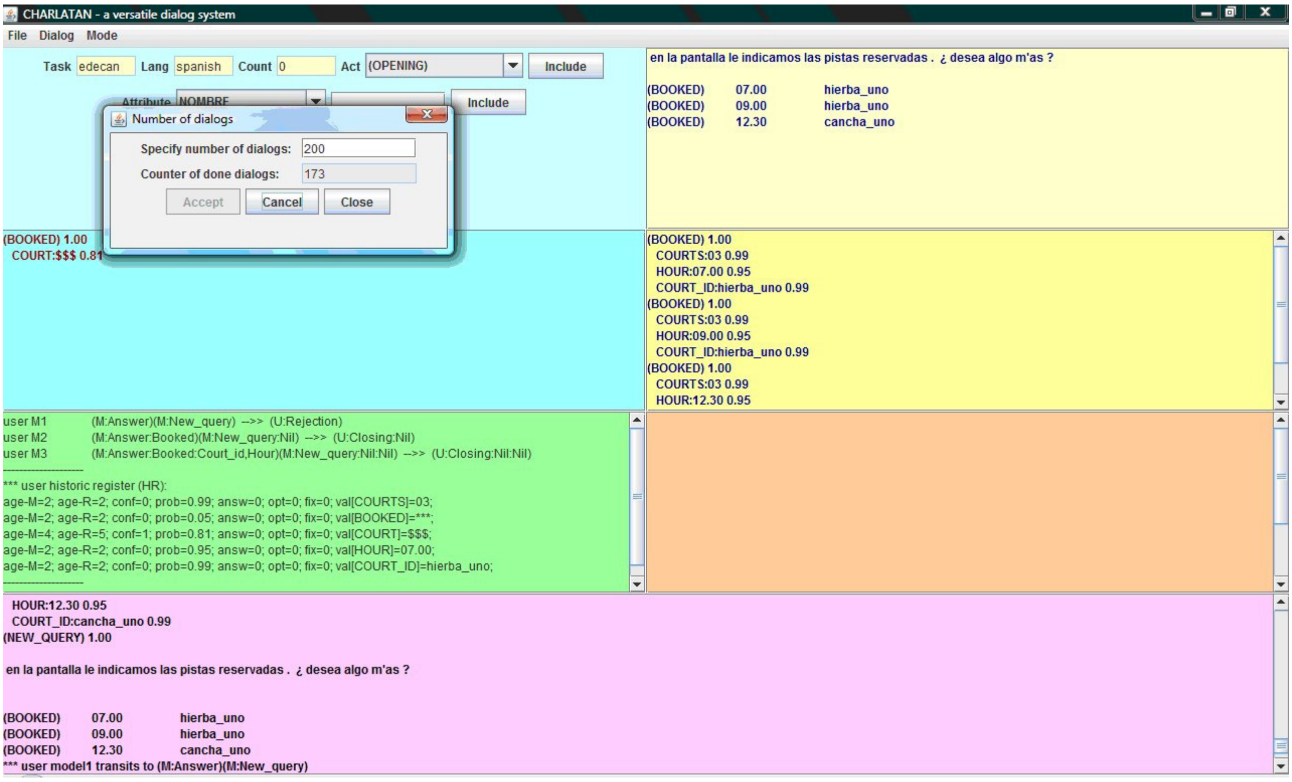


Figure 2. *The* C*HARLATAN platform acquiring a set of simulated dialogs*

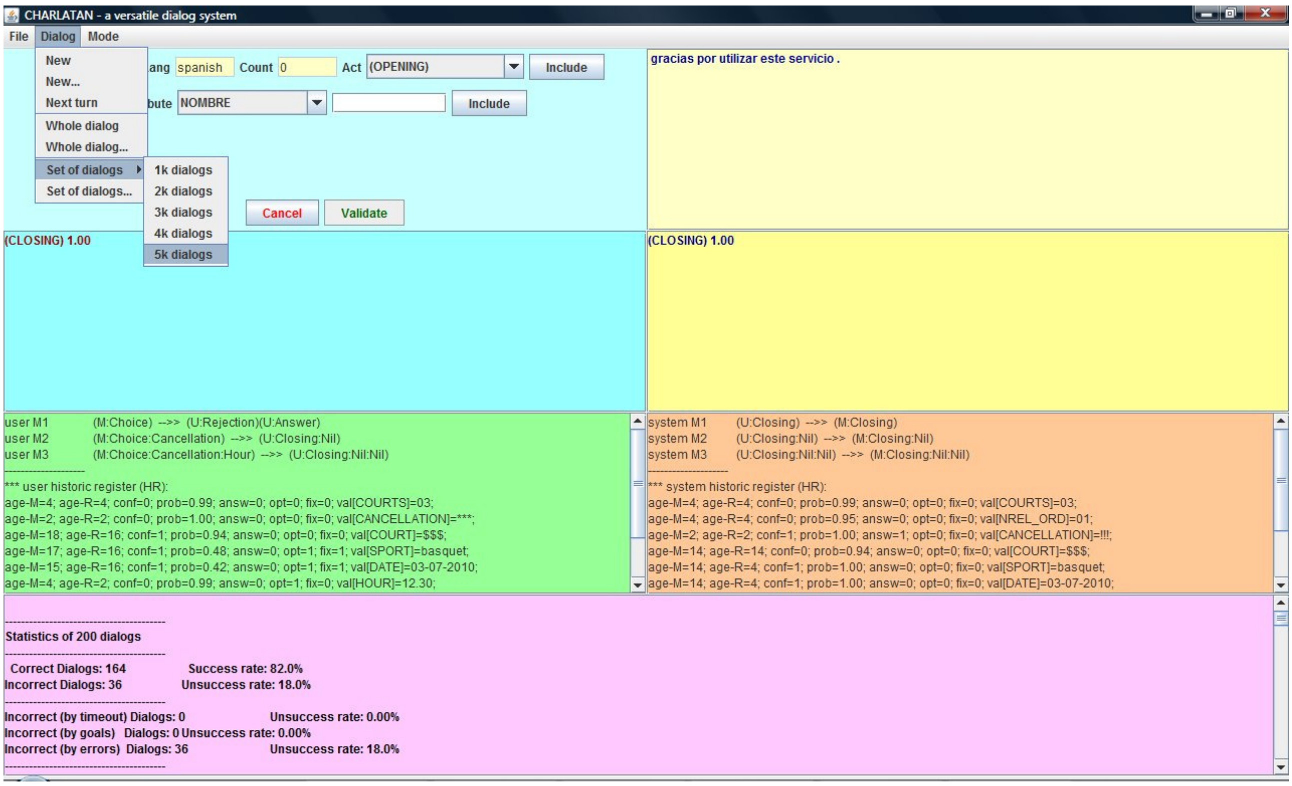


Figure 3. *The* C*HARLATAN platform at the end of a simulation*

3. Interface description

In this section, the interface of the CHARLATAN dialog platform is described by means of several screenshots (Figures 2, 3, and 4) that show the operation of the platform while acquiring dialogs in the EDECAN-SPORT task.

The CHARLATAN applet interface consists of seven areas of text. The left side of the applet corresponds to the user, and the right side corresponds to the system. The three areas on the left, from top to bottom, are the real user graphical interface, the output of the UDM (user frames), and its internal state (BM transitions, and HR content). The three areas on the right, from top to bottom, are the output of the SLG (system

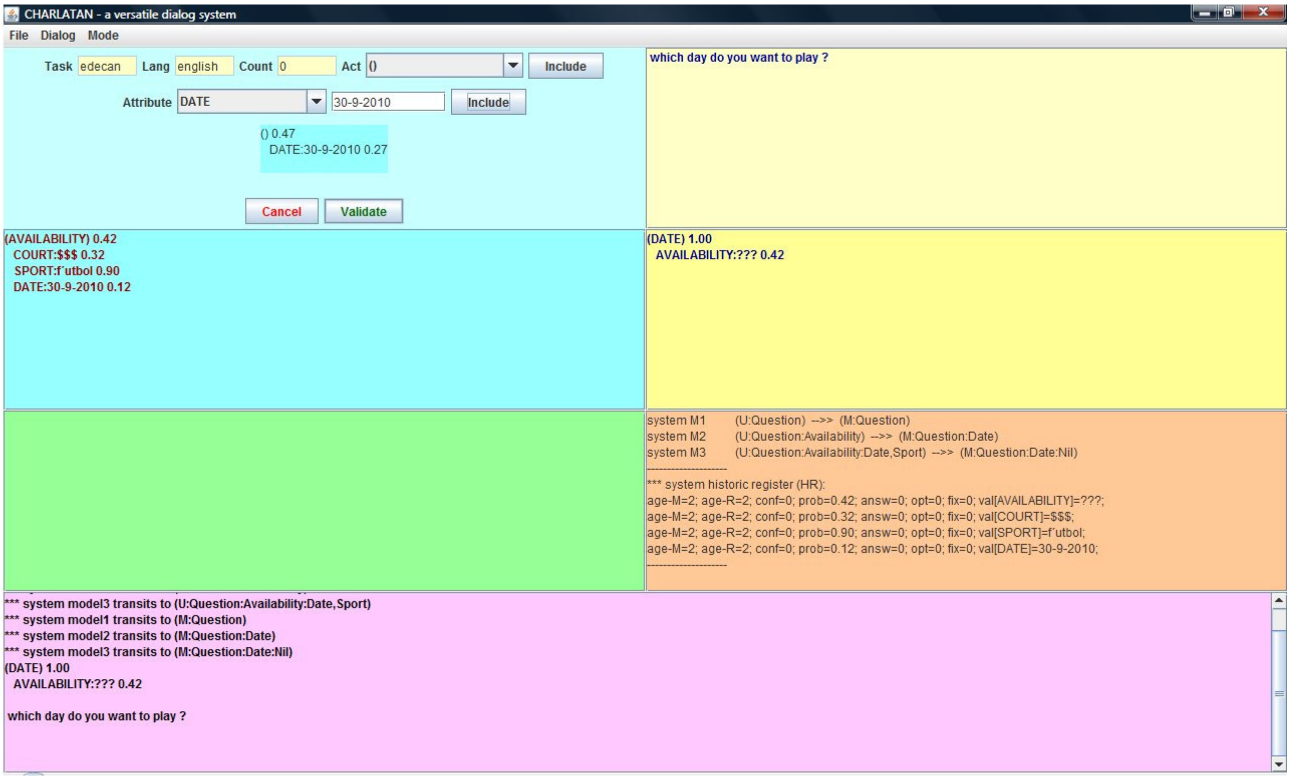


Figure 4. *The* C*HARLATAN platform interacting with a real user*

sentences), the output of the SDM (system frames), and its internal state (BM transitions, and HR content). In the bottom text area, the whole dialog is collected, or statistics of a series of dialogs are collected, depending on the way of operation.

Figure 2 shows a screenshot in which CHARLATAN is acquiring a set of 200 simulated dialogs of the EDECAN task, using Spanish as natural language, with simulation of errors in input frames. The screenshot corresponds to the dialog number 173, at the moment in which the user simulator has asked for booked courts, and the system has answered providing a list with information of three booked courts. The bottom text area is collecting all the turns of this dialog.

Figure 3 shows a screenshot in which CHARLATAN has finished a previous simulation. In such a case, the bottom text area shows the collected statistics of the series. In the menu bar, the *Dialog menu* is expanded, showing the commands to initiate new dialogs, run dialog turns (turn by turn simulation), acquire whole dialogs (choosing the scenarios to be simulated), and acquire sets of dialogs.

Another menu, the *Mode menu*, is available in the menu bar. The *Mode menu* allows us to select the task, the natural language, the type of user (real or simulated), the simulation of input errors (activation or deactivation), and some advanced options like the selection of dialog manager strategy (full stochastic strategy, or hybrid strategy –half stochastic, half heuristic–), the selection of training or test modes in using the dialog models, and the selection of thresholds for adjusting the number of simulated input errors.

Figure 4 shows a screenshot in which CHARLATAN is interacting with a real user, carrying out a dialog of the EDECAN task, using English as natural language. In the previous turn, the user has asked for courts availability (see user frame in the second left text area). The system has replied with a question about the date (see system sentence, system frame, and BM transitions and HR content, in the right text areas), due to the low confidence of this attribute. At the moment of the screenshot, the user is answering, providing again the date value (see user graphical interface, in the first left text area). This graphical interface consists of combo boxes, text fields and buttons, and it allows the user to specify the dialog acts, the attributes and their values (i.e., to build the user frames). The confidence scores of these user frames are randomly generated.

4. An example of a dialog

In order to illustrate the dialogs that the CHARLATAN platform can acquire, we present an example in this section. The following dialog corresponds to a simulation of the scenario 1 of the EDECAN task, in which the user has to obtain the availability of courts, given some restrictions (sport, date, court-type), and then s/he has to choose one of these courts for booking it. Figure 5 shows this dialog, which log has been simplified, and we only present the most relevant information. For each turn, we have included the user frames, the system BM transitions, the system HR content, the system frames, and the system sentences.

The system opens the dialog (in S0 turn) asking for the user queries. The user simulator answers providing an availability frame with the values of the attributes that correspond to its goal. In S1 turn, the system asks for the sport because of its low confidence. It also makes an implicit confirmation of the court-type attribute. Once the sport has been confirmed, the system needs to confirm the date attribute, which is also not enough reliable. The system does it in S2 turn, and it also makes an implicit confirmation of the hour attribute. Given that the user simulator confirms all the mandatory items, at the following turn (S3 turn), the system accesses the database and provides the availability frames. In U3 turn, the user simulator chooses one of these courts for booking it. Such a booking is confirmed by the system in S4 turn. After this, the user simulator has achieved all the goals of the scenario. Thus, it generates a closing frame. And the system also closes the dialog.

S0: (OPENING) 1.00
S0: Welcome to the UPV sport courts booking system. How may I help you?

U0: (AVAILABILITY) 1.00
SPORT: tennis 0.50
DATE: 25-06-2010 0.47
COURT_TYPE: exterior 0.51

*** **System L3-BM** transitions:
(U:Question:Availability:Court_type,Date,Sport) →
(M:Question:Sport:Court_type)
*** **System HR** content:
[AVAILABILITY] confidence = 1.00 value = ???
[SPORT] confidence = 0.50 value = tennis
[DATE] confidence = 0.47 value = 25-06-2010
[COURT_TYPE] confidence = 0.51 value = exterior
S1: (SPORT) 1.00
AVAILABILITY: ??? 1.00
COURT_TYPE-ACTUAL: exterior 0.51
S1: On exterior, which sport do you want to play?

U1: () 1.00
SPORT: tennis 0.50
DATE: 25-06-2010 0.47
HOUR: 16.00-20.00 0.64

*** **System L3-BM** transitions:
(U:Answer:Date,Hour,Sport:Date,Hour,Sport) →
(M:Confirmation:Date:Date,Hour)
*** **System HR** content:
[AVAILABILITY] confidence = 1.00 value = ???
[SPORT] confidence = 1.00 value = tennis
[DATE] confidence = 0.47 value = 25-06-2010
[HOUR] confidence = 0.64 value =16.00-20.00
[COURT_TYPE] confidence = 0.51 value = exterior
[INT_HOUR] confidence = 0.64 value = 16.00-20.00
S2: (CONFIRMATION) 1.00
DATE: 25-06-2010 0.47
HOUR-ACTUAL: 16.00-20.00 0.64
S2: Between four o'clock p.m. and eight o'clock p.m., do you want to play on June the twenty-fifth?

U2: (ACCEPTANCE) 1.00

*** **System L3-BM** transitions:
(U:Acceptance:Date:Nil) →
(M:Answer:Availability:Court_id,Hour)(M:New_query:Nil:Nil)
*** **System HR** content:
[AVAILABILITY] confidence = 1.00 value = ???
[SPORT] confidence = 1.00 value = tennis
[DATE] confidence = 1.00 value = 25-06-2010
[HOUR] confidence = 0.64 value = 16.00-20.00
[COURT_TYPE] confidence = 0.51 value = exterior
[INT_HOUR] confidence = 0.64 value = 16.00-20.00
S3: (AVAILABILITY) 1.00
COURTS: 03 0.99
HOUR: 17.00 0.99
COURT_ID: olympic_one 0.95
(AVAILABILITY) 1.00
COURTS: 03 0.99
HOUR: 18.30 0.99
COURT_ID: grass_two 0.95
(AVAILABILITY) 1.00
COURTS: 03 0.99
HOUR: 18.30 0.99
COURT_ID: clay_one 0.95
(NEW_QUERY) 1.00
S3: You can see the available courts on the screen.
17.00 olympic_one
18.30 grass_two
18.30 clay_one
Do you want anything else?

U3: () 1.00
BOOKING: !!! 0.08
HOUR: 17.00 0.99
COURT_ID: olympic_one 0.95

*** **System L3-BM** transitions:
(U:Answer:Booking:Court_id,Hour) →
(M:Choice:Booking:Court_id)
*** **System HR** content:
[COURTS] confidence = 0.99 value = 03
[NREL_ORD] confidence = 0.95 value = 01
[AVAILABILITY] confidence = 1.00 value = !!!
[BOOKING] confidence = 0.08 value = !!!
[SPORT] confidence = 1.00 value = tennis
[DATE] confidence = 1.00 value = 25-06-2010
[HOUR] confidence = 0.99 value = 17.00
[COURT_TYPE] confidence = 0.51 value = exterior
[COURT_ID] confidence = 0.95 value = olympic_one
[INT_HOUR] confidence = 0.64 value = 16.00-20.00
S4: (CHOICE) 1.00
BOOKING: !!! 0.08
COURT_ID: olympic_one 0.95
S4: I confirm to you that the olympic_one court has been booked.

U4: (CLOSING) 1.00

*** **System L3-BM** transitions:
(U:Closing:Nil:Nil) → (M:Closing:Nil:Nil)
S5: (CLOSING) 1.00
S5: Thank you for using this service.

Figure 5. *Example of an EDECAN dialog*

5. Conclusions

This dialog platform, called CHARLATAN, is currently a robust, versatile and easy-to-use tool to generate consistent dialogs in different semantic-restricted tasks. At the beginning, we started adapting a BASURDE-task-dependent dialog manager to appropriately attend both tasks (BASURDE and EDECAN). Then, we followed working in the generalization and complete parameterization of the dialog managers. Thus, we have achieved, with our CHARLATAN prototype, a task-independent dialog platform.

6. Acknowledgements

This work has been partially funded by CICYT under project TIN2008-06856-C05-02/TIN, Spain.

7. References

[1] Lleida, E., et al., "EDECÁN: sistema de diálogo multidominio con adaptación a contexto acústico y de aplicación", Jornadas en Tecnología del Habla (JTH), Zaragoza, Spain, 291–296, 2006.
[2] Benedí, J.M., et al., "Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA", Proc. of 5th LREC, Genove, Italy, 1636–1639, 2006.
[3] Torres, F., et al., "Error handling in a stochastic dialog system through confidence measures", Speech Communication 45, 211–229, 2005.
[4] Torres, F., Sanchis, E., Segarra, E., "User simulation in a stochastic dialog system", Computer Speech & Language, 22, 230–255, 2008.
[5] Torres, F., "Prototype of the dialog system", available in website http://www.laesteladetanit.es/prototype.htm, 2010.

Information search engine for multilingual audiovisual contents: BUCEADOR

Asunción Moreno

TALP Research Center, Technical University of Catalonia, Spain
asuncion.moreno@upc.edu

Abstract

BUCEADOR is a three years project focused on advanced research in the core Spoken Language Technologies (SLT) such as diarization, speech recognition, speech machine translation, and text-to-speech conversion, and the successful joint integration of all of them in a multilingual and multimodal information retrieval system.

In order to show the achievements of the project in the above mentioned technologies and their successful joint integration, a show case, consisting of a search engine for multilingual audiovisual contents, will be created. Specifically, the audiovisual content includes broadcast news from several TV, radio, and internet channels in all the official Spanish languages (Spanish, Catalan, Basque and Galician) plus English.

Index Terms: Diarization, speech recognition, speech synthesis, speech to speech translation.

1. Introduction

BUCEADOR is a three years project funded by the Spanish Government. The project is driven by three public Spanish Universities with the collaboration of Barcelona Media – Centre d'Innovaciò (BMCI) research center. The project focuses on advanced research in all core Spoken Language Technologies (SLT), (speech recognition, speech machine translation, and text-to-speech conversion), and the successful joint integration of all of them in a multilingual and multimodal information retrieval system.

The goal of the project is to achieve improvements in all the SLT components and voice search applications to improve human-machine and human-to-human communication among all the official languages spoken in Spain and between them and English. The aim of the project is to obtain research advances in each SLT technology. New techniques will be explored for the diarization of speeches; in speech recognition, unconstrained conversational speech recognition systems will be implemented in several languages; in machine translation, new machine learning algorithms and linguistic knowledge will be incorporated; in speech synthesis, new acoustic and prosodic models will be implemented..

In order to show the achievements in the above mentioned technologies and their successful joint integration capability, the project will create a show case consisting of a search engine for multilingual audiovisual contents. Specifically, broadcast news from several TV, radio, and internet channels in all the official Spanish languages (Spanish, Catalan, Basque and Galician) plus English will be utilized. The system query and the audiovisual files, oral and textual, can be in any of these languages. By means of diarization, speech recognition and machine translation techniques, the search engine and the information retrieval system can explore the contents of all the files independently of the language in which they were created. Finally, the user of the system can choose the output format, either text or speech, in the language he/she chooses.

This paper is structured as follows: Section 2 describes the current state of the art in the mentioned technologies, Section 3 describes the general objectives of the project and the specific objectives to be achieved in each technology; Section 4 briefly describes the groups involved in the project. The paper ends with acknowledgments in section 5 and the references list in section 6.

2. Involved Technologies

The speech technologies involved in the BUCEADOR project show different research achievements and challenges

2.1. Speech Recognition

The objective of achieving truly human-level automatic transcription of speech, has received a vast amount of research efforts during past decades. As a result, the automatic speech recognition (ASR) technology has converged to an almost universal standard based in HMMs (Hidden Markov Models) and statistical N-Gram language models (LM). With this baseline technology, systems today are capable of achieving quite impressive performance levels in complex tasks, but with extreme sensitivity to mismatches between training and real conditions. Current challenge in ASR technology is to achieve universal systems, robust enough to resist changing acoustic conditions, and adaptable to different speakers and language styles. Most of the research conducted on projects like TC-STAR [1], CHIL [2], AMIDA [3] and EMIME [4] or in public evaluations like NIST [5] is directed to this objective.

Modern recognition systems are very complex programs, comprised of several interconnected modules which perform a variety of tasks: audio diarization, baseline decoding, acoustic model adaptation, language model adaptation, N-Best rescoring, etc. The task of audio diarization is to segment the audio stream into voice and non-voice regions, and to divide an audio stream into speaker homogeneous segments, determining which segments come from the same speaker. The state-of-art speaker diarization systems can de divided into three categories according to the approach used for speaker segmentation: metric-based methods, model-based techniques and hybrid one. The segmentation process can be done in a single pass or in multiple passes through the acoustic data. The most widely applied approach of clustering used in diarization systems is the hierarchical agglomerative clustering with a BIC based stopping criterion.

In the decoding process, the main problem to solve is to achieve an adequate treatment for the high variability of the acoustics, language and vocabulary. Acoustic mismatches appear due to variations in the environmental noise or because of the inherent inter-speaker variability. Modern recognizers include parameterizations that partially incorporate characteristics of the human ear, as well as speech enhancement techniques that increase the robustness against noise (MFCC or PLP front-ends). Regarding the inter-speaker variability, normalization of the length of vocal tract (VTLN) and the adaptive training of speaker are common approaches. Most modern recognition systems also employ a multi-pass strategy, in which the acoustic models are adapted using MLLR,or MAP algorithms, in an unsupervised way. A confidence measure may be used to exclude the audio segments that are more likely to be incorrectly transcribed.

Linguistic mismatches, mainly caused by changes in style and topic, are also commonly tackled using unsupervised adaptation strategies, while vocabulary mismatches (a particularly serious problem with agglutinative languages like Basque, or inflecting languages like Spanish and Galician) are palliated using statistical methods to compose and select the vocabulary. The recognition of spontaneous speech shows a specific and complex problematic due to the presence of "filled pauses", hesitations, repetitions, false starts, etc. These phenomena are very difficult to model, both from the lexical point of view, but also from the linguistic one. Current research includes the use of prosodic information and specialized language modelling.

2.2. Machine Translation Technologies.

Machine Translation (MT) constitutes a research area that has gained much attention worldwide during the last years. Specifically in the European Community, where the language diversity still represents an important drawback for the integration process, a large amount of resources has been invested in R&D in this technology. As a representative of such an effort, the following projects can be mentioned: C-Star [6], Eutrans [6], Verbmobil [8], LC-Star [9], Nespole! [10], Fame [11], TC-Star [1], SMART [12] and EURO-MATRIX [13]. However, no matter the recent progress, MT technology is still far from achieving satisfactory performance and quality levels.

From the point of view of practical applications, MT can be categorized into two specific problems: written language machine translation (WLMT) and spoken language machine translation (SLMT). The problem of spoken language translation (SLMT), in addition to the specific MT complexities, involves two additional problems: first, those related to the nature of spoken language, such as spontaneity and poor structured (or even lack of) syntax; and second, those related to the state-of-the-art in automatic speech recognition technologies (ASR), such as recognition errors. These problems explain why statistical methods outperform rule-based methods in speech translation. Additionally, in recent years, the second problem had a relevant impact in the integration of ASR and MT systems. In the case of WLMT written language is much more controlled with respect to sentences' grammar content. This allows MT algorithms to explore richer linguistic information starting from lexical level up to the syntactical and semantic levels. However, during the last years the results of international evaluation campaigns carried out by NIST [5] or promoted by specialized conferences (IWSLT09, ACL08) showed that the performance of statistical machine translation systems is comparable to the performance of ruled-based systems, and can be even better when working on restricted domains.

This project considers the MT problem for both, WLMT and SLMT. It focuses on translation tasks among three languages spoken in Spain (Spanish, Galician and Catalan), as well as on translation tasks between them and English, as far as bilingual resources and methodologies are available to the project. Continuing the previous work of the consortium on MT, the selected approach is Statistical Machine Translation. Further effort will be invested in incorporating, using, and adapting available knowledge-based tools and linguistic resources. In particular, special attention will be paid to the reordering problem and to the integration of morpho syntactic knowledge and bilingual dictionaries.

2.3. Speech Synthesis

The synthetic speech produced by state-of-the-art TTS systems does not always fit the actual speech task. Considering speaking styles in synthetic speech produces an improvement in its quality and naturalness. If the desired style deviates only slightly from the standard speech of the TTS system, the style can be simulated with adapted prosody or generating the correct prosody using corpus based techniques. In other cases deeper adaptations are needed and spectral features as well as prosodic features must be controlled. Some works propose speaking style interpolation and adaptation for HMM-based speech synthesis. In most cases flexible and high quality speech synthesis techniques like HMMs are used.

In the building of voice transformation systems different speech models and analysis/reconstruction techniques allowing spectral transformations have been successfully applied: LP-PSOLA or FD-PSOLA, sinusoidal harmonic models, hybrid models, and STRAIGHT. In general, the voice conversion functions are trained from a set of aligned acoustic features that captures the source-target phonetic correspondence. Several alignment strategies can be adopted, depending on the requirements of the spectral envelope transformation method applied by the system. These strategies involve mapping between acoustic classes, or frame-level alignment when a parallel training corpus is available, or alignment techniques for non-parallel training corpora, although some systems do not require an explicit source-target alignment. There is still a trade-off between the quality of the converted speech and the similarity between converted and target voices. Despite the recent appearance of new methods with very good similarity-versus-quality balance, a higher quality is still desirable in some applications.

There is a lack of general agreed criteria for the evaluation of TTS systems. One attempt to establish a common framework for TTS evaluation is the Blizzard Challenge organized since 2006 [14] where the participants must build a TTS system using the same database. In a similar way, Albayzin 2008 TTS evaluation campaign was organized to evaluate Spanish TTS systems. This kind of evaluation campaigns is costly in time and effort and some objective evaluation method that complements them is desired.

3. Objectives

The objectives of the project can be sumarized as making advanced research in all the spoken language techologies necessary to create a multilingual and multimodal information retrieval system. This ambitious objective implies the improvement of all the involved technologies and the specific problems created due to their interaction. Specifically:

- In speech recognition it is necessary to work in unconstrained domains. It is necessary to improve the robustness of the system against speakers, noise, and unknown vocabulary. Diarization techniques should be improved and closely interact with the ASR system.
- In speech translation is necessary to improve the current performance when the SMT system works in unconstrained domains and with automatic transcriptions from oral discourses. Lack of punctuation, false starts, hesitations and recognition errors reduce the performance of the SMT systems. New algorithms and knowledge integration should be incorporated.
- In speech synthesis, expresive speech and high quality text to speech conversion are the main goals to be achieved. Machine learning techniques will be applied to improve prosody and speaking styles. In a speech to speech translation system, the purpose is that the translated voice be able to imitate the style, and even the voice of the source speaker.

In the field of information retrieval, a state of the art system will be implemented. New techniques will be studied to improve the rank of the retrieved information.

3.1. Specific objectives in Diarization, Segmentation and Speech Recognition

The objectives in this area are to improve the performance of the University of Vigo (UVIGO) ASR system for Spanish and Galician languages, to develop a new system for Basque language and to implement a spoken query recognizer. These main objectives may be split in the following sub-objectives

- Development of a robust audio pre-processing module for audio classification and speaker segmentation and classification
- Improvement of the UVIGO LVCSR performance for spontaneous speech by means of automatic detection of speech disfluencies.
- Development of confidence measures for continuous speech recognition.
- Incorporation of unsupervised acoustic adaptation mechanisms using confidence measures in the LVCSR.
- Development of statistical methods for optimizing the recognition vocabulary in flexible and agglomerative languages.

3.2. Specific objectives in Speech translation

The main objective in Speech Translation is to develop high quality statistical machine translation (SMT) systems for all the pairs among the Spanish, Galician, Catalan and English languages. This main objective can be divided into the following sub-objectives:

- Development of algorithms concerning both, word reordering and discriminative training to improve the translation system performance.
- Integration of morpho-syntactic knowledge into SMT.
- Development of an efficient methodology to use Spanish as a pivot language.
- Develop a speech to speech translation system. For this purpose, the integration of all the speech technologies involved in the project is needed

3.3. Specific objectives in Speech Synthesis

The main objective is to develop high quality expressive text-to-speech synthesis systems for the five languages considered in this project. Sub-objectives include the following,

- Characterization and generation of speaking styles, mainly by the use of the correct modelling of the prosody.
- Development of a voice transformation system allowing the modification of any aspect of the synthetic speech signal that conveys expressivity and speaker identity.
- Development of flexible and high quality TTS systems, considering harmonic/stochastic models and HMM techniques.
- Organization and participation in TTS evaluation campaigns with the goal of defining objective criteria that complement subjective evaluation results.

3.4. Technologies integration

In this task, speech recognition, translation, synthesis technologies and the information retrieval engine are integrated in order to build a complete search system. The achievement of his general objective implies the fulfilment of the following sub-objectives:

- Development of bilingual text resources for Spanish – English/Catalan/Basque/Galician language pairs, including parallel texts and language processing tools.
- Specification and development audiovisual signal processing tools to generate metadata for an audiovisual repository. Metadata includes diarization, speech recognition and translation of the audiovisual material.
- Implementation of a web based architecture to put into communication the different subsystems.
- Definition and implementation of efficient interfaces among technologies that improve the performance of the overall system.

- Implementation of tools for building a voice and text driven engine for audiovisual information retrieval search among the official languages in Spain (Spanish, Catalan, Galician and Basque) and English. The engine will be a demonstrator of the developed technologies. This demonstrator will be part of the project dissemination.

4. The partners of the project

The objectives described in this project are multilingual, multimodal and interdisciplinary. They require players and organizations with varying skills and expertise in a wide variety of domains; e.g. speech recognition, translation, synthesis, speech and text processing, implementation of search engines, information retrieval, production of language resources, etc. The requirements to carry out such a project are therefore considerable and the feasibility of such a project relies on the close cooperation of different institutions and the support of external companies. In this project each group will contribute with its specific and complementary technological expertise as well as with its previous experience with the various languages involved.

The involved groups are the TALP research center from the Technical University of Catalonia (TALP-UPC), the Aholab Signal Processing Laboratory from the Universidad del País Vasco (EHU-Aholab) and, the Signal Theory Group from the Universidad de Vigo (UVIGO). The Information Retrieval Group from Barcelona Media – Centre d'Innovaciò (BMCI) will collaborate in the project as an external partner and will contribute to the objectives of the project with their experience in Information Retrieval tools.

TALP-UPC [15] provides expertise in three areas: speech recognition, with competitive systems for telephony applications, and a continuous speech recognition system, both in Catalan, Castilian and English; speech translation in Catalan, Castilian and English; and speech synthesis in Castilian and Catalan. TALP-UPC coordinates the translation activities and integration technologies.

UVIGO [16] brings experience in continuous speech recognition in a complex task such as broadcast news in Galician and Castilian, supplementing TALP-UPC system perfectly. The group has sound experience in speech synthesis in both languages Galician and Spanish. UVIGO coordinates the work of speech recognition.

EHU-Aholab [17] is an expert group in speech synthesis and specifically, in research on prosody. Their synthesis system works in Euskera and Castilian. The group has also developed a speech recognition system in Euskera. EHU-Aholab coordinates the activities of speech synthesis.

The Information Retrieval Group of Barcelona Media – Centre d'Innovaciò (BMCI) [18] is composed of a multidisciplinary team of advanced and starting researches that allows for concentrating, into a single research group, first class expertise and experience in the areas of: information storage and retrieval, natural language processing, machine translation, artificial intelligence, signal processing and data mining technologies. They will set up the search engine in the show case.

5. Acknowledgements

This Project is supported by the Spanish Government grant number TEC2009-14094-C04.

The text of this paper has been extracted from the Technical Annex of the BUCEADOR project. Consequently, all the research team has contributed to it

6. References

[1] http://www.tc-star.org (2004/2007)

[2] http://chil.server.de

[3] http://www.amidaproject.org (EC IST Integrated Project., 2006-2009)

[4] http://www.emime.org (ICT-2007.2.1)

[5] NIST http://www.nist.gov/speech/tests/mt/

[6] http://www.c-star.org/, (1991/2004)

[7] http://cordis.europa.eu/esprit/src/30268.htm (SPRIT-20268-96; SPRIT-30268, 1997/2000)

[8] http://verbmovil.dfki.de/overview-us.html, (1993/2000)

[9] http://www.lc-star.com (IST-2001-32216, 2002/2004)

[10] http://nespole.itc.it/, (IST-1999-11562, 2001/2004)

[11] http://isl.ira.uka.de/fame (IST-2000-28323, 2001/2004)

[12] http://www.smart-project.eu/node/1

[13] http://www.euromatrix.net/ (P6-IST-5-034291-STP)

[14] A.W: Black,.K. Tokuda "The blizzard challenge - 2005: evaluating corpus-based speech synthesis on common datasets", In Proc. INTERSPEECH, pp. 77-80, 2005.

[15] http://www.talp.cat/talp/

[16] http://www.gts.tsc.uvigo.es/web/index.php

[17] https://bips.bi.ehu.es/aholab/

[18] http://*www.barcelonamedia.org*

PhD Dissertation: Intra-lingual and Cross-lingual Voice Conversion using Harmonic plus Stochastic Models

Daniel Erro[1,2], supervised by Asunción Moreno[1]

[1] TALP Research Center, Technical University of Catalonia (UPC), Barcelona
[2] AHOLAB Signal Processing Laboratory, University of the Basque Country (UPV/EHU), Bilbao
derro@aholab.ehu.es, asuncion.moreno@upc.edu

Abstract

This PhD dissertation, written by Daniel Erro and supervised by Asunción Moreno, was defended on June 16th 2008 at UPC. The committee members were Antonio Bonafonte (UPC), Helenca Duxans (TID), Inma Hernáez (EHU), Eduardo Rodríguez-Banga (UVIGO), and Xavier Serra (UPF). Qualification: "sobresaliente cum laude".

Index Terms: voice transformation and conversion, speech synthesis, harmonic plus stochastic model

1. The Voice Conversion Problem

Voice conversion (VC) is the technology used to transform the voice of one speaker (the source speaker) for it to be perceived by listeners as if it had been uttered by a different specific speaker (the target speaker). Among all the speaker-dependent voice characteristics, voice conversion focuses mainly on the acoustic ones: the spectral characteristics and the fundamental frequency. During training, given a certain amount of training data recorded from specific source and target speakers, the system determines the optimal transformation for converting one voice into the other one. Then, the system can apply this transformation to convert new input utterances of the source speaker.

VC has a wide variety of applications, including the design of multi-speaker speech synthesis systems without strong memory requirements, the customization of speaking devices, the design of speaking aids for people with speech impairments, film dubbing using the original actors' voices, the creation of virtual clones of famous people for videogames, masking identities in chat rooms, etc.

2. State-of-the-Art before this Thesis

At the time of beginning this thesis, state-of-the-art VC systems were mainly based on the statistical framework proposed by Stylianou [1] and modified by other authors [2][3][4]. Such systems produced satisfactory results in terms of similarity between converted and target voices, but degraded noticeably the quality of the speech signals. In fact, considering different VC systems and methods, a trade-off could be observed between these two performance dimensions (i.e. converted-target similarity and quality). Therefore, the design of new VC methods characterized by a good similarity-quality balance was one of the main challenges.

The versatility of state-of-the-art VC systems was often limited by their requirements for estimating adequate transformation functions from the training data. A vast majority of them could operate only when a parallel training corpus was available (in other words, when the training sentences uttered by the source and target speaker were exactly the same and therefore showed a clear phonetic correspondence). Although some techniques compatible with non-parallel corpora (different source and target utterances) had been already proposed, being some of them valid for cross-lingual VC, they required some extra information for a correct performance (other pre-trained VC functions [5], phonetic labels [4], etc.) or either the performance scores of the VC system decayed under some conditions [6]. In order to build more versatile VC systems capable of learning transformation functions flexibly under different training conditions, a new source-target alignment procedure was desirable.

Other unsolved problems related to VC were prosody conversion and robustness against data reduction.

3. Objectives and Methodology

The general objective of this thesis was to research into VC systems and methods in order to improve their quality and versatility. The specific objectives were the following:

- Design of new VC methods that succeeded at converting the source voices into the target voices without degrading significantly the quality of the signals.
- Design of training methods that made the VC system capable of operating in all possible training scenarios: intra-lingual scenario with parallel corpus, intra-lingual scenario without parallel corpus, and cross-lingual scenario.
- Integration of the resulting VC system into a text-to-speech (TTS) synthesis system, so that it could operate not only as a conversion device, but also as a stand-alone TTS that generated different voices from a single synthesis database.

The fulfillment of the described objectives was to be verified by means of perceptual tests: both the similarity between converted and target voices and the quality of the converted speech had to be rated by human listeners. The Mean Opinion Scores (MOS) were chosen as figure of merit.

The database used in the experiments, which had been specifically recorded for research on voice conversion in the framework of the TC-STAR project, contained around 200 sentences in Spanish and 170 in English, uttered by 4 different professional bilingual speakers (2 male + 2 female speakers). The average duration of the sentences was 4 seconds (10-15 minutes of audio per speaker and language).

The research work carried out during the thesis can be summarized in the following steps:

- Design and implementation of a speech model suitable for speech analysis-reconstruction, pitch and duration modification, synthesis, and spectral manipulation.
- Implementation and subjective evaluation of a baseline VC system using the mentioned speech model, state-of-the-art VC techniques, and parallel training corpora.
- Research on strategies for improving the quality scores of the baseline system without worsening the conversion scores. Subjective evaluation of the resulting method.
- Research on strategies to allow the system to train VC functions from non-parallel corpora while maintaining its subjective performance scores. Evaluation under both, intra-lingual and cross-lingual conditions.
- Implementation and evaluation of a multi-speaker TTS system and optimization of the interaction between the synthesis process and the VC process.

The next section presents an overview of the main contributions of this thesis.

4. Main Contributions of the Thesis

4.1. Flexible Harmonic/Stochastic Model

A new speech model based on a harmonic plus stochastic decomposition was developed during the thesis. This model allowed the manipulation of all kind of signal features with a high degree of flexibility, which was desirable for implementing VC systems. The novelty of the model lay in the algorithms for time-scale manipulations, pitch-scale manipulations, and concatenation of units, which were compatible with a non-pitch-synchronous analysis scheme. The reasons for preferring a constant analysis frame rate rather than a pitch-synchronous rate were augmenting the flexibility and simplifying the analysis (because the accurate separation of the signal periods was not a previous requirement). In exchange, in order to make artifact-free speech modification possible, the problem of estimating and manipulating the linear-in-frequency phase term of the speech frames without producing artifacts was faced. In contrast to previous non-pitch-synchronous models based on sinusoidal or hybrid decompositions, it was not necessary to use onset times or pitch-synchronous epochs as a reference. The use of computationally expensive techniques such as inverse filtering was also avoided. Instead, amplitude and phase envelopes were used as estimators of the vocal tract, assuming a simplified speech production model. A new method for removing the linear phase term from a set of measured harmonics was also proposed.

In order to validate the suitability of the new model for high-quality speech transformations, it was integrated into the waveform generation module of a concatenative TTS system which was then compared to an equivalent TD-PSOLA-based system. The results showed that the listeners had a clear preference for the new system when the synthesis required applying high modification factors. It was concluded that the speech model was valid for high-quality speech transformation and concatenation and provided a very good framework for research on voice conversion.

4.2. Weighted Frequency Warping

As mentioned before, state-of-the-art VC systems were based on statistical methods. Most of them followed Kain's approach [2], in which a set of paired vectors was extracted from parallel utterances of the source and target speakers and a joint Gaussian mixture model (GMM) was fitted to them. The mean vectors and covariance matrices provided by the GMM were used to define the conversion function, a weighted combination of linear transforms:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i(\mathbf{x})\left(\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx}\boldsymbol{\Sigma}_i^{xx^{-1}}\left(\mathbf{x}-\boldsymbol{\mu}_i^x\right)\right) \tag{1}$$

where $\mathbf{x}$ is the source acoustic vector to be converted; $p_i(\mathbf{x})$ is the probability that $\mathbf{x}$ belongs to the i-th mixture of the GMM, given by the mean $\boldsymbol{\mu}_i^x$ and the covariance matrix $\boldsymbol{\Sigma}_i^{xx}$; $\boldsymbol{\mu}_i^y$ is the target mean vector and $\boldsymbol{\Sigma}_i^{yx}$ is the cross-covariance matrix. This method was known to yield a successful conversion of voices but significant speech quality degradation due to statistical over-smoothing.

When the above mentioned method was implemented as baseline system, a high correlation was observed between the spectral envelopes given by the source and target mean vectors, $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$, for all the mixtures of the GMM. A simple frequency-warping transformation (a mapping between source and target frequency axes) of the source mean envelopes appeared to yield good estimates of the target mean envelopes. This was an interesting finding, since frequency warping of spectra was known to introduce very little quality degradation. Since vectors inside each Gaussian component of the GMM contained parametric representations of phonemes with similar formant structures, it was assumed that a single frequency warping function could be valid for all the spectra belonging to that class.

The resulting VC method, which was called Weighted Frequency Warping (WFW), can be described as follows. During training, once a joint GMM has been estimated from a set of paired vectors, the spectra given by $\boldsymbol{\mu}_i^x$ and $\boldsymbol{\mu}_i^y$ are used to estimate a warping function $W_i(f)$ for each mixture of the GMM (see Figure 1). During conversion, given the spectrum of the k-th frame, $S^{(k)}(f)$, and its parametric representation $\mathbf{x}^{(k)}$, the following transformation is applied to obtain the converted spectrum $S'^{(k)}(f)$:

$$S'^{(k)}(f) = G^{(k)}(f)\cdot S_{fw}^{(k)}(f) \tag{2}$$

where $S_{fw}^{(k)}(f)$ is a frequency-warped version of the original spectrum, calculated by applying a weighted combination of the trained warping trajectories $\{W_i(f)\}$,

$$S_{fw}^{(k)}(f) = S^{(k)}\left(W^{(k)^{-1}}(f)\right) \tag{3}$$

$$W^{(k)}(f) = \sum_{i=1}^{m} p_i\left(\mathbf{x}^{(k)}\right)\cdot W_i(f) \tag{4}$$

and $G^{(k)}(f)$ is a smoothed correction filter that compensates the differences in amplitude between the warped spectrum $S_{fw}^{(k)}(f)$ and the real target spectrum, estimated from vector $F(\mathbf{x}^{(k)})$ given by expression (1).

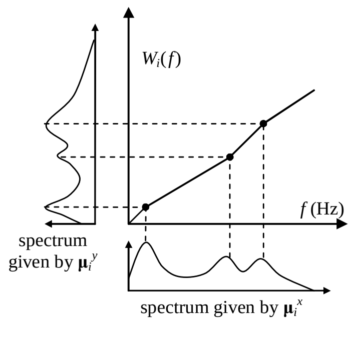


Figure 1: *Estimating a frequency warping function from the mean spectra of the i-th mixture of a GMM.*

The subjective evaluation of a WFW-based VC system confirmed that the method provided much better quality than standard GMM-based systems. The average improvement was around 0.7 points in a 1-to-5 MOS scale, whereas the similarity scores remained almost unaltered. The average quality level achieved by WFW in these experiments was close to 3.5 in a 1-to-5 MOS scale, which could be considered acceptable for real voice conversion applications. The system achieved also excellent results in the 3rd international TC-STAR evaluation campaign.

4.3. Iterative Alignment Method

A new solution was proposed to allow the system to train VC functions when no parallel corpus was available. In other words, it made possible the correct alignment between source and target acoustic vectors extracted from text- and language-independent utterances. The idea came from the observation that when a simple nearest neighbour (NN) criterion was used to pair the source and target vectors during training, an intermediate converted voice was obtained (closer to the target, but not very different from the source). If a NN alignment was carried out again on the intermediate converted vectors and the target vectors, the resulting voice got slightly closer to the target voice. Therefore, the proposed method consisted of iteratively repeating these two steps: 1) NN alignment; 2) training of a VC function, and conversion of the source vectors. This idea is described graphically in Figure 2.

Some experiments using parallel test corpora and several objective measures were conducted in order to check that the alignment got more accurate (compared to the parallel corpus case) as the number of iterations increased. At the same time, several aspects of the method were studied using the same objective measures: configuration of the NN search, convergence, stopping criterion, initialization, etc. The objective performance of the alignment system resulted to be close to that of parallel alignment in most cases. Although the speech frames were being aligned using only acoustic information extracted directly from the signal, the experiments showed that the alignment was acceptable also from a phonetic point of view.

In order to evaluate the new alignment method through perceptual tests, it was integrated into the training module of a VC system. In an intra-lingual context, the results were found to be highly satisfactory, similar to those obtained by an equivalent voice conversion system using parallel training corpora. The performance under cross-lingual conditions was slightly worse, due to the phonetic differences between languages. A cross-lingual VC system that resulted from the combination of the described alignment method and WFW (see section 4.2) participated in a public evaluation campaign organized under the authority of the European TC-STAR project. It achieved the best results considering both the similarity of voices and the quality of the converted speech.

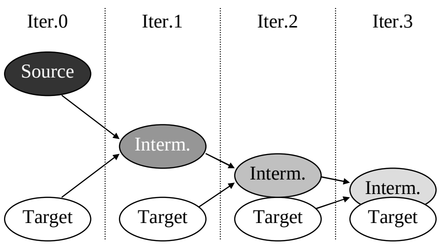


Figure 2: *Underlying idea of the iterative alignment method.*

4.4. Multi-speaker TTS System

A multi-speaker synthesis system was built by combining the UPC TTS system with a VC system based on the methods and algorithms presented throughout this thesis. The full system was evaluated by means of perceptual tests. As expected, the scores indicating the similarity between converted and target voices were very close to those obtained when converting natural speech utterances (approximately 3.1 in a 1-to-5 MOS scale). The quality of the converted synthetic signals was affected by both synthesis and conversion, so the resulting average score was 2.5 (approximately 1 point below the one obtained for converted natural speech and 0.5 points below the one obtained for non-converted synthetic speech). The results revealed also that the choice of the source voice had a direct influence on the performance of the system.

5. General Conclusions

As a result of this thesis, contributions were made in many parts of the voice conversion process. The analysis, manipulation and reconstruction of speech signals were improved through the design of a new speech model based on a harmonic/stochastic decomposition. A new spectral conversion method, WFW, was proposed to increase the quality of the converted speech with respect to state-of-the-art methods and alleviate the over-smoothing effect. An iterative alignment method was also proposed to overcome the lack of versatility of VC systems when no parallel training corpus was available. All these ideas were put into practice in a multi-speaker TTS system. All these methods and systems yielded highly satisfactory performance scores as shown by different subjective listening tests.

6. Publications and Merits

The main contributions of the thesis [7], summarized in section 4, were published in two journal papers in IEEE Trans. Audio, Speech, and Lang. Proc. [8][9], and also in several papers in the most important conferences related to speech technologies, being the most relevant ones the following: [10][11][12][13]. In addition, the author was invited by the Music Technology Group of the Pompeu Fabra University to give a talk on the results of this thesis [14]. Some indirect results of the thesis contributed to the publication of two more journal papers: in [15] the methods described in section 4.1 and 4.2 were used in an emotion conversion application; in [16] the VC system was used to evaluate the robustness of a speaker recognizer. In one of the two papers awarded during last JTH [17] (best paper award), the parameter extraction module of the VC system was applied to build an HMM-based synthesizer, outperforming other traditional methods. Moreover, the VC system developed during this thesis took part in different public evaluation campaigns. It is worth mentioning the diploma obtained in Albayzin 2006 evaluation (which consisted in cheating a biometric system based on speaker recognition), and the excellent results achieved in the 3rd international evaluation campaign of the EU funded project TC-STAR [18] (best results in cross-lingual voice conversion categories). Finally, the findings in this thesis contributed to make progress in the framework of three different funded projects: TC-STAR (FP6-506738, European Comission), AVIVAVOZ (TEC2006-13694-C03, Spanish Ministry of Science and Education) and Tecnoparla (Catalan Government).

7. References

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Trans. Speech and Audio Proc., vol. 6, pp. 131–142, 1998.

[2] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Sci. and Eng., Beaverton, OR, 2001.

[3] T. Toda, A.W. Black, K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter", Proc. IEEE ICASSP, vol.1 pp.9-12, 2005.

[4] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," IEEE Trans. Audio, Speech and Lang. Proc., vol. 14, no. 4, pp. 1301–1312, 2006.

[5] A. Mouchtaris, J. Van der Spiegel, P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach", IEEE Trans. Audio, Speech and Lang. Proc., vol. 14, no. 3, pp. 952–963, 2006.

[6] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, A. W. Black, S. Narayanan, "Text-independent voice conversion based on unit selection", Proc. IEEE ICASSP, vol. 1, pp. 81–84, 2006.

[7] D. Erro, "Intra-lingual and Cross-lingual Voice Conversion using Harmonic plus Stochastic Models", PhD thesis, Technical University of Catalonia (UPC), 2008.

[8] D. Erro, A. Moreno, A. Bonafonte, "Voice Conversion based on Weighted Frequency Warping", IEEE Trans. Audio, Speech, and Lang. Proc., vol. 18(5), pp. 922-931, 2010.

[9] D. Erro, A. Moreno, A. Bonafonte, "INCA Algorithm for Training Voice Conversion Systems from Nonparallel Corpora", IEEE Trans. Audio, Speech, and Lang. Proc., vol. 18(5), pp. 944-953, 2010.

[10] D. Erro, T. Polyakova, A. Moreno, "On combining statistical methods and frequency warping for high-quality voice conversion", Proc. IEEE ICASSP, pp. 4665-4668, 2008.

[11] D. Erro, A. Moreno, "Weighted Frequency Warping for Voice Conversion", Proc. Interspeech, pp. 1965-1968, 2007.

[12] D. Erro, A. Moreno, "Frame Alignment Method for Cross-Lingual Voice Conversion", Proc. Interspeech, pp. 1969-1972, 2007.

[13] D. Erro, A. Moreno, A. Bonafonte, "Flexible Harmonic/ Stochastic Speech Synthesis", Proc. 6th ISCA Speech Synthesis Workshop, 2007.

[14] D. Erro, "Voice conversion: state of the art, recent improvements and remaining challenges", seminar at Pompeu Fabra University, Barcelona, 2009.

[15] D. Erro, E. Navas, I. Hernáez, I. Saratxaga, "Emotion Conversion based on Prosodic Unit Selection", IEEE Trans. Audio, Speech, and Lang. Proc., vol. 18(5), pp. 974-983, 2010.

[16] M. Farrús, M. Wagner, D. Erro, J. Hernando, "Automatic Speaker Recognition as a Measurement of Voice Imitation and Conversion", International Journal of Speech, Language and the Law, vol. 17(1), pp. 119-142, 2010.

[17] E. Banos, D. Erro, A. Bonafonte, A. Moreno, "Flexible Harmonic/Stochastic Modelling for HMM-based Speech Synthesis", Proc. V Jornadas en Tecnologías del Habla, pp. 145-148, 2008.

[18] D. Mostefa, O. Hamon, N. Moreau, K. Choukri, Evaluation Report, Deliverable D30 of the EU funded project TC-STAR, http://www.tc-star.org, 2007.

Feature analysis and evaluation for automatic emotion identification in speech

Iker Luengo, Eva Navas

University of the Basque Country
Alda. Urquijo s/n 48013 Bilbao (SPAIN)
iker.luengo@ehu.es, eva.navas@ehu.es

Abstract

This PhD dissertation was written by Iker Luengo under the advice of Dr. Eva Navas. It was successfully defended in the University of the Basque Country the 1st of June, 2010. The doctoral thesis committee was composed by Prof. José B. Mariño as president, Prof. Inmaculada Hernáez as secretary, Prof. Carmen García, Dr. Valentín Cardeñoso and Dr. Laura Docío.

Index Terms: emotion identification, parameterisation, feature selection

1. Introduction

Features extracted from the speech signal have a great effect on the reliability of an emotion identification system. Depending on these features, the system will have a certain capability to distinguish emotions, and will be able to deal with speakers not seen during the training. Many works in the field of emotion recognition are aimed to find the most appropriate parameterisation, yet there is no clear agreement on which feature set is best.

Going over the literature shows that each research group uses a different parameter set. Features extracted from the spectral envelope, prosodic characteristics, glottal flow or even the linguistic content are used indiscriminately, in an attempt to detect those that really are useful. The approach that is most often used is to define a large parameter set and feed it to an automatic feature selection algorithm that heuristically selects the most discriminant ones. i.e., let the data tell you what is relevant and what not [1, 2]. Unfortunately, this features selection step is usually seen as an unavoidable nuisance needed in order to maximise the accuracy, not as a useful tool for the enlightening of the relation between the features and the emotions. Only few papers show the results of this selection, and almost none discuss them, so it is not possible to know if prosody, spectral envelope or voice quality features were the ones that really provided information to the system.

It is widely accepted that prosodic features carry most of the emotional information in the speech. For many years automatic identification systems have used prosody almost exclusively, leaving spectral characteristics in a second term [3, 4]. There are several reasons behind this idea. On the one hand, the theories describing the physiological changes caused by an emotional state [5, 6] focus mostly on changes in the air pressure and vocal cord tension. On the other hand, many studies directly compare speech signals with different emotional content in order to find measurable differences [7, 3]. These studies usually conclude that, taken individually, the differences from one emotion to another are larger for prosodic features than for spectral ones. But the behaviour of the complete set of features is not analysed, so the complete set of spectral characteristics may provide more information than the complete set of prosodic features. In fact, it is nowadays very usual to find papers successfully using spectral features [8, 9, 10] or voice quality characteristics [11, 12], thus proving that these kind of parameters are also important in the classification of emotions.

Furthermore, no systematic study of the effectiveness of each parameterisation has been performed. Such a study could identify which is the most appropriate parameterisation for the automatic identification of emotions. There are some works in the literature that analyse the behaviour of different feature sets, but they do not provide a complete view of the problem. Many of these studies treat each feature individually, which may provide conclusions that are not generalisable when they are combined with others. Other works provide experimental results with complete parameterisations, but not separately, so it is not possible to deduce whether the combination of features was really better than the features (or a feature subset) alone.

Finally most of the works in the literature are not comparable among them, as they use different speech databases, different methodologies or different number of emotions. As a result, it is impossible to build a complete view of the properties of different parameterisations comparing the results of such works.

2. Objectives

The work presented in this dissertation attempts to fill the existing gap regarding the effectiveness of the different acoustic features for the recognition of emotions in speech. It presents a systematic analysis of the acoustic parameterisations that are used most for emotion identification (spectral, prosodic and voice quality characteristics), providing a complete description of their effectiveness for this task.

The purpose is to describe the effectiveness of isolated features as well as the behaviour of different sets of features. Therefore, individual parameterisations and their combinations have been analysed.

A special care was taken so that the obtained results can be comparable among different features. A common database and methodology was used all along the process in order to ensure this.

Furthermore, these features are supposed to work in real-life automatic emotion identification systems. Therefore all the parameterisation process had to be made completely automatic, without manual corrections. Some of the algorithms used during this process had to be modified, or new ones had to be developed, in order to make the process robust enough to work with natural emotions and spontaneous speech.

3. Methodology

3.1. Databases

The analysis of the features and the experiments were repeated using two publicly available databases. The first one, *Berlin EMO-DB* [13], is an acted emotional speech database, that contains recordings from 10 male and 10 female speakers. Each speaker repeated the same 10 sentences simulating seven different emotional states: anger, boredom, disgust, fear, happiness, neutral and sadness.

As it contains parallel corpora for each emotion, this database makes it easier to compare the different characteristics of the emotions. Therefore, it was used to make the first estimation of the discriminality of the parameterisations.

In order to validate the results obtained with acted speech, and to see whether the conclusions can be applied to natural emotions, a second analysis was carried out using the *FAU-Aibo* database [14]. This one contains spontaneous speech and natural emotions recorded from 21 boys and 30 girls while they played with the Sony Aibo pet robot. The database contains almost 18,000 recordings distributed in four speaking styles: anger, emphatic, neutral and positive.

3.2. Processing of the recordings

The recordings were processed in order to get the characteristic curves and labellings needed for the extraction of the features. The processing included detection of the vocal activity, estimation of the glottal source signal and of the intonation curve, voiced-unvoiced labelling, pitch-period marking and vowel detection.

All this processing was performed automatically without manual corrections. In order to obtain reliable results under these conditions, some new algorithms had to be developed, and others had to be modified. These algorithms included:

- A new *vocal activity detector* (VAD), based on the LTSE-VAD [15]. The algorithm was modified so that the result is independent from the SNR of the signal, which is an important factor when using spontaneous speech. The resulting algorithm is described in [16].
- A new F_0 *estimator* and *voiced-unvoiced labeller*. The intonation curve was computed with the *Cepstrum Dynamic Programming* (CDP) algorithm [17], which uses the cepstrum transform and dynamic programming in order to estimate the F_0 value and the voice-unvoiced labelling at once. The paper describing the algorithm also presents experiments comparing it to other well-known pitch estimators, concluding that CDP has the best robustness with low SNR signals.
- A new *vowel detector* based on a phoneme recogniser working with models of clustered phonemes [18]. The clustering provides a consistent and very robust set of models that achieves high detection accuracy.

3.3. Considered features

The presented analysis is focused on acoustic parameters that can be extracted directly from the speech signal without a recognition step: spectral envelope, prosodic and glottal flow features. The parameters are divided according to their temporal structure into segmental and supra-segmental features. The diagram in Fig. 1 presents a schematic view of the parameterisation process.

Segmental features describe the evolution of the parameter over time. LFPC values [5] were used as representative of the spectral envelope, whereas instantaneous F_0 and intensity values were selected as *prosody primitives*, i.e., instantaneous '*prosodic*' features.

Supra-segmental features collect long-term information, estimated over time intervals longer than a frame. In this work, this interval was defined as the time between two consecutive pauses, as detected by the VAD algorithm.

For the long-term characterisation of the spectrum, different statistics of the LFPC were estimated. Similarly, prosodic information was extracted in the form of long-term statistics of the prosodic primitives. Last, voice quality features were also considered, and they were again defined as long-term statistics of various glottal source signal parameters, which were estimated by inverse filtering of the speech.

The combinations of the different information types were studied at parameter level (applying early fusion techniques) as well as at classifier level (using late fusion techniques). This allowed to analyse the discrimination capacity of different temporal structures created with the same parameterisations.

3.4. Feature analysis methods

The emotion discriminality of each parameterisation was studied using various techniques, each one of them providing a different insight about the characteristics of the features. This analysis was carried out both for individual parameters as well as for combinations of features, in order to check whether these combinations were useful or not.

3.4.1. Inter-class and intra-class dispersion

The intra-class dispersion represents the width of the distribution for a certain feature or feature set and for a given emotion. Therefore, it is a measure of the variability of the parameterisation. The inter-class dispersion represents the separation among the emotions that a feature or set of features provides. The relation between both dispersions provides a measure of the overlapping of the class distributions, i.e., the confusion probability among the emotions. This relation can be estimated using the J_1 criterion:

$$J_1 = \mathrm{tr}(S_W^{-1} \cdot S_B) \tag{1}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix and S_W and S_B are the intra-class and inter-class dispersion matrices respectively. The more separated the features are for each emotion, the higher this value is. Therefore, J_1 values were computed for each feature family, in order to estimate their capability to discriminate emotions.

3.4.2. Unsupervised clustering

Unsupervised clustering aims to divide the feature vectors into clusters according to their distribution, so that vectors that are close to each other are assigned to the same cluster, and vectors that are far away are assigned to different ones. Given a set of parametrised emotional speech recordings, if the emotional classes are correctly separated with that parameterisation, the resulting clusters should correspond to each emotion. Generally speaking, the fewer clustering errors that occur, the better the discrimination is. So, the clustering error can be used as another measurement for the estimation of the emotion discrimination. For this purpose, a k-means clustering was performed for each feature family, and the results were compared.

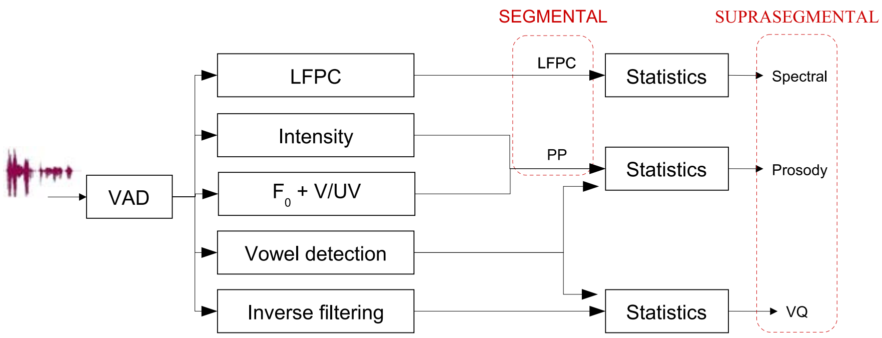


Figure 1: Schematic diagram of the parameterisation process.

3.4.3. Feature selection

A feature selection algorithm can help identifying the truly useful features, reducing the dimensionality of the parameterisation and making the classifier run faster and more accurately. Furthermore, detecting the discriminative features may provide a deeper understanding about the influence of the emotions in the acoustic characteristics of the voice.

The minimal-redundancy-maximal-relevance (mRMR) algorithm [19] was used to get a ranking of the features, from the most to the least significant one. mRMR was applied to all parameterisation families, as well as to their combinations, and the resulting rankings were carefully analysed in order to detect which features were most relevant individually and in combination with others.

3.4.4. Experimental evaluation

The final validation of the results obtained with the previous analysis was given by a series of experiments on automatic emotion identification using the various parameterisations and combinations of parameters that were considered. The experimental evaluation was carefully designed to be speaker independent, so that the results can be extrapolated to real-life conditions.

4. Results

4.1. Inter-class and intra-class dispersion

The J_1 values that were calculated show that spectral features provide larger separation among emotions than prosodic features. The emotion overlapping is highest when voice quality characteristics are used, which suggests that these kind of parameters are not suitable for automatic emotion identification.

As expected, the separation among the classes is smaller with real emotions and spontaneous speech than with acted speech. Nevertheless, the relation among feature sets is kept: spectral features separate emotions most and voice quality separates them least in both cases.

4.2. Unsupervised clustering

The results from the unsupervised clustering analysis are in accordance with J_1 values. Again, the number of clustering errors is larger with spontaneous speech than with acted speech. But in both cases the result using spectral features is better, with only a few recordings assigned to the wrong cluster, whereas the confusion is higher with prosodic features.

4.3. Feature selection

Results from the feature selection process are most interesting when the algorithm is applied to the combination of different feature sets. In these cases, the obtained ranking may show that the preferred features are of a certain nature, i.e., if spectral, prosodic or voice quality features are ranked in higher positions.

Regarding suprasegmental parameterisations, long-term spectral statistics are overall placed higher in the ranking than prosodic values. Voice quality features, instead, come out in the last positions. This effect is more evident with real emotions than with acted speech.

When the results of segmental parameterisations are analysed, it can be seen that short-term prosodic primitives are placed in quite good positions, although not in the best ones. In fact, when the ranking is performed over natural speech, prosodic primitives stay lower in the list, but still in a good place.

4.4. Experimental evaluation

The automatic emotion identification experiments confirm that features extracted from the spectral envelope of the speech are indeed more suitable for the task than parameters derived from the prosody or the voice quality. In fact, voice quality features have a really bad performance, and are almost of no use, even in combination with other kinds of parameters.

Prosodic characteristics, instead, may be useful if they are combined with spectral features. Nevertheless, this combination is relevant only in the case of acted speech. In the case of natural emotions, spectral features alone reach an accuracy similar to the one obtained with the combination.

5. Conclusions

The results from the analysis of the features reveal that, contrary to the most widely accepted theory, prosodic or voice quality features are not the most suitable ones for the automatic identification of emotions in speech. At least, not the kind of features that are typically used and that have been considered in this work. Spectral characteristics alone provide higher discrimination, and the combination of these spectral features with prosodic or voice quality ones does not improve the result. Although prosodic features may be useful at some extent when dealing with acted speech, they provide no accuracy improvement with natural emotions.

Most of the works that are focused on the analysis of features consider each parameter individually [20]. This way, they conclude that the variation from one emotion to another is larger for prosodic features than for spectral ones. But that is only applicable to individual parameters. The results from the inter-class and intra-class dispersion measures and the blind clustering suggest that, when the features are taken as a whole set, spectral characteristics provide more information about the emotion than prosodic ones.

The poor performance of prosodic and voice quality features is most probably due to the lack of robustness during their estimation from the speech signal. The low reliability of these features is more apparent with spontaneous speech. This effect is also shown in the described analysis, with prosodic and voice quality features performing better in acted speech than in spontaneous speech. The low reliability is far more noticeable for voice quality parameters, which are very difficult to obtain with enough robustness unless a manual supervision is applied.

It is not that features extracted from prosody or voice quality are useless. Several papers show that humans are able to identify emotions in prosodic copy-synthesis experiments [21], confirming that prosody does carry a great amount of emotional information, at least for some emotions. But the traditional prosodic representations may not be well suited to capture this information. On the one hand, long-term statistics estimated over the whole sentence lose the information of specific characteristic prosodic events. On the other hand, short-term prosodic primitives do not capture the prosodic structure correctly, which is suprasegmental by definition. The results suggest that a new more elaborate representation is needed to effectively extract the emotional information contained in the prosody, and that new and more robust algorithms are needed in order to capture this information with enough robustness.

Finally, the behaviour of the different sets of parameters for the discrimination of emotions has been similar both using acted and natural speech, with the expected loss of performance due to the more challenging problem that spontaneous speech poses. Therefore, acted speech databases can be used to perform the feature selection process and tune the automatic emotion identification system even if it has to deal with real emotions.

A paper describing part of this dissertation, and with the title "*Feature analysis and evaluation for automatic emotion identification in speech*" has been recently published in the IEEE Transactions on Multimedia, vol. 12, pp.490-501.

6. Acknowledgements

The author would like to thank the University of the Basque Country (UPV/EHU) for providing the grant and funding necessary to complete this work.

7. References

[1] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotional recognition," *Signal Processing*, vol. 88, pp. 2956–2970, Dec. 2008.

[2] S. Casale, A. Russo, and S. Serano, "Multistyle classification of speech under stress using feature subset selection based on genetic algorithms," *Speech Communication*, vol. 49, no. 10, pp. 801–810, Aug. 2007.

[3] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology*, vol. 26, no. 4, pp. 317–325, 2005.

[4] I. Luengo, E. Navas, I. Hernáez, and J. Sanchez, "Automatic emotion recognition using prosodic parameters," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 493–496.

[5] T. L. Nwe, S. W. Foo, and L. C. de Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, June 2003.

[6] C. E. Williams and K. N. Stevens, "Vocal correlates of emotional speech," in *Speech evaluation in Psychiatry*, J. K. Darby, Ed. New York, USA: Grune and Stratton, 1981, pp. 189–220.

[7] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Pathology*, vol. 70, no. 3, pp. 614–636, 1996.

[8] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features," in *IEEE Workshop on Multimedia Signal Processing*, Crete, Oct. 2007, pp. 48–51.

[9] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," *Lecture Notes on Computer Science*, vol. 4738, pp. 139–147, 2007.

[10] E. Navas, I. Hernáez, I. Luengo, I. Sainz, I. Saratxaga, and J. Sanchez, "Meaningful parameters in emotion characterisation," *Lecture Notes on Artificial Intelligence*, vol. 4775, pp. 74–84, 2007.

[11] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of all things the measure is man. automatic classification of emotions and inter-labeller consistency," in *ICASSP*, Philadelphia, USA, Mar. 2005, pp. 317–320.

[12] I. Luengo, E. Navas, and I. Hernáez, "Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge," in *Interspeech*, Brighton, UK, Sep. 2009, pp. 332–335.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, Lisbon. Portugal, Sep. 2005, pp. 1517–1520.

[14] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Information Society - Language TechnologiesConference (IS-LTC)*, Ljubljana (Slovenia), Oct. 2006, pp. 240–245.

[15] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long term speech information," *Speech Communication*, vol. 42, pp. 271–287, Apr. 2004.

[16] I. Luengo, E. Navas, and I. Hernáez, "Modified LTSE VAD algorithm for applications requiring reduced silence frame misclassification," in *Language Resources and Evaluation Conference (LREC)*, 2010, p. (To appear).

[17] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sánchez, and I. Sainz, "Evaluation of pitch detection algorithms under real conditions," in *ICASSP*, Honolulu, USA, Apr. 2007, pp. 1057–1060.

[18] I. Luengo, E. Navas, J. Sánchez, and I. Hernáez, "Detección de vocales mediante modelado de clusters de fonemas," *Procesado del Lenguaje Natural*, vol. 43, pp. 121–128, Sep. 2009.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[20] E. Navas, I. Hernáez, A. Castelruiz, J. Sánchez, and I. Luengo, "Acoustic analysis of emotional speech in standard Basque for emotion recognition." *Lecture Notes on Computer Science*, vol. 3287, pp. 386–393, Oct. 2004.

[21] E. Navas, I. Hernáez, and I. Luengo, "An objective and subjective study of the role of semantics in building corpora for emotional TTS," *IEEE transactions on Audio Speech and Language Processing*, vol. 14, no. 4, pp. 1117–27, Jul. 2006.

FALA 2010
VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Hierarchical Phrase-based Translation with Weighted Finite-State Transducers

Gonzalo Iglesias[⋆‡]
Advisors: *Adrià de Gispert*[⋆] and *Eduardo R. Banga*[‡]
[⋆]Department of Engineering, University of Cambridge, CB2 1PZ, U.K.
[‡]Dept. of Signal Theory and Communications, University of Vigo, Spain
gi212@eng.cam.ac.uk

Abstract

This thesis [1] on Statistical Machine Translation (SMT) has shown it is possible to combine succesfully hierarchical decoding with finite-state technology. On the other side, refinements on hierarchical grammars lead to faster translations with similar performance. The combined strategy has worked very well on several language pairs, including Spanish-English translation tasks, leading in many cases to systems that are search-error free, with state-of-the-art performance.

Index Terms: statistical machine translation, hierarchical decoding, finite-state transducers.

1. Introduction

Hierarchical phrase-based translation (*Hiero*) is one of the dominant current approaches to SMT [2]. *Hiero* systems not only incorporate many of the strengths of phrase-based systems, but also allow flexible word reordering based on a hierarchical grammar, which is a specific instance of a synchronous context-free grammar consisting of a set of rules $X \rightarrow \langle\gamma,\alpha\rangle$ and two special 'glue' rules to allow monotonic concatenation [3].

Hierarchical systems apply these rules guided by a context-free parser. Although the underlying idea is that both source and target languages should have very similar 'syntactic' trees, the use of abstract non-terminal symbols rather than more linguistically motivated ones endows this strategy with several advantages: importantly, it is possible to extract automatically the grammar in a similar way to phrase-based rules, but now also considering word gaps, and with several added constraints to ensure feasibility [4].

The cube-pruning decoder is a common decoding strategy used to handle these grammars. For these systems, a monolingual source parse is performed first; then, by traversing backpointers of the parse forest, the translation search space is built. During this construction the cube-pruning technique is used in each cell, in order to make the procedure tractable; but of course not only this comes at a cost in terms of speed. As search errors inevitably appear, there is also a risk of degrading performance.

In this context, this thesis [1] shows that it is possible to use compact efficient lattice representations of the translation hypotheses within hierarchical decoding that lead to search-error free translation strategies. Furthermore, we use weighted finite-state transducers (WFSTs) to represent these lattices, with the advantage of powerful and efficient operations such as determinization, minimization or composition over WFSTs [5]. The result is a decoder named *HiFST*. Combined with a special refinement to hierarchical grammars which we call *shallow grammars*, *HiFST* has performed very well on several translation tasks.

This paper is organized as follows. In Section 2 we describe *HiFST*. In Section 3 we explain shallow-N grammars. In order to briefly assess this strategy, in Section 4 we discuss results in Chinese-English, Arabic-English and Spanish-English translation tasks, after which we conclude.

2. Hierarchical Decoding with WFSTs

The first step of this translation system is based on a variant of the CYK algorithm closely related to CYK+ [6]. We keep backpointers and employ hypotheses recombination without discarding rules. The underlying model is a synchronous context-free grammar consisting of a set $\mathbf{R} = \{R^r\}$ of rules R^r : $N \rightarrow \langle\gamma^r,\alpha^r\rangle$ / p^r, with 'glue' rules, $S \rightarrow \langle X,X\rangle$ and $S \rightarrow \langle S\ X,S\ X\rangle$. If a rule has probability p^r, it is transformed to a cost c^r; here we use the tropical semiring, so $c^r = -\log p^r$. N denotes a priori any non-terminal (S,X,V, etcetera), $N \in \mathbf{N}$. $\mathbf{T}$ denotes the terminals (words), and the grammar builds parse forests based on strings $\gamma, \alpha \in \{\mathbf{N} \cup \mathbf{T}\}^+$. Each cell in the CYK grid is specified by a non-terminal symbol and position in the CYK grid: (N, x, y), which spans s_x^{x+y-1} on the source sentence.

In effect, the source language sentence is parsed using a context-free grammar with rules $N \rightarrow \gamma$. The generation of translations is a second step that follows parsing. For this second step, we describe a method to construct word lattices with all possible translations that can be produced by the hierarchical rules. Construction proceeds by traversing the CYK grid along the backpointers established in parsing. In each cell (N, x, y) in the CYK grid, we build a target language word lattice $\mathcal{L}(N, x, y)$. This lattice contains every translation of s_x^{x+y-1} from every derivation headed by N. These lattices also contain the translation scores on their arc weights.

The ultimate objective is the word lattice $\mathcal{L}(S, 1, J)$, which corresponds to all the analyses that cover the source sentence s_1^J. Once this is built, we can apply a target language model to $\mathcal{L}(S, 1, J)$ to obtain the final target language translation lattice [7].

2.1. Lattice Construction Over the CYK Grid

In each cell (N, x, y), the set of rule indices used by the parser is denoted $R(N, x, y)$, *i.e.* for $r \in R(N, x, y)$, $N \rightarrow \langle\gamma^r,\alpha^r\rangle$ was used in at least one derivation involving that cell.

For each rule $R^r, r \in R(N, x, y)$, we build a lattice $\mathcal{L}(N, x, y, r)$. This lattice is derived from the target side of

⋆‡The work in this thesis has been done while Gonzalo Iglesias was still working for the Department of Signal Theory and Communications, at the University of Vigo.

the rule α^r by concatenating lattices corresponding to the elements of $\alpha^r = \alpha^r_1 ... \alpha^r_{|\alpha^r|}$. If an α^r_i is a terminal, creating its lattice is straightforward. If α^r_i is a non-terminal, it refers to a cell (N', x', y') lower in the grid identified by the backpointer $BP(N, x, y, r, i)$; in this case, the lattice used is $\mathcal{L}(N', x', y')$. Taken together,

$$\mathcal{L}(N, x, y, r) = \bigotimes_{i=1..|\alpha^r|} \mathcal{L}(N, x, y, r, i) \tag{1}$$

$$\mathcal{L}(N, x, y, r, i) = \begin{cases} \mathcal{A}(\alpha_i) & \text{if } \alpha_i \in \mathbf{T} \\ \mathcal{L}(N', x', y') & \text{else} \end{cases} \tag{2}$$

where $\mathcal{A}(t)$, $t \in \mathbf{T}$ returns a single-arc acceptor that accepts only the symbol t. The lattice $\mathcal{L}(N, x, y)$ is then built as the union of lattices corresponding to the rules in $R(N, x, y)$:

$$\mathcal{L}(N, x, y) = \bigoplus_{r \in R(N,x,y)} \mathcal{L}(N, x, y, r) \otimes c^r \tag{3}$$

This slight abuse of notation indicates that the cost c^r is applied at the path level to each lattice $\mathcal{L}(N, x, y, r)$; the cost can be added to the exit states, for example. This could as well be done at Equation 1.

2.2. Avoiding Pruning in Search

Equation 2 leads to the recursive construction of lattices in upper-levels of the grid through the union and concatenation of lattices from lower levels. If Equations 1 and 3 are actually carried out over fully expanded word lattices, the memory required by the upper lattices will increase exponentially.

To avoid this, we use special arcs that serve as pointers to the low-level lattices. This effectively builds a skeleton of the desired lattice and delays the creation of the final word lattice until a single replacement operation is carried out in the top cell $(S, 1, J)$. To make this exact, we define a function $g(N, x, y)$ that returns a unique tag for each lattice in each cell, and use it to redefine Equation 2. With the backpointer $(N', x', y') = BP(N, x, y, r, i)$, these special arcs are introduced as:

$$\mathcal{L}(N, x, y, r, i) = \begin{cases} \mathcal{A}(\alpha_i) & \text{if } \alpha_i \in \mathbf{T} \\ \mathcal{A}(g(N', x', y')) & \text{else} \end{cases} \tag{4}$$

The resulting lattices $\mathcal{L}(N, x, y)$ are a mix of target language words and lattice pointers. However, each still represents the entire search space of all translation hypotheses covering the span.

At the upper-most cell, the lattice $\mathcal{L}(S, 1, J)$ contains pointers to lower-level lattices. A single FST replace operation [5] recursively substitutes all pointers by their lower-level lattices until no pointers are left, thus producing the complete target word lattice for the whole source sentence. The use of the lattice pointer arc was inspired by the 'lazy evaluation' techniques developed by Mohri et al. [8]. Its implementation uses the infrastructure provided by the OpenFST libraries for delayed composition, etc.

Importantly, operations on these cell lattices — such as lossless size reduction via determinization and minimization — can still be performed. Owing to the existence of multiple hierarchical rules which share the same low-level dependencies, these operations can greatly reduce the size of the skeleton lattice; Figure 1 shows the effect on the translation example. This strategy is a key aspect to avoid pruning in search as much as possible. As stated, size reductions can be significant. However, not all redundancy is removed, since duplicate paths may arise through the concatenation and union of sublattices with different spans.

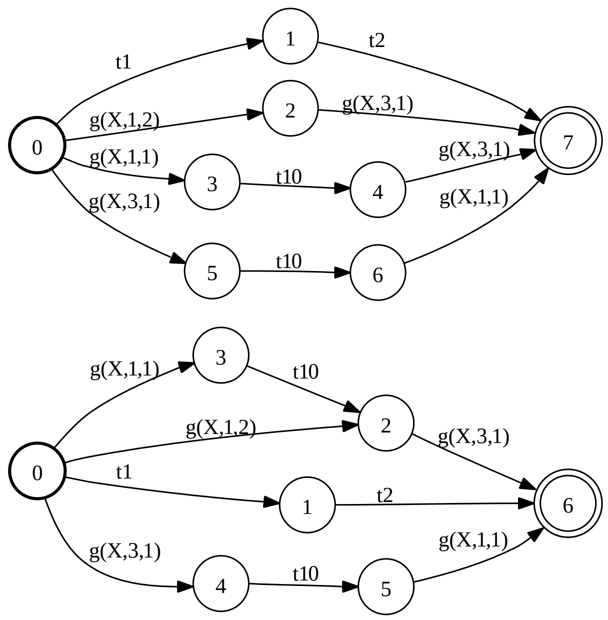


Figure 1: Delayed translation WFST with derivations from Figure 1 and Figure 2 before [t] and after minimization [b].

One interesting issue is where to use and where *not* to use pointer arcs. Several WFST operations are quite efficient due to the use of epsilon arcs. Unfortunately, combining carelessly these operations introduces an excessive number of epsilon arcs that very easily lead to intractable lattices. In many cases, removing epsilons is enough. But the expansion is a single operation that recursively traverses all the arcs substituting pointers to lower lattices by adding at least two epsilons per substitution. So, the issue is not only about making the lattice construction fast, but delivering a tractable skeleton for posterior steps. We decide which cell lattice will be replaced by a single arc depending on the non-terminal this cell is associated to. As a rule of thumb, the S cell lattices should never be replaced by pointer arcs, as they are used recursively many times for each translation hypothesis. A lattice construction doing so would return a minimal FST of two states binded by one single pointer arc, from which the complete search space lattice (possibly with millions of derivations) must be created, including at least twice as many epsilons as glue rules used within each derivation.

3. Shallow-N Grammars

Hierarchical grammars have shown good performance on translation tasks that demand lots of word reordering, such as Chinese-to-English. But for translation tasks between closer languages – e.g. Spanish-English – a search space defined by a hierarchical grammar leads to overgeneration, i.e. nonsensical translation hypotheses allowed by this powerful word reordering. Whereas hierarchical grammar allows any number of nestings through X rules up to a given word span threshold (typically 10 words), shallow-N grammars limit the size of the search space by defining a parameter N that controls directly the number of rule nestings allowed. More formally, a shallow-

N translation grammar can be defined as:

1. the usual non-terminal S
2. a set of non-terminals $\{X^0, \ldots, X^N\}$
3. two glue rules: $S \rightarrow \langle X^N, X^N \rangle$ and $S \rightarrow \langle S\ X^N, S\ X^N \rangle$
4. hierarchical translation rules for levels $n = 1, \ldots, N$:
R: $X^n \rightarrow \langle \gamma, \alpha, \sim \rangle$, $\gamma, \alpha \in \{\{X^{n-1}\} \cup \mathbf{T}\}^+$
with the requirement that α and γ contain at least one X^{n-1}
5. translation rules which generate lexical phrases:
R: $X^0 \rightarrow \langle \gamma, \alpha \rangle$, $\gamma, \alpha \in \mathbf{T}^+$

Table 1 illustrates the shallow grammars for $N = 1, 2, 3$. As is clear, with larger N the expressive power of the grammar grows closer to that of full Hiero.

grammar	rules included
S-1	S$\rightarrow\langle X^1, X^1 \rangle$ S$\rightarrow\langle$S X^1,S $X^1\rangle$ $X^0 \rightarrow \langle \gamma, \alpha \rangle$, $\gamma, \alpha \in \mathbf{T}^+$ $X^1 \rightarrow \langle \gamma, \alpha, \sim \rangle$, $\gamma, \alpha \in \{\{X^0\} \cup \mathbf{T}\}^+$
S-2	S$\rightarrow\langle X^2, X^2 \rangle$ S$\rightarrow\langle$S X^2,S $X^2\rangle$ $X^0 \rightarrow \langle \gamma, \alpha \rangle$, $\gamma, \alpha \in \mathbf{T}^+$ $X^1 \rightarrow \langle \gamma, \alpha, \sim \rangle$, $\gamma, \alpha \in \{\{X^0\} \cup \mathbf{T}\}^+$ $X^2 \rightarrow \langle \gamma, \alpha, \sim \rangle$, $\gamma, \alpha \in \{\{X^1\} \cup \mathbf{T}\}^+$
S-3	S$\rightarrow\langle X^3, X^3 \rangle$ S$\rightarrow\langle$S X^3,S $X^3\rangle$ $X^0 \rightarrow \langle \gamma, \alpha \rangle$, $\gamma, \alpha \in \mathbf{T}^+$ $X^1 \rightarrow \langle \gamma, \alpha, \sim \rangle$, $\gamma, \alpha \in \{\{X^0\} \cup \mathbf{T}\}^+$ $X^2 \rightarrow \langle \gamma, \alpha, \sim \rangle$, $\gamma, \alpha \in \{\{X^1\} \cup \mathbf{T}\}^+$ $X^3 \rightarrow \langle \gamma, \alpha, \sim \rangle$, $\gamma, \alpha \in \{\{X^2\} \cup \mathbf{T}\}^+$

Table 1: Rules contained in shallow-N grammars for $N = 1, 2, 3$.

Actually, shallow grammars are created by a trivial rewriting procedure of the full grammar. Consider the following example with a source sentence '$s_1\ s_2$' and a full grammar defined by these four rules:

R[1]: $S \rightarrow \langle X, X \rangle$
R[2]: $X \rightarrow \langle s_1\ s_2, t_2\ t_1 \rangle$
R[3]: $X \rightarrow \langle s_1\ X, X\ t_1 \rangle$
R[4]: $X \rightarrow \langle s_2, t_2 \rangle$

We can easily rewrite these rules according to a shallow-1 grammar:

R[1]: $S \rightarrow \langle X^1, X^1 \rangle$
R[2]: $X^0 \rightarrow \langle s_1\ s_2, t_2\ t_1 \rangle$
R[3]: $X^1 \rightarrow \langle s_1\ X^0, X^0\ t_1 \rangle$
R[4]: $X^0 \rightarrow \langle s_2, t_2 \rangle$
R[5]: $X^1 \rightarrow \langle X^0, X^0 \rangle$

One interesting feature comes from the topology of shallow grammars: as they use several non-terminals, it is possible to set different constraints on each of them, e.g. enforce a minimum source word span. This is useful to speed up translation systems with shallow grammars allowing more than one nested rule [9].

4. Results

We have assessed the validity of our combined strategy throughout several translations tasks, such as Chinese-to-English, Arabic-to-English, and Spanish-English, amongst others. In first place, we have shown that *HiFST*, even when pruning-in-search is required, never makes more search errors than a hierarchical cube-pruning decoder, overall resulting in better performance. This is specially notable for rescoring steps, after which an improvement of more than 1 BLEU point for both Arabic-to-English and Chinese-to-English is achieved [9]. Whereas with a complex task such as Chinese-to-English a *full* hierarchical grammar is needed to achieve the best performance, a nesting of 2 or even 1 is enough for other closer language pairs. All these findings are discussed in detail throughout several conference papers [10, 11, 12] and a journal paper [9]. *HiFST* is also the core of the CUED system, which ranked first in the Arabic-to-English NIST 2009 Constrained Data Track[1].

In this section we will focus on some relevant results of the Spanish-English translation tasks.

The training was performed using lower-cased data. Word alignments were generated using GIZA++ [13] over a stemmed version of the parallel text. After unioning the Viterbi alignments, the stems were replaced with their original words, and phrase-based rules of up to five source words in length were extracted [14]. Hierarchical rules with up to two non-contiguous non-terminals in the source side are then extracted applying the usual restrictions [4]. The Europarl language model is a Kneser-Ney [15] smoothed default cutoff 4-gram back-off language model estimated over the concatenation of the Europarl and News language model training data.

After translating with optimized feature weights, we carry out the two following rescoring steps to the output lattice:

- *Large-LM rescoring (5g).* We build sentence-specific zero-cutoff stupid-backoff [16] 5-gram language models.
- *Minimum Bayes Risk (MBR).* We rescore the first 1000-best hypotheses with MBR [17], or the lattice with Lattice MBR (LMBR) [18], taking the negative sentence level BLEU score as the loss function.

4.1. Experiments on the Shared task of WMT08

In this subsection we present experiments for Spanish-to-English on the shared task of the ACL 2008 Workshop on Statistical Machine Translation [12].

As we had already discovered for the Arabic-to-English task [10], we found that the shallow-1 grammar already had the same performance as hierarchical full system at much greater speed, as pruning during search is avoided entirely. Table 2 shows results for our shallow-1 model and subsequent rescoring steps. Gains from large language models are more modest than MBR, possibly due to the domain discrepancy between the EuroParl and the additional newswire data.

Scores are comparable to the top submissions in the $WMT08$ shared-task results [19].

	dev2006	*test2008*
HiFST(S-1)	33.6/7.85	33.8/7.90
+5g	33.7 /7.90	33.9/7.95
+5g+MBR	33.9 /7.90	34.2/7.96

Table 2: EuroParl Spanish-to-English translation results (lower-cased IBM BLEU / NIST) after MET and subsequent rescoring steps

[1] See http://www.itl.nist.gov/iad/mig/tests/mt/2009/ResultsRelease for full MT09 results.

Task	System	*nwtest08*	*nwtest09*	*nwtest10*
SP → EN	HiFST(S-1)	24.6	26.0	29.1
	+5g+LMBR	25.4	27.0	**30.5**
EN → SP	HiFST (S-1)	23.9	24.5	28.0
	+5g+LMBR	24.7	25.5	**29.1**

Table 3: Translation Results for the Spanish-English (SP-EN) language pair, shown in lowercase IBM BLEU. Bold results correspond to submitted systems.

4.2. Shared task of WMT10

We have participated in the ACL 2010 Workshop of Machine Translation [20] on several translation tasks [21]. Table 3 shows the excellent results for both Spanish-to-English and English-to-Spanish tasks. In both directions, using *HiFST* with shallow-1 grammars allows a search-error free decoding. In turn, this allows rescoring steps to increase the performance in more than 1 BLEU point for both directions.

5. Conclusions

In the context of hierarchical translation, this thesis has proposed a novel translation system that uses WFSTs within hierarchical decoding, capable of compact and efficient representations of the translation search space. A refinement to the hierarchical grammars, which we call shallow-N grammars, has also been introduced. This refinement allows a simple tuning to the word-reordering requirements of each particular translation task, thus avoiding overgeneration. Taken both strategies together, it is possible to build state-of-the-art translation systems between close languages – such as Spanish, English, French or even Arabic – without pruning in search, leading to faster decoding times, search-error free translation lattices and improved rescoring performance. This thesis is available for download at `http://www.eng.cam.ac.uk/~gi212/thesis.pdf`.

6. Acknowledgements

This work has been supported by Spanish Government research grant BES-2007-15956, project AVIVAVOZ (TEC2006-13694-C03-03) and project BUCEADOR (TEC2009-14094-C04-04). Also supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

7. References

[1] G. Iglesias, "Hierarchical phrase-based translation with weighted finite-state transducers," Ph.D. dissertation, University of Vigo, Spain, 2010.

[2] A. Zollmann, A. Venugopal, F. Och, and J. Ponte, "A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT," in *Proceedings of COLING*, 2008, pp. 1145–1152.

[3] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proceedings of ACL*, 2005, pp. 263–270.

[4] ——, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[5] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *Proceedings of CIAA*, 2007, pp. 11–23.

[6] J.-C. Chappelier and M. Rajman, "A generalized CYK algorithm for parsing stochastic CFG," in *Proceedings of TAPD*, 1998, pp. 133–137.

[7] C. Allauzen, M. Mohri, and B. Roark, "Generalized algorithms for constructing statistical language models," in *Proceedings of ACL*, 2003, pp. 557–564.

[8] M. Mohri, F. Pereira, and M. Riley, "The design principles of a weighted finite-state transducer library," *Theoretical Computer Science*, vol. 231, pp. 17–32, 2000.

[9] A. de Gispert, G. Iglesias, G. Blackwood, E. R. Banga, and W. Byrne, "Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars," *Computational Linguistics*, vol. 36, no. 3, pp. 505–533, 2010.

[10] G. Iglesias, A. de Gispert, E. R. Banga, and W. Byrne, "Rule filtering by pattern for efficient hierarchical translation," in *Proceedings of EACL*, 2009, pp. 380–388.

[11] ——, "Hierarchical phrase-based translation with weighted finite state transducers," in *Proceedings of NAACL-HLT*, 2009, pp. 433–441.

[12] ——, "The HiFST system for the europarl Spanish-to-English task," in *Proceedings of SEPLN*, 2009, pp. 207–214.

[13] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[14] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of NAACL-HLT*, 2003.

[15] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proceedings of ICASSP*, vol. 1, 1995, pp. 181–184.

[16] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large language models in machine translation," in *Proceedings of EMNLP-ACL*, 2007, pp. 858–867.

[17] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proceedings of NAACL-HLT*, 2004, pp. 169–176.

[18] R. Tromble, S. Kumar, F. J. Och, and W. Macherey, "Lattice Minimum Bayes-Risk decoding for statistical machine translation," in *Proceedings of EMNLP*, 2008, pp. 620–629.

[19] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "Further meta-evaluation of machine translation," in *Proceedings of WMT*, 2008, pp. 70–106.

[20] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," in *Proceedings of WMT*, 2010, pp. 17–53.

[21] J. Pino, G. Iglesias, A. de Gispert, G. Blackwood, J. Brunning, and W. Byrne, "The CUED HiFST System for the WMT10 Translation Shared Task," in *Proceedings of WMT*, 2010, pp. 155–160.

On-line Personalization and Adaptation to Disorders and Variations of Speech on Automatic Speech Recognition Systems

Oscar Saz

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Spain
oskarsaz@unizar.es

Abstract

This thesis deals with the research and development of speech technology-based systems for the requirements of users with different impairments and disabilities, with the final aim of improving their quality of life. Speech disorders are shown to be a major challenge in the work with these users. This work performs all the steps in the research in speech technologies: starting with the acquisition of an oral corpus from young impaired speakers, the analysis of the acoustic and lexical variations in the disordered speech and the characterization of speaker dependent Automatic Speech Recognition (ASR) systems adapted to the acoustic and lexical variants introduced by these speakers. Furthermore, automated methods for detection and correction of lexical mispronunciations are also evaluated. The results of the experiments show the on-going possibility for developing a fully personalized ASR system for handicapped users that learns the speaker's speech characteristics on-line: while the user interacts with the recognition system. The development of speech therapy tools based on the knowledge gained is another outcome of the present thesis, where the development of "Comunica" aims to improve the possibilities for semi-automated speech therapy in Spanish.

Index Terms: speech disorders, speaker personalization, language learning

1. Introduction

Communication disorders are a heavy limitation for those who suffer them, and the impossibility to communicate with others impedes the social inclusion and development of impaired people. Difficulties in the access to education, impossibility in the access to work or social exclusion are some of the consequences of communicative impairments.

New technologies can facilitate communication and knowledge as they make access to information universal, immediate and ubiquitous. Unfortunately, current interfaces based on peripherals like mouse or keyboard are not accessible to disabled users as they require good levels of motor control and cognitive capabilities. Research on new forms on human computer interfaces towards more natural, adaptable and accessible interaction is currently being done in many facets like eye gaze tracking, movement tracking, brain computer interface or speech, which is the subject of this thesis.

1.1. Speech Technology for the Handicapped

Different systems based on Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) synthesis have already been researched for their use by the handicapped community. The STARDUST project aimed to provide oral commands and control of a home environment [1, 2]; the Vocal Joystick was designed to provide accessibility to computers by heavily impaired users [3, 4]; and, finally, the VIVOCA project created communicative aids able to recreate the speech from a disordered user [5, 6].

Education is another area of work where speech technologies can help the handicapped community; Computer-Aided Speech and Language Therapy (CASLT) tools, a subset within the broader domain of Computer-Aided Language Learning (CALL) tools [7], bring speech therapy to individuals with communication difficulties [8, 9].

One of the main difficulties for the research in these lines of work is the lack of resources (corpora, databases) to characterize and study the main features of disordered speech and to develop and evaluate automated speech systems. Corpora like the Whitaker [10] and Nemours [11] databases were collected at the early stages of the interest in disordered speech. Currently, the Universal Access Database [12] is the largest database covering the domain of dysarthric and disordered speech.

1.2. Motivation and Objectives

This thesis has grown up thanks to the interest of different people and institutions in the creation of software and systems based in speech technologies for education, inclusion and assistance purposes. These institutions like the Public School for Special Education (CPEE) "Alborada", CADIS-Huesca, ASPACE-Huesca or the Vienna International School (VIS) are the great motivators of this thesis.

The objectives of this thesis cover two different aspects in the technological work. From a scientific point of view, it aims to provide the community with a fully functional corpus of speech disorders in Spanish, to learn and study different adaptation techniques to these disorders in ASR systems, and to develop techniques for the assessment of speech proficiency. From a practical point of view, it aims to develop devices for control and interaction based on speech and voice which are accessible for handicapped users and CALL tools for students with different needs in learning communication through language.

2. Methodology and Corpus

The methodology of work was as presented in Figure 1. The corpus collected with different impaired voices served as the foundations of the work. Posteriorly, a main line of work appeared towards a better knowledge of the properties of speech

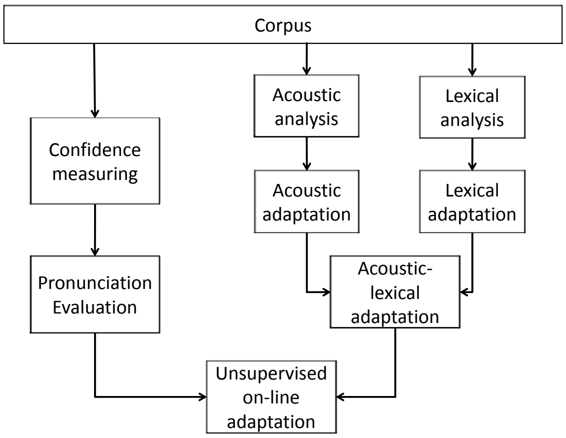


Figure 1: *Methodology of work.*

impairments through different analyses. These led to proposals for speaker adaptation in the acoustic and lexical models of the ASR, which were merged in acoustic-lexical speaker adaptation. A parallel line studying how to evaluate correctly the pronunciation of the impaired speakers and transform this in confidence measures was finally merged with the speaker adaptation into a proposal for unsupervised on-line personalized system.

2.1. The "Alborada-I3A" corpus

The "Alborada-I3A" corpus was recorded with the intention of filling the gap in resources and databases of disordered speech in Spanish [13]. It was recorded in collaboration with different educative institutions and is fully available for research.

The core of the corpus were 14 young speakers suffering different physical and cognitive handicaps and also suffering very different speech and language disorders. These speakers were 7 boys and 7 girls from 11 to 21 years old. A set of 4 isolated word sessions were recorded from each speaker, with a vocabulary of 57 words per session, for a total of 3,192 isolated word utterances. The set of words used, the Induced Phonological Register (RFI) [14], is well known among speech therapists in Spanish and contains a all phonemes of Spanish language in different positions and contexts. Furthermore, 232 young unimpaired speakers in the same age range than the impaired speakers (10-18 years old) were later recorded to model and characterize the speech in children and young adults with a proper correct speech.

In order to characterize phonologically the production of errors in the impaired speakers, a labeling was carried out by a set of experts to determine whether each phoneme in the pronunciation of the speakers was correctly pronounced, mistaken or substituted, or deleted by the speaker. The results showed a big relevance of the disorders, as around 10% of the phonemes were substituted and 7% were deleted.

The baseline ASR experiments with this corpus pointed out the big influence of the disorders on the performance of the system. While the unimpaired speakers were in a 4% of Word Error Rate (WER) on adult speech models, the impaired peers reached 37% WER. Task dependent models, retrained on the 232 unimpaired speakers allowed for a reduction of the WER to 28%, which was isolated as been due to acoustic and lexical disorders.

3. Analyses of Disordered Speech

Before starting any experimentation with the corpus, several studies were carried out to understand how the disorders affected all the facets of speech production. These studies were expected to be relevant on how to face all the problematic regarding recognition and evaluation of disordered speech.

3.1. Acoustic Analysis

A study on vowel production by the impaired speakers was made to find any differences between unimpaired and impaired speakers. The features which were under study were: First two formant frequencies (F_1 and F_2), fundamental frequency value (F_0), intensity value and duration. Speech processing methods like Linear Predictive Coding (LPC) and autocorrelation were used to calculate these features.

A degradation of the quality of vowel production was measured in some of these features, showing the inability of the impaired speakers for a precise control of articulation. This degradation affected in the following ways: Reduction of distance between vowels in the formant map, especially between /a/, /e/ and /o/; loss of distinction in the production of intensity between stressed and unstressed vowels; and a higher variability in vowels duration [15].

3.2. Lexical Analysis

The patterns of production of mistakes in the speakers were studied from a phonological point of view. A first approach did not show an influence of the phonological properties of phonemes (point and manner of articulation). On the contrary, the context and position of the phoneme in the syllable were the most prominent factors in mispronunciations. The main findings of this analysis were the significant reductions of vowels in diphthongs, consonants in coda position and consonant clusters.

The comparison of the pronunciation patterns of this impaired speakers with the pattern in young children in learning stages (3-6 years old) showed how young impaired speakers presented similar errors to small children. This pointed out that the possible origin of this language disorders might be a delay in language acquisition due to the cognitive delays of these speakers.

4. Techniques for Personalization

Personalization is a key point when developing speech interfaces for the handicapped. The influence of the specific impairment of each user is so unique that they require that the system is fully adaptive to the user instead of the user adapting to the system.

4.1. Speaker Adaptation

The studies carried out in the effects of speech disorders have shown the existence of acoustic distortion in the speakers' voice as well as phonological and lexical disorders leading to mispronunciations at this level [16]. Adaptation was proposed at three levels:

Acoustic adaptation for these speakers can provide a better matching of acoustic models to their speech properties. These properties not only include inter speaker variability as for the rest of the population, but the modeling of their speech disorders and how they modify the acoustic production of speech. Lexical adaptation provides a correct modeling of the pronunciations that these speakers are uttering, characterized by a large number of substitutions and deletions at the phoneme level. While lexical adaptation is usually not required for unimpaired speakers, these speakers might find relevant benefits with it. Acoustic-lexical adaptation merges the two previous approaches, but it is important to understand the correlations between them as they provide different ways to model similar effects of speech.

The proposal for acoustic adaptation was Maximum A Posteriori [17] and Maximum Likelihood Linear Regression (MLLR) [18], while the proposal for lexical adaptation was a data-driven approach, which learned the speaker's transcriptions through Acoustic Phonetic Decoding (APD). After several experiments, the joint use of acoustic and lexical adaptation produced a relative improvement in the WER of 46%.

4.2. Pronunciation Evaluation

Pronunciation verification is a relevant issue in the work with speech disorders, it can serve to correct the speaker's pronunciation or to discard incorrect pronunciations in speech systems. Among several other proposals [19, 20], the work focused on normalization of the phoneme scores to different combinations of competing phonemes in the same way that the test normalization (or t-norm) method for speaker verification tasks.

The results showed that a correct selection of the competing phonemes [21] could achieve better results in terms of Equal Error Rate (EER) in the pronunciation verification tasks than known techniques like the Goodness of Pronunciation (GOP) [22]. GOP can also be seen as normalization technique where all the competing phonemes are considered. Better result of EER was lowered to 16%.

4.3. Proposal of On-line Unsupervised Personalized System

Finally, a proposal for on-line unsupervised personalization was provided in Figure 2. In this system, the user is fully unaware of the procedure in which the personalization is performed constantly in the underlying loop. During the use of the ASR system, the confidence measuring algorithm discard those transcriptions which might be inaccurate (either due to recognition mistakes or due to pronunciation mispronunciations) and stores in a buffer all the signals which are considered useful for a posterior adaptation. Once the buffer has sufficient amount of reliable data, speaker adaptation is performed according to the transcriptions obtained by the ASR system validated by the confidence measure. The new models (acoustic and/or lexical) are inserted in the ASR system to improve the recognition accuracy of the user's speech. This process can be repeated to keep providing further adaptation to the speaker.

After some preliminary experiments in this proposal, it was seen how an iterative procedure of adaptation like the one depicted in Figure 2 could outperform in terms of improvement of the WER a similar proposal using the same amount of adaptation data in one single stage. Several concerns have to be taken into account in this proposal, especially the influence of the initial ASR stage to obtain the transcriptions and the accuracy of the confidence measure algorithm.

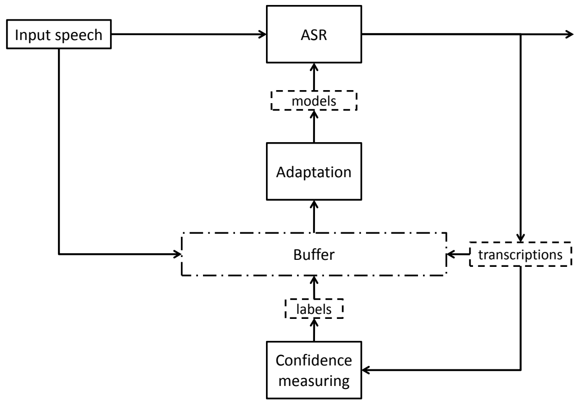


Figure 2: *On-line personalization.*

5. Speech Based Applications

The experimental work in the thesis has been accompanied by work on the development and deployment of speech-based tools to improve the quality of life of the handicapped. "VozClick", developed for ASPACE-Huesca, aims to substitute physical switches for severely handicapped people in their access to computer applications [23]. It transforms a pulse of vocal emission by the user into a event in the computer substituting peripherals like mouse or keyboard.

It has been in the development of CALL tools where greater results have been achieved during this work where "Comunica" has been a framework on the research and development of CASLT tools in Spanish [24]. "Comunica" consists of "PreLingua" for the training of phonatory skills in small handicapped children [25], "Vocaliza" for the training of articulation skills [26] and "Cuéntame" for linguistic skills. Their open and free distribution through the Internet[1] has shown the great interest of speech therapists in Spain and Latin America for this kind of technical aids and their appreciation for the tools in "Comunica"

The results in pronunciation verification achieved in the thesis were introduced in "Vocaliza" to provide a tool for the training of Spanish as a Second Language (L2) to children. The results of an experience carried out in this field showed the usefulness of computer-aided tools with a dedicated interface and a correct use of speech technologies [27, 28, 29].

6. Conclusions

This thesis has supposed a relevant effort in all the objectives proposed at the beginning of the work. The corpus has showed to be useful and it has put the interest of the community on this specific task. Different personalization techniques have been evaluated and the performance of them has been framed. Confidence measuring and pronunciation assessment has shown significant improvements and solid results.

The thesis has also discussed several subjects regarding the origins of the speech disorders, their affection in the speech production of the users and their effect on the performance of automated speech recognition and assessment systems. More precisely, disorders at the lexical level have had a special treatment in their analysis and evaluation in the thesis, compared to previous works focusing mostly on the acoustic side of speech.

Finally, further work which has arisen from the thesis has to be oriented towards the study on personalization techniques which can work in cases of unsupervised data and data sparsity. Techniques which take into account the mutual information of different sources might be useful for further improvement in confidence measuring tasks. Finally, all this work still has open road to be introduced in real systems to provide accessibility and inclusion for handicapped people.

[1] http://www.vocaliza.es

7. Acknowledgments

This thesis was supported by the FPU program of the Spanish Ministry of Education. Part of the work was funded under national projects TIN 2005-08660-C04-01 and TIN2008-06856-C05-04. The thesis was carried out under the supervision of Prof. Eduardo Lleida at the Department of Electrical Engineering and Communications of the University of Zaragoza (Spain). Work was shared with Prof. Phil Green and Dr. Stuart Cunningham at the Department of Computer Science at the University of Sheffield (United Kingdom) in June-September 2006 and with Prof. Richard Rose at the Department of Electrical and Computer Engineering at McGill University (Montreal, Canada) in June-September 2007.

8. References

[1] P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, and M. Parker, "Automatic Speech Recognition with sparse training data for dysarthric speakers," in *Proceedings of the 8th Eurospeech*, Geneva, Switzerland, September 2003, pp. 1189–1192.

[2] M. Hawley, P. Enderby, P. Green, S. Brownsell, A. Hatzis, M. Parker, J. Carmichael, S. Cunningham, P. O'Neill, and R. Palmer, "STARDUST Speech Training And Recognition for Dysarthric Users of aSsistive Technology," in *Proceedings of the 7th Conference of the Association for the Advancement of Assistive Technology in Europe, AAATE*, Dublin, Ireland, August 2003.

[3] J. Bilmes, J. Malkin, X. Li, S. Harada, K. Kilanski, K. Kirchhoff, R. Wright, A. Subramanya, J. Landay, P. Dowden, and H. Chizeck, "The vocal joystick," in *Proceedings of the 2006 ICASSP*, Toulouse, France, May 2006, pp. 625–628.

[4] S. Harada, J. Landay, J. Malkin, X. Li, and J. Bilmes, "The Vocal Joystick: evaluation of voice-based cursor control techniques for assistive technology," *Disability and Rehabilitation: Assistive Technology*, vol. 3, no. 1, pp. 22–34, 2008.

[5] S. Creer, S.-P. Cunningham, P.-D. Green, and K. Fatema, "Personalizing synthetic voices for people with progressive speech disorders: Judging voice similarity," in *Proceedings of the 11th Eurospeech*, Brighton, UK, September 2009, pp. 1427–1430.

[6] S. Creer, P. Green, S. Cunningham, and J. Yamaghisi, "Building personalized synthesized voices for individuals with dysarthria using the HTK toolkit," in *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, J.-W. Mullenix and S.-E. Stern, Eds. Hershey (PA), USA: IGI Publishing, 2010.

[7] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[8] K. Vicsi, P. Roach, A. Oester, Z. Kacic, P. Barczikay, and I. Sinka, "SPECO: A multimedia multilingual teaching and training system for speech handicapped children," in *Proceedings of the 6th Eurospeech*, Budapest, Hungary, September 1999, pp. 859–862.

[9] A.-M. Oester, D. House, A. Protopapas, and A. Hatzis, "Presentation of a new EU project for speech therapy: OLP (Ortho-Logo-Paedia)," in *Proceedings of the XV Swedish Phonetics Conference (Fonetik 2002)*, Stockholm, Sweden, May 2002, pp. 45–48.

[10] J.-R. Deller, M.-S. Liu, L.-J. Ferrier, and P. Robichaud, "The whitaker database of dysarthric (cerebral palsy) speech," *Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3516–3518, 1993.

[11] X. Menéndez-Pidal, J.-B. Polikoff, S.-M. Peters, J. Lorenzo, and H.-T. Bunnell, "The Nemours database of dysarthric speech," in *Proceedings of the 4th ICSLP*, Philadelphia (PA), USA, October 1996, pp. 1962–1965.

[12] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. France, "Dysarthric speech database for universal access research," in *Proceedings of the 10th ICSLP*, Brisbane, Australia, September 2008, pp. 1741–1744.

[13] O. Saz, W.-R. Rodríguez, E. Lleida, and C. Vaquero, "A novel corpus of children's impaired speech," in *Proceedings of the 2008 Workshop on Children, Computer and Interaction*, Chania, Greece, October 2008.

[14] M. Monfort and A. Juárez-Sánchez, *Registro Fonológico Inducido (Tarjetas Gráficas)*. Madrid, Spain: Ed. Cepe, 1989.

[15] O. Saz, J. Simón, W.-R. Rodríguez, E. Lleida, and C. Vaquero, "Analysis of acoustic features in speakers with cognitive disorders and speech impairments," *EURASIP Journal on Advances in Signal Processing*, vol. Special Issue on Analysis and Signal Processing of Oesophageal and Pathological Voices, 2009.

[16] O. Saz, E. Lleida, and A. Miguel, "Combination of acoustic and lexical speaker adaptation for disordered speech recognition," in *Proceedings of the 11th Eurospeech*, Brighton, United Kingdom, September 2009, pp. 544–547.

[17] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[18] C.-J. Legetter and P.-C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of the parameters of continous density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[19] S.-C. Yin, R. Rose, O. Saz, and E. Lleida, "Verifying pronunciation accuracy from speakers with neuromuscular disorders," in *Proceedings of the 10th ICSLP*, Brisbane, Australia, September 2008, pp. 2218–2221.

[20] ——, "A study of pronunciation verification in a speech therapy application," in *Proceedings of the 2009 ICASSP*, Taipei, Taiwan, April 2009, pp. 4609–4612.

[21] O. Saz, E. Lleida, and W.-R. Rodríguez, "Avoiding speaker variability in pronunciation verification of children disordered speech," in *Proceedings of the 2009 Workshop on Children, Computer and Interaction*, Cambridge (MA), USA, November 2009.

[22] S.-M. Witt and S.-J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[23] O. Saz, E. Lleida, L. Abarca, and S. Mejuto, "Mouseclick: Acceso al ordenador a través de la voz," in *Proceedings of the IV Jornadas Iberoamericanas de Tecnologías de Apoyo a Discapacidad*, Madrid, Spain, October 2009.

[24] O. Saz, S.-C. Yin, E. Lleida, R. Rose, W.-R. Rodríguez, and C. Vaquero, "Tools and technologies for computer-aided speech and language therapy," *Speech Communication*, vol. 51, no. 10, pp. 948–967, 2009.

[25] W.-R. Rodríguez, C. Vaquero, O. Saz, and E. Lleida, "Speech technology applied to children with speech disorders," in *Proceedings of the 4th Kuala Lumpur International Conference on Biomedical Engineering*, Kuala Lumpur, Malaysia, June 2008, pp. 247–250.

[26] C. Vaquero, O. Saz, E. Lleida, and W.-R. Rodríguez, "E-inclusion technologies for the speech handicapped," in *Proceedings of the 2008 ICASSP*, Las Vegas (NV), USA, April 2008, pp. 4509–4512.

[27] O. Saz, V. Rodríguez, E. Lleida, W.-R. Rodríguez, and C. Vaquero, "An experience with a Spanish Second Language learning tool in a multilingual environment," in *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE)*, Wroxall Abbey Estates, United Kingdom, 2009.

[28] O. Saz, E. Lleida, V. Rodríguez, W.-R. Rodríguez, and C. Vaquero, "The use of synthetic speech in language learning tools: Review and a case study," in *Computer Synthesize Speech Technologies: Tools for Aiding Impairment*, J.-W. Mullenix and D.-E. Stern, Eds. Hershey (PA), USA: IGI Global Publishing, 2010.

[29] O. Saz, V. Rodríguez, E. Lleida, W.-R. Rodríguez, and C. Vaquero, "The use of multimodal tools for pronunciation training in second language learning of preadolescents," in *Language Teaching: Techniques, Developments and Effectiveness*, F. Columbus, Ed. Hauppauge (NY), USA: Nova Science Publishers, 2010.

New reordering and modeling strategies for Statistical Machine Translation

Marta R. Costa-jussà[1]*; Supervisor: José A. R. Fonollosa*[2]

[1] Barcelona Media Innovation Center
marta.ruiz@barcelonamedia.org
[2] Universitat Politècnica de Catalunya
adrian@gps.tsc.upc.edu

Abstract

Nowadays, translation may be the bottleneck of the pretended information globalisation. While surfing the Internet, for instance, sometimes we come across languages and characters we do not understand. Statistical machine translation (SMT) constitutes a research sub-area of machine translation (MT) that has recently gained much popularity. In fact, this technology has experienced real growth motivated by the development of computer resources needed to implement translation algorithms based on statistical methods. This thesis focuses on the SMT framework and primarily on the definition and experimentation of novel algorithms for building a correct structural reordering for translated words. Moreover, challenging techniques regarding language modeling and system combination are successfully applied to state-of-the-art SMT systems. This thesis should shed some light on the SMT approach and on the word ordering challenges and should be specially useful to natural language processing researchers having non or some expertise in machine translation.

Index Terms: Statistical machine translation, Word reordering, Language modeling, System combination, Rescoring, Word graphs

1. Introduction

This thesis focuses on the statistical machine translation (SMT) framework and primarily on the definition and experimentation of novel algorithms for building a correct structural reordering for translated words. Moreover, challenging techniques regarding language modeling and system combination are successfully applied to state-of-the-art SMT systems.

To begin, a thorough study of the SMT state-of-the-art is performed. Ngram- and phrase-based SMT feature functions are described. The former, which has been developed in our research group, is used as a baseline system and the latter, given its popularity, is used to deepen the new techniques during experimentation.

This thesis proposes the introduction of novel statistical reordering techniques in an SMT system. The first approach is based on an algorithm that detects, learns and infers pairs of words in the source language that swap in the target language providing accurate local reorderings. The second approach consists of generating weighted reordering hypotheses using the same powerful techniques of SMT systems in order to undo the source language structure and to make it more similar to the target language structure. Therefore, the translation challenge is divided into two steps: predicting the order of the words in the target language and substituting these words in the target language. In order to infer new reorderings that were not learnt during training, the NbR system uses word classes instead of words themselves. In order to correctly integrate the NbR and SMT systems, both are concatenated, by using a word graph. This approach is an elegant and efficient reordering approach that is capable of achieving significantly improved translation in the target language.

Then, the introduction of continuous space language models is reported and analyzed in an Ngram-based system that uses translation and target language models. The continuous space language modeling technique is based on projecting word indices onto a continuous space. The resulting probability functions are smooth functions of the word representation. Events are better estimated than in standard smoothing methods, which is shown by the significant reduction in perplexity. This better probability estimation allows for an improvement in translation quality.

Moreover, this thesis performs a two-system combination considering the phrase and Ngram-based systems. Multiple outputs of both systems with their corresponding score are concatenated, and for each system translation the score given by the opposite system is computed. The final translation is properly chosen by simultaneously considering the scores given by both systems.

This paper reviews the main thesis ideas and it is organized as follows. Next section describes the nature of the machine translation problem. Section 3 presents the relevant theory from a qualitative point of view. Section 4 explains the thesis goals. Section 5 describes the novel methods proposed in the thesis. Finally, section 6 reports the project framework of this thesis and section 7, the achievements.

2. Nature of the problem

This PhD thesis focuses on the framework of statistical machine translation (SMT), which is a specific approach to machine translation (MT). The main goal of MT is to be able to translate from a source language s to a target language t. MT is a difficult task, mainly because natural languages are highly complex. Many words have more than one meaning and sentences may have various readings. Certain grammatical relations in one language might not exist in another language. Moreover, there are non-linguistic factors such as the problem that performing a translation might require world knowledge. Additional challenges arise when dealing with spoken language translation like confronting non-grammatical texts.

In order to face the MT challenge, many dependencies have to be taken into account. Often, these dependencies are weak and vague, which makes it rarely possible to describe simple and relevant rules that hold without exception for different language pairs. SMT treats MT as a decision problem, where we have to decide upon several target sentences, given a source sentence and among all possible target sentences, we will choose the sentence with the highest probability according to a statistically-learned model. SMT technology has received increasing interest leading to improved algorithms and it has been justified by various successful comparative evaluations since its revival by the work of the famous IBM research group more than fifteen years ago. It has proved to be a competitive approach, which shows greater robustness than other methods for the translation of spontaneous speech. Particularly, SMT translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. However, translations generated by SMT systems still have several significant challenges to pursue, like word reordering or word correspondences as we will see in the next sections.

3. Relevant theory

The SMT framework formulates the problem of translating a sentence from a source language s into a target language t as the maximization problem of the conditional probability $p(t|s)$. During the translation process, a statistical score based on the probabilities of the feature functions is assigned to each translation candidate, and the one with the highest combination score is selected as translation output. However, the SMT system might not be able to correctly score translations due to statistical models limitations.

Reordering is understood as the word order redistribution of the translated words as shown in Figure 1. In initial SMT systems, this different order is only modeled within the limits of translation units.

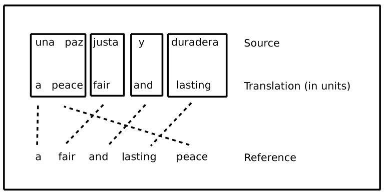


Figure 1: *Source, translation (in units) and reference example. Translation and reference differ in word order.*

Relying only in the reordering provided by translation units may not be good enough in most language pairs, which might require longer reorderings (as shown in Figure 1). Therefore, additional techniques may be deployed to face the reordering challenge. That is why many extended approachespropose to face statistical machine translation as a concatenation of two sub-tasks: predicting the collection of words in a translation and deciding the order of the predicted words. Introducing more complex feature functions that facilitate scoring the translation may not be easily introduced during decoding; for example, those feature functions that use the entire sentence to produce a score. One straightforward solution to this problem is a two-step decoding approach: in the first step, the decoder is run in n-best mode to produce n-best lists with N hypotheses per sentence. In the second step, the n-best lists are rescored with additional models. This technique reevaluates the n-best translation hypotheses of an MT system by introducing additional feature functions that should add information not included during decoding. Figure 2 shows a standard rescoring framework.

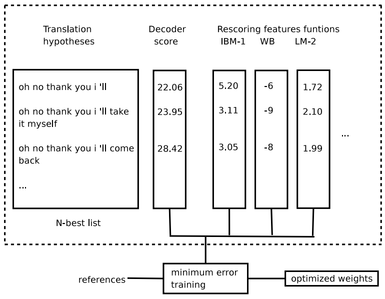


Figure 2: *Standard rescoring framework.*

Multiples translations can be computed by one MT system or by different MT systems. We may assume that different MT systems make different errors due to using different models, generation strategies or tweaks. An investigated technique, inherited from Automatic Speech Recognition, is the so-called system combination that is based on combining the outputs of multiples MT systems.

Given that most experiments of this PhD were done with phrase- and Ngram-based SMT systems, we combined the outputs of both systems using statistical criteria and additional rescoring features.

4. Thesis' goals (hypothesis to be tested)

This PhD is focused on achieving one main objective. Additionally, it addressed complementary challenges without deviating from the original scope.

- **To introduce novel reordering statistical techniques which are able to produce the translation in the correct word order**. Source and target languages may have different order structures. Languages may differ in the basic word order of verbs (V), subjects (S), and objects (O) in declarative clauses. For example, Spanish and English are both SVO languages. German, by contrast, is an SOV language while Classical Arabic and Urdu are VSO languages. Furthermore, there may be other structural differences in word orders between constituents; for instance, modifiers for nouns or verbs may be located in different places. Actually, an important deficit of current SMT systems is the difficult introduction of reordering capabilities. Incorporating them in the search process implies a high computational cost. However, reordering plays an important role, especially in some language pairs, such as Arabic or Chinese to English. The main issue of this thesis consists in developing an efficient reordering technique that can solve statistically the difference in word order of any language pair.

Other complementary objectives are:

- **To propose rescoring and system combination strategies which captures the best quality translations**. Current translation algorithms segment the given source sentence into units and then translate each unit. Therefore, it can become extremely complex to introduce feature functions that deal with information of the entire translated sentence. In those cases, translation may be performed in two steps. In the first step, we compute an N-best list. In the second step, the N-best list is reranked by applying additional features functions. Furthermore, the reranking of hypotheses allows for easier system combination. Different SMT systems approaches lead to different translations. In order to merge N-best lists which have been provided by different systems, we can use feature functions which decide which is the best translation. We contemplate developing and/or introducing feature functions to discriminate translations.

- **To gain efficiency and accuracy in the translation unit vocabulary**. By complementing the reordering objective, the extraction process and reordering techniques must be combined, either at the word or unit levels. Hence, the way the SMT system learns bilingual units plays an important role in translation quality. Here, the main idea is to further study the extraction of bilingual units from parallel corpora taking practical aspects, such as the translation vocabulary sparseness, into account. For some applications with limited memory space (PDAs, mobiles), the number of bilingual units stored in the device should be limited without affecting the quality of translation. In addition, the fewer bilingual units allows for faster translation time. Hence, it is important to look for the best extraction algorithm linking efficiency and quality.
- **To build and maintain a state-of-the-art phrase-based system to compare its performance with the Ngram-based system.** We improved the SMT translation by using the Ngram-based system. To have impact and be relevant to the community, our improvements were demonstrated, when possible, in an in-house phrase-based system that is the most widely used system in SMT.

5. New methods and analysis

This section briefly describes the methods that were proposed to achieve the PhD objectives.

Reordering

This thesis proposes two novel reordering techniques [1]. The first and less complex is briefly described as follows. Given a word alignment, we identify those pairs of consecutive source blocks (sequences of words) whose translation is swapped, i.e. those blocks which, if swapped, generate a correct monotonic translation. Afterwards, we classify these pairs into groups, following recursively a co-occurrence block criterion, in order to infer reorderings. Inside the same group, we allow new internal combination in order to generalize the reorder to unseen pairs of blocks. Then, we identify the pairs of blocks in the source corpora (both training and test) which belong to the same group. We swap them and we use the modified source training corpora to realign and to build the final translation system. We have evaluated this first reordering approach both in alignment and translation quality. In addition, we have used two state-of-the-art SMT systems: a Phrased-based and an Ngram-based. Experiments on the EPPS task show improvements almost over 1 point in BLEU (the standard MT evaluation metric) [2].

The second is the so-called Ngram-based Reordering (NbR) approach which uses the powerful techniques of SMT systems to generate a weighted reordering graph (see the schema in Figure 3). Thus, statistical criteria reordering constraints are supplied to an SMT system, and this allows an extension to the SMT decoding search. NbR allows for a reduction of the vocabulary sparseness of the SMT system

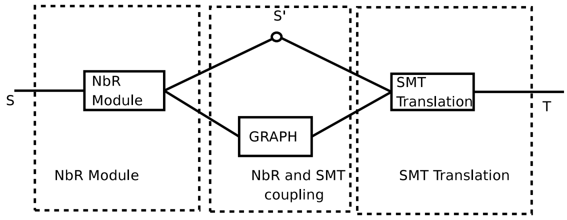


Figure 3: Schema of the NbR and SMT coupling.

during the training phase. The fact of using classes to train the reordering hypothesis (instead of words themselves) allows to generalize in the test phase. Therefore, the NbR technique is able to generate reordering hypotheses of sequences of words which were not seen during training. Additionally, the NbR technique provides a smoothed context-based weight to each reordering hypothesis by taking advantage of the highly developed language model techniques. Although introducing reordering abilities increases the system computational cost, experiments show that using the NbR technique guides the final translation decoding in an efficient manner. Reordering with the NbR technique highly outperforms our monotonic baseline system and a non-monotonic baseline system with a standard distance-based reordering. Improvement in translation performance has been demonstrated with the EPPS task (Spanish and German to English) and the BTEC task (Arabic to English), achieving improvements of 4 point BLEU [3].

Rescoring

The rescoring techniques proposed in this thesis use the continuous space LM which performs probability estimation in a continuous space. Since the resulting probability functions are smooth functions of the word representation, better generalization to unknown n-grams can be expected.

The continuous space LM is introduced as a target language model to rescore the n-best lists of a phrase- and Ngram-based statistical machine translation system. We have studied 4-gram language models because we had limited data. With more data, it would be easy to train a continuous space LM with much longer contexts, since the complexity of our approach increases only slightly with the size of the context. Results are provided on the BTEC tasks of the 2006 IWSLT evaluation for the translation direction Chinese, Arabic, Japanese and Italian to English. These tasks provide a very limited amount of resources in comparison to other tasks. Therefore, new techniques must be employed to take the best advantage of limited resources. The results show significant improvement for four different languages pairs and for both systems. The new approach achieves good improvements on the test data; the BLEU score increases by up to 1.9 points.

The continuous space language modeling was successfully extended to smoothing the bilingual language model of an Ngram-based system. The continuous space language model is trained on a bilingual sequence of tuples and it is introduced in the Ngram-based system rescoring. Our method is distinguished by two characteristics: better estimation of the numerous unseen n-grams; and a discriminative estimation of the tuple probabilities. Results are provided on the BTEC task of the 2006 IWSLT evaluation for the translation direction Italian to English. We have chosen the Italian to English task because it is challenging to improve the already good quality of the translation task (over 40 BLEU). Using the neural model for the translation and target language model, an improvement of 1.5 BLEU points on the test data was observed. The described smoothing method was explicitly developed to tackle the data sparseness problem in tasks like the BTEC corpus. Recently, continuous space language modeling applied on the target model (of a phrase-based system) has shown significant improvements when large amounts of data are available (see LIUM site participation in the 2008 NIST evaluation).

System combination
We propose a straightforward system combination method using several well-known feature functions for rescoring the 1-best output of the phrase- and Ngram-based SMT systems, using several n-gram language models, a word bonus and the IBM Model 1 for the whole sentence. The combination seems to obtain clear improvements in BLEU score. We report a structural comparison between the phrase- and Ngram-based system. On the one hand, the Ngram-based system outperforms the phrase-based in terms of search time efficiency by avoiding the overpopulation problem presented in the phrase-based approach. On the other hand, the phrase-based system shows a better performance when decoding under a highly constrained search. We carry out a detailed error analysis in order to better determine the differences in performance of both systems. The Ngram based system produces more accurate translations, but also a larger amount of extra (incorrect) words when compared to the phrase-based translation system. We present another system combination method which consists of concatenating a list of the respective system outputs and rescoring them using the opposite system as a feature function, i.e. the Ngram-based system is used for the phrase-based system and vice-versa. For both systems, including the probability given by the opposite system as a rescoring feature function leads to an improvement of BLEU score [4].

State-of-the-art SMT system
All methods presented were studied to gain efficiency and accuracy in translation and were contrasted with best state of the art systems in multiple MT International Evaluation Campaigns. The work in this events has to be understood in most cases as a team work and it complements the PhD scopes. We particpated in: IWSLT (2005-2008), WMT (2005-2008), NIST (2006 and 2008) and TC-STAR (2005-2007), which means a total of 11 evaluations, presenting the winning system in more than 5 tasks.

6. Project framework

The research presented in this thesis is mainly based on work carried out in several research projects on spoken machine translation: Aliado (2004-2006) funded by the Spanish Government (TIC2002-04447-C02); TC-STAR (2004-2007) funded by the European Union (IST-2002-FP6-506738); Avivavoz (2007-2009) funded by the Spanish Government (TEC2006-13964-C03); and tecnoparla (2007-2009) funded by the Catalan Government.

7. Achievements

Actively pursuing the major thesis objective has lead to the following main research contribution to the SMT field:

- A novel approach for solving the word reordering challenge in an SMT system: a first-pass translation is performed on the source-text, converting it to an intermediate representation, in which source-language words are presented in an order that more closely matches that of the target language. This first translation is performed using an Ngram-based system. Reordering is coupled with translation, which then allows a choice among multiple reordering paths.

Further research contributions are:

- A novel approach for solving local word reorderings in an SMT system. The main limitation is that it addresses reordering in a deterministic way (a fixed reordering is given to the SMT system).
- Experimental work to introduce continuous space language models both in phrase- and Ngram-based SMT systems and its influence in translation.
- A study of two state-of-the-art SMT systems mentioned above. This study leads to system combination at the rescoring level.
- Construction of several SMT systems which were presented at International Evaluation Campaigns. Building a machine translation system is a serious undertaking. The participants are usually provided with a common set of training and test data. Therefore, systems are evaluated under similar conditions, generally with automatic and human measures.

The main techniques and results of this PhD thesis have given place to aproximately 40 international publications [1] including the following publications in JCR journals [3, 2, 4, 1, 5, 6]. The [3] work won the 2009 RTTH BEST PAPER AWARD. Additionally, the PhD itself has been published by VDM Verlag [7]. Finally, a SMT demo system where the PhD author participated very activily is available at *http://www.n-ii.org*. This demo received the 2009 Iberian SLTech BEST DEMO AWARD.

8. References

[1] M. R. Costa-jussà and J. A. R. Fonollosa, "State-of-the-art word reordering approaches in statistical machine translation," *IEICE Transactions on Information and Systems*, vol. 92, no. 11, pp. 2179–2185, November 2009.

[2] M. R. Costa-jussà, J. A. R. Fonollosa, and E. Monte, "Recursive alignment block classification technique for word reordering in statistical machine translation," *Language Resources and Evaluation*, p. Accepted for publication, 2010.

[3] M. R. Costa-jussà and J. A. R. Fonollosa, "An ngram-based reordering model," *Comput. Speech Lang.*, vol. 23, no. 3, pp. 362–375, 2009.

[4] M. R. Costa-jussà and J. A. R. Fonollosa, "Phrase and ngram-based statistical machine translation system combination," *Applied Artificial Intelligence: An International Journal*, vol. 23, no. 7, pp. 694–711, August 2009.

[5] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa, and M. R. Costa-jussà, "N-gram based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, December 2006.

[6] M. R. Costa-jussà, "New reordering and modeling approaches for statistical machine translation," *Knowledge and Engineering Review*, p. Accepted for publication, 2010.

[7] ——, *Improving reordering and modeling in statistical machine translation*. ISBN 978-3-639-23568-5 VDM Verlag,Saarbruecken, Germany, 2010.

[1] Please visit the author personal web page at *http://gps-tsc.upc.es/veu/personal/mruiz/*

Coded-speech recognition over IP networks

Ph.D. José L. Carmona
Advisors: Antonio M. Peinado, José L. Pérez-Córdoba

Dpto. Teoría de la Señal, Telemática y Comunicaciones, University of Granada
{maqueda, amp, jlpc} @ ugr.es

Abstract

In this Ph.D. dissertation the influence of packet losses on speech recognition is analyzed and different solutions to prevent, reduce and conceal their effects are developed. The performance of remote speech recognition will be subject to the robustness of the speech coding scheme used. Conventional speech codecs achieve to reduce the bit-rate by making use of predictive techniques that exploit temporal speech correlations. Thus, to decode a frame, a correct decoding of the previous ones is required. However, this inter-frame dependency reduces considerably the robustness against packet losses because it originates an error propagation in addition to the associated information loss. Furthermore, speech decoders integrate their own packet loss concealment algorithms, which are based on perceptual considerations that are unsuitable for speech recognition. In order to combat these degradations, we propose a set of mechanisms that can be divided into sender-driven and receiver-based techniques.

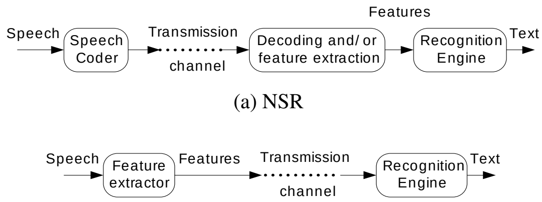
Index Terms: Network speech recognition, robust speech recognition, packet loss concealment.

1. Nature of the problem

Thanks to the convergence of wireless technologies, access to information on the move is nowadays a technological reality on the increase. Nevertheless, mobile phone constraints, such as the lack of keypad for size reasons, hinder the access to remote services. Oral interaction with such services arises as a new faster and more natural means of access to information with the help of automatic speech recognition that offers an interactive service and fast access to information, what benefits the user, doing without the assistance operator at the other side, what benefits the provider. Unfortunately, there are several problems to install an automatic speech recognition subsystem into mobile terminals. It is mainly their size restrictions what limits the computation capacity and, therefore, the recogniser power and flexibility. The possibility of remote speech recognition, i.e. outside the terminal, emerged to overcome these obstacles.

Remote Speech Recognition (RSR) allows to circumvent these hardware constraints by moving the most complex computational tasks of speech recognition to a remote server. Moreover, the structure of a remote recognition system is well suited for the IP model, since it is the provider that implements the recogniser depending on its needs. Thus, the provider can incorporate new services adapted to the present needs of users. Under this point of view, low cost terminals with limited features are connected to powerful remote computers that carry out more complex tasks for them, what leads to an optimum use of centralised resources. As shown in Fig. 1, there are two possibilities for the implementation of an RSR system [1]:

1. Network Speech Recognition (NSR). In this approach, the whole recognition system resides in the network. Thus, the client sends the speech signal, employing a conventional speech codec, to the server where recognition is carried out.
2. Distributed Speech Recognition (DSR). The client includes a local front-end that processes the speech signal in order to obtain the specific features used by the remote server (back-end) to perform recognition.



(a) NSR

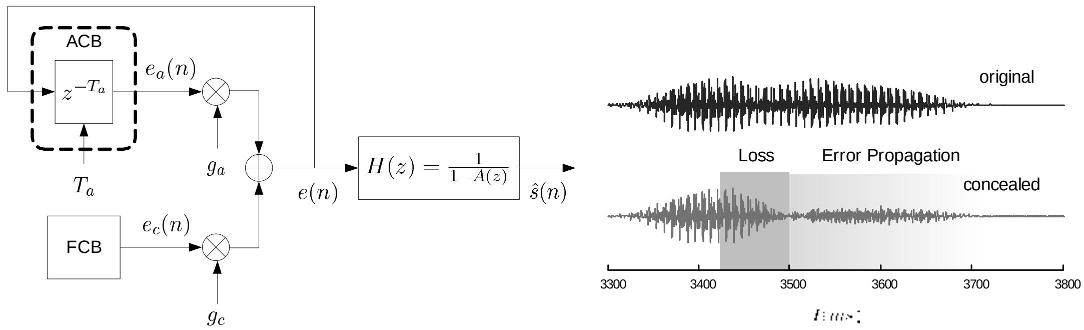
(b) DSR

Figure 1: Different architectures for the implementation of a Remote Speech Recognition (RSR) system. a) Network Speech Recognition (NSR); b) Distributed Speech Recognition (DSR).

In DSR, the feature extractor is applied directly to the speech signal to obtain a low dimensional representation with less redundant information. Although during the last years several standards have been issued, the lack of DSR codecs in the existing devices supposes a barrier for its deployment.

On the other hand, the most direct implementation of an RSR system is the speech transmission to the extreme server where the recognition task is performed, i.e., the NSR architecture. In this case, RSR is considered a value-added service of VoIP, since the coded speech is not transmitted to establish a call session, but to have access to a particular service. The main advantage of this type of application is the use of emerging IP platforms, without modifying in any respect the client terminal. However, the application has at the same time some disadvantages, since the loss information that speech coding involves may affect performance. There are also some implicit problems in remote recognition. Among them, two of the most outstanding ones are acoustic noise (the acoustic context of the terminal may vary) and degradations introduced by the communication channel. This dissertation is focused on the second one, since IP networks design, which offer a best-effort service, does not guarantee to meet real time requirements (delay and jitter) nor reliability requirements of multimedia flow transmission (packet losses).

This thesis [2] aims to develop a set of mechanisms to improve the performance of NSR systems considering the degradation of the current IP networks. In particular, such mechanisms should overcome the loss of VoIP packets over the network, optimising the performance of recognition tasks. In this sense, two types of measures can be adopted: sender-driven or



(a) (b)

Figure 2: Impact of a packet loss on a CELP-based codec: (a) Decoder structure; (b) Speech synthesis applying the packet loss concealment included in the speec codec.

receiver-based. The first type of these measures attempts to make speech coding schemes more robust against packet loss effects. In this case, modifications in the coding scheme will be applied taking into account a double objective: improve both the subjective perceptual quality of coding systems and speech recognition. The second type of measures will develop loss concealment techniques for coded speech recognition. In this case, these techniques only require to modify the server structure, and therefore they focus on the improvement of the recogniser performance disregarding perceptual aspects.

The rest of this extended abstract is organized as follows. In Section II we described the methodology that we followed to develop this Ph.D. thesis. Section III is devoted to describe the experimental framework and the most relevant results. In Section IV we summarize our conclusions. Finally, Section V collects the most important references and publications derived from this Ph.D. dissertation.

2. Methodology

The development of this research has been carried out by means of a methodology based on three stages:

1. Preliminary study of basic NSR systems. In this part we analyzed the performance degradation caused by packet losses in the speech decoding process.
2. Proposals for robust speech encoding. At this stage we took as initial hypothesis that if the proposed techniques achieve strengthen coding schemes from a perceptual point of view, then also increase the performance of recognition systems. Finally, we verified this hypothesis by carrying out speech recognition tests.
3. Proposals for packet loss concealment algorithm at the receiver. Here, we proposed different algorithms for the concealment of those degradations originated by packet losses at the receiver side.

2.1. Speech decoding in presence of packet losses

Most current speech codecs are based on the CELP (Code Excited Linear Prediction) paradigm, since it provides a high-quality synthesis at a low bit-rate. In particular, the speech synthesis is carried out by filtering the excitation signal $e(n)$ through the LPC (Linear Predictive Coding) filter $H(z)$. The general structure of a CELP decoder is shown in Fig. 2a. As we can see, the excitation signal is produced as the sum of the signals $e_p(n)$, obtained from an adaptive codebook (ACB), and $e_c(n)$, obtained from a fixed codebook (FCB), weighted by their corresponding gains g_p and g_c. The fixed codebook contains a number of innovation sequences, whereas the adaptive codebook models the long-term correlation of the excitation signal (related to the pitch period). For this reason, the adaptive codebook is dynamically built from the previous excitation samples by means of a long-term predictor (LTP).

Unfortunately, the inter-frame dependencies in the encoding process endanger the performance in packet networks. Fig. 2b illustrates this issue. As shown, once a packet loss occurs, the predictive schemes used by the encoder prevent from obtaining the correct decoded parameters. Furthermore, even when the decoder has already the correct parameters, there exists an error propagation caused by the ACB contribution to the excitation signal [3].

2.2. Robust speech coding techniques

Sender-driven loss concealment techniques for NSR are only justifiable if they optimise both perceptual quality and speech recognition accuracy. The main advantage of the NSR architecture is that no terminal changes are required, so modifications can not be justified only by recognition improvements. For this reason, our proposals are based on modifications of the coding scheme to improve the speech synthesis and, subsequently, the recognition performance.

iLBC is a speech codec specially conceived for packet networks, such as Internet, since it was designed to combat packet losses. To achieve this goal, iLBC does not exploit the correlation between adjacent frames in the excitation encoding. Thus, iLBC removes the interframe dependencies at the cost of a higher bit-rate than other coding techniques [4].

We proposed combining iLBC and CELP schemes in order to obtain a robust performance against packet losses while reducing the bit-rate of iLBC. The idea is based on using independent (iLBC) and dependent (CELP) frames in the same way that video codecs do. Thus, in case of packet losses, the error propagation of CELP frames is limited by the iLBC frames (key frames), which act as firewalls. At the same time, the insertion of CELP frames reduces the bit-rate [5].

Our second approach is also oriented to remove the error propagation caused by packet losses. In this sense, a FEC (Forward Error Correction) code consisting of the previous excitation samples for every frame will remove the possible error propagation. Of course, this completely alienates the CELP

ASR Architecture	*Bit-Rate (kbps)*	*Channel Conditions*					*Avg. Value*
		C0	*C1*	*C2*	*C3*	*C4*	
G.729 baseline	*8*	98.81	98.02	89.87	83.13	75.98	89.16
AMR baseline	*12.2*	98.68	97.93	93.97	88.55	83.07	92.44
iLBC baseline	*15.2*	98.96	98.56	96.35	92.43	87.11	94.68
G.729 B-NSR	*8*	98.82	98.52	97.64	95.69	92.89	96.71
AMR B-NSR	*12.2*	98.79	98.59	97.69	95.96	93.61	96.93
iLBC B-NSR	*15.2*	98.94	98.85	98.22	96.29	93.32	97.12
DSR FE	*4.75*	99.04	99.04	98.65	97.10	94.10	97.59

Table 1: Summary of the most relevant results.

coding idea, increasing the bitrate up to unusable limits. Instead, we proposed to encode only the most representative excitation samples by means of a multipulse scheme [6, 7].

2.3. Receiver-based PLC techniques

The packet loss concealment algorithms implemented in the decoders are unsuitable for recognition tasks. Such algorithms are based on perceptual considerations that are not appropriate for recognition. Thus, when several consecutive packets are lost, decoders progressively mute, leading to an increase on the insertion errors in the recogniser (artificial silences).

Under the perspective of an NSR system, packet losses and error propagation can be jointly treated on the feature vectors extracted by the speech recognizer. In particular, we proposed solutions based on a Bayesian MMSE (Minimum Mean Square Error) estimation of those feature vectors affected by packet losses. In general, this estimate can be expressed as $\hat{\mathbf{x}}_t = E[\boldsymbol{x}_t|\Lambda]$, where $\boldsymbol{x}_t$ corresponds to the original feature vector and Λ represents the *a priori* knowledge about $\boldsymbol{x}_t$ [8, 9, 10]. Thus, the corrupted feature vectors are replaced by the expected value of the uncorrupted ones given some additional information Λ. This information consists of those feature vectors before and after a given loss, which is refined using a hidden Markov model (HMM). This model allows us to consider the distortions introduced by packet losses (including error propagation) and the temporal correlations of the speech signal [11, 12, 13].

One of the virtues of our MMSE estimate is that we can obtain some information on the confidence associated with each of the reconstructions performed. Additionally, this information can be used by recognizer-based techniques, such as the Soft-Data approach and the Weighted Viterbi algorithm (WVA), in order to consider this uncertainty in the recognition process [14]. These techniques require that the recognizer must be fed, as usual, with the feature vectors provided by the PLC block plus a reliability measure for those features. In this Ph.D. dissertation we proposed different techniques to compute the reliability factors, carrying out a comparative study of the results obtained by both approaches [11].

The architecture NSR can be modified in order to extract the feature vectors from the codec parameters (see Fig. 1a). This variant, called B-NSR (Bitstream-based NSR) or transparameterization, avoids the speech signal reconstruction by introducing a bitstream-based feature extraction that directly transforms the received codec parameters into recognition features. There are several reasons why the B-NSR approach can be attractive, among which are the following:

- Speech codecs usually include some type of post-processing at the decoder in order to obtain a decoded signal perceptually improved. However, this post-processing is not optimized for an objective performance measure as in speech recognition.
- It is not necessary to reconstruct the speech signal. This provides a computational saving.

For these reasons, in this Ph.D. thesis we proposed several transparameterization schemes for different popular speech codecs [11, 15]. This approach also allows us to develop an efficient adaptation, in terms of word accuracy and computational resources, of those techniques based on MMSE estimation and uncertainty treatment (soft-data and WVA), which were described above [11].

3. Experimental results

The experimental setup is based on the framework proposed by the ETSI STQ-Aurora working group in [16]. The Aurora DSR front-end [17] provides a 14-dimension feature vector containing 13 MFCC (Mel Frequency Cepstral Coefficients) plus log-Energy. Furthermore, these vectors are extended by appending the first and second derivatives of the features. The recognizer is the one provided by Aurora and uses eleven 16-state continuous HMM word models, (plus silence and pause, that have 3 and 1 states, respectively), with 3 gaussians per state (except silence, with 6 gaussians per state). The training and testing data are extracted from the Aurora-2 database (connected digits). Training is performed with 8400 clean sentences and test is carried out over set *A* (4004 clean sentences distributed into 4 subsets).

In this work we have used two widely used CELP-based codecs: G.729A and AMR (Adaptive Multi-Rate) mode 12.2 kbps. In addition, iLBC (internet Low Bit-rate Codec) is also included because its design is oriented to increase the robustness against packet losses.

The channel burstiness exhibited by lossy packet networks was modelled by a 2-state Markov model. In particular, we obtained a wide set of channel conditions from a packet loss rate of 0% (condition C0) to 20% with a mean burst duration of 4 consecutive packets (condition C4), which simulate realistic situations of wired and wireless channels.

Since this paper is an extended abstract of a wide research work, we will only show the speech recognition results obtained by the best of our proposals. However, further readers can find a complete analysis of results in [2]. Table 1 collects a brief summary of our experimental results. In first place, we show the baseline results for G.729, AMR 12.2 and iLBC. These results correspond to carry out the recognition task using directly the decoded speech. On the other hand, the following rows show the results obtained by the combination of a B-NSR scheme, as feature extraction method, and the best proposed receiver-based PLC technique. In particular, this PLC technique carries out the recontruction of those feature vectors affected by packet losses by means of an MMSE estimation, which additionally assigns confidence values to the estimates. These confidence values are taken into account during the recognition process to achieve further improvements. The last row shows the results obtained by the DSR approach described in [17], which can be considered as an upper limit for NSR. In this comparison, we must highlight that our proposals clearly outperform the speech recognition from decoded speech and they reduce considerably the differences between NSR and DSR.

4. Conclusions

This research work presents a thorough study of the degradations suffered by coded-speech recognition systems in presence of packet losses. In this sense, we have identified a new degradation source that is given by the inter-frame dependencies introduced by speech codecs. Thus, when a packet loss happens, in addition to the loss of information, a propagated error appears after the loss. In order to tackle these problems, we have proposed a wide range of packet loss concealment techniques based on the sender and the receiver. The first group of techniques modifying the structure of the client in order to improve the perceptual quality of the speech synthesis and, subsequently, the speech recognition accuracy. On the contrary, the proposed receiver-based techniques focus exclusively on improving the speech recognition accuracy by replacing those features vectors affected by packet losses and considering the remaining uncertainty in the recognizer. This Ph.D. dissertation, in conclusion, just to facilitate the understanding of degradation suffered by speech codecs in presence of packet losses and the wide range of proposed solutions, should serve as a reference work to develop new packet loss concealment techniques.

5. Acknowledgments

This work was supported by the Spanish MEC in the project FEDER TEC2007-66600 and the national plan for mobility of researchers.

6. References

[1] A.M. Peinado and J.C. Segura. *Speech recognition over digital channels. Robustness and standards*. Wiley, 2006.

[2] J.L. Carmona. *Coded-speech recognition over IP networks*. Ph.D. thesis, University of Granada, 2009. http://digibug.ugr.es/bitstream/10481/2349/1/18056052.pdf

[3] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, V. Sánchez, A.M. Gómez, "Rendimiento perceptual y reconocimiento con codificadores VoIP sobre redes de paquetes," *Actas de las IV Jornadas de la Red de Tecnologas del Habla*, pp. 249-254, Zaragoza (Spain), 2006.

[4] A.M. Gómez, J.L. Carmona, A.M. Peinado, V. Sánchez, "Intelligibility evaluation of Ramsey-derived interleavers for Internet voice streaming with the iLBC codec," *in proceedings of Interspeech 2008*, Brisbane (Australia), 2008.

[5] J.L. Carmona, J.L. Pérez-Córdoba, A.M. Peinado, A.M. Gómez, J.A. González, "A scalable coding scheme based on interframe dependency limitation," *in proceedings of ICASSP'08*, pp. 4805-4809, Las Vegas (USA), 2008.

[6] A.M. Gómez, J.L. Carmona, A.M. Peinado, V. Sánchez, "A multipulse-based forward error correction technique for robust CELP-coded speech transmission over erasure channels," *IEEE Trans. Audio Speech Lang. Process.*, Vol. 8, No. 6, pp. 1258-1268, 2010.

[7] J.L. Carmona, A.M. Gómez, A.M. Peinado, J.L. Pérez-Córdoba, J.A. Gonzaĺez, "A multipulse FEC scheme based on amplitude estimation for CELP codecs over packet networks,"*in proceedings of Interspeech 2010*, Makuhari (Japan), Septiembre, 2010.

[8] A.M. Gómez, A.M. Peinado, V. Sánchez, J.L. Carmona, "Interleaving and MMSE estimation with VQ replicas for distributed speech recognition over lossy packet networks," *in proceedings of Interspeech 2006*, Pittsburgh (USA), 2006.

[9] J.A. González, A.M. Peinado, A.M. Gómez, J.L. Carmona, J.A. Morales, "Efficient VQ-based MMSE estimation for robust speech recognition," *in proceedings of ICASSP'10*, Dallas (USA), 2010.

[10] J.A. González, A.M. Peinado, A.M. Gómez, J.L. Carmona, "Efficient MMSE estimation and uncertainty processing for multienvironment robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, in press, pp. 1-15, 2010.

[11] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba and A.M. Gómez, "MMSE-based packet loss concealment for CELP-coded speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, Vol. 8, No. 6, pp. 1341-1353, 2010.

[12] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, A.M. Gómez, J.A. González,"Robust encoded speech recognition over IP networks,"*in proceedings of SIMPE 2010*, Lisbon (Portugal), Septiembre, 2010.

[13] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, A.M. Gómez, J.A. González,"Dealing with acoustic noise and packet loss in VoIP recognition systems," *to appear in proceedings of FALA 2010*, Vigo (Spain), 2010.

[14] A.M. Gómez, A.M. Peinado, V. Sánchez, J.L. Carmona, "A robust scheme for distributed speech recognition over loss-prone packet channels," *Speech Communication*, Vol. 51, pp. 390-400, 2009.

[15] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, V. Sánchez and A.M. Gómez, "iLBC-based transparameterization: A real alternative to DSR for speech recognition over packet networks," *in proceedings of ICASSP'07*, Honolulu, USA, 2007.

[16] H.G. Hirsh and D. Pearce. *The Aurora experimental framework for the performance evaluations of speech recognitions systems under noise conditions*. ISCA ITRW ASR, 2000.

[17] ETSI ES 201 108. *Front-end feature extraction algorithm; Compression algorithms*, 2000.

Analysis, design and application of flexible, contextual and dynamic dialogue management solutions based on Bayesian Networks

Fernando Fernández-Martínez, Javier Ferreiros

Grupo de Tecnología del Habla,
Universidad Politécnica de Madrid, Madrid, Spain.
ffm@die.upm.es, jfl@die.upm.es

Abstract

In this thesis we tackle the problem of identifying the best practices when designing and evaluating a spoken dialogue system. With the purpose of demonstrating that a more natural, flexible and robust dialogue is possible, and introducing a spoken dialogue system for controlling a Hi-Fi audio system as the selected prototype, we propose a Bayesian Networks (BNs) based solution for dialogue modelling combined with carefully designed contextual information handling strategies. Dynamic capabilities are also provided to keep the dialogue context permanently updated according to the evolution of the dialogue. All the thesis contributions have been evaluated finding an experimental support enough to demonstrate their relevance.

Index Terms: spoken dialogue systems, mixed initiative, Bayesian Networks, contextual information, usability, real users evaluation, electronic devices control

1. Introduction

Speech is the most widely used natural means of communication between people. Speech also is of increasing importance as a user-machine interface. As a result of the knowledge and the experience accumulated during almost half a century of speech technology research, now the time has come to design automated dialogue systems that make use of the communicative aspects of speech. In particular, it is essential to incorporate to the design of such systems some ideas related to the concept of "ambient intelligence" (AmI), for providing intelligent interfaces that are able to conduct a natural dialogue, including negotiations in order to achieve the goals required by users.

A dialogue system can be seen as a computer application that enables interaction and communication between users and machines as naturally as possible. Besides the typical recognition and text-to-speech conversion modules and other components, dialogue systems usually contain a module called Dialogue Manager (DM). This module is responsible for a dual task: to interpret the intention of the user and to decide how to continue the dialogue.

To successfully provide users with answers resembling a human-human interaction as much as possible, we believe that the design of a dialogue system should be approached from both a theoretical and practical point of view [1]. Thus, we must pay attention not only to dialogue management and modelling, but also to the enhancement of such models with knowledge about the specific tasks of the dialogue and the application domain (i.e. task and domain models). That way, it is feasible to develop procedures that support the user-machine interaction by useful elements of communication for realizing a collaborative and cooperative dialogue.

2. Dialogue management based on BNs

2.1. On the spoken dialogue system

Our conversational interface allows users to drive a Hi-Fi system from natural language sentences, differentially from other typical control systems based on simple commands. A detailed description of this system, its architecture and the implemented dialogue strategies can be found in [3][5][7].

2.2. On the dialogue management solution

As an alternative to classical dialogue systems (finite state automata or FSMs, script based systems or dialogue plans, etc.), we are presenting a dialogue solution based on BNs, that allows a greater flexibility and naturalness by appropriately defining dialogue as the interaction with an inference system [2].

The first task of the Dialogue Manager (DM) module is to identify the intention of the user (i.e. dialogue goals) considering the relevant information extracted by a semantic parser from the last utterance (i.e. available concepts) [6][13], together with the dialogue context. Then, according to the inferred goals the DM has to make a decision regarding how the dialogue should continue. Both tasks can be accomplished using BNs.

2.2.1. *"Forward Inference"*

As can be observed in Figure 1, BNs can be adopted to model the existing causal relation between the goals and the concepts [2][3][7]. Typically, both of them are assumed to be binary (i.e. a concept is true or "present" only when it is observed in the sentence). Thus, from the whole set of available evidences, e.g. $E = \{C_1 = 0, C_2 = 1, ..., C_N = 1\}$ for N defined concepts, a posterior probability $P(G_i = 1|E)$ can be obtained for each goal using the "Forward Inference" (FI) technique [2].

Subsequently, a decision is made for each goal on the comparison of the posterior with a defined threshold, θ. As a result of that comparison, one goal is "active" or "present" if the corresponding posterior is over the threshold ("absent" if not).

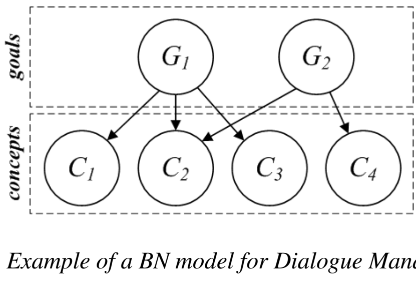


Figure 1: *Example of a BN model for Dialogue Management.*

Table 1: *Concept analysis used to drive the dialogue.*

	$P(C_j = 1\|E^*) < \theta$	$P(C_j = 1\|E^*) \geq \theta$
C_j absent ($C_j = 0$)	C_j **unnecessary** (No action)	C_j **missing** (Prompt to request C_j)
C_j present ($C_j = 1$)	C_j **wrong** (Prompt to clarify or notify about C_j)	C_j **required** (C_j is stored in the dialogue memory)

2.2.2. *"Backward Inference"*

After the FI process, and assuming the inferred results (i.e. those goals which were decided to be "present", $G_i = 1$) as new evidences, Bayesian inference can be applied again but this time aimed at the estimation of $P(C_j = 1|E^*)$, the probability that each concept should be present where E^* refers to the updated set of evidences (i.e. E also including goal evidences obtained through the FI process but removing the evidence corresponding to the target concept, C_j). This process is known as the "Backward Inference" (BI) technique [2]. Making a similar binary decision on the value of $P(C_j = 1|E^*)$, it is possible to check whether that concept should be present or not.

2.2.3. *Concept analysis*

The BI result can be compared with the actual occurrence of the concept enabling the classification presented in Table 1. As a result of that analysis [2] every concept can be properly classified allowing the DM to perform a suitable action. A possible dialogue proceeding strategy has been suggested for each possible result. For example, the system can drive the dialog prompting for the "missing" concepts.

3. Main thesis contributions

In this thesis, we have completed a thorough and comprehensive study of the BNs. In particular, we have focused on their possible application as new dialogue management solutions.

In the following subsections we will sucessively highlight the main contributions with regard to each explored solution.

3.1. Regarding the BN approach

The main advantages that we can highlight in this regard are:

- The BNs based inference system enables, through the FI process, **a better identification of the dialogue goals according to the intention of the user** (i.e. actions that the user may request the system to perform) **from the available concepts consistently with the context of the ongoing dialogue**. FI evaluation results at [9] showed a F-measure of $92,29\%$ regarding goal identification.
- BN models are defined at a semantic level. This allows their design **with independence of the language used**.
- **BNs can be automatically learnt from training data**. This favors portability and scalability across domains.
- BNs allow **a simple way of incorporating human knowledge**, e.g. refining the topology by hand.
- BNs allow to conduct an **analysis of congruence** between the goals assumed by the system to have been requested by the user, and all data collected during the interaction. Based on this analysis, the system can **determine the flow of interaction** and react according to the semantics of the application domain. In particular, through the BI process, it is possible to automatically detect **which concepts are needed** (available or not), **erroneous** or **optional** with regard to the inferred goals. Thus, the dialogue could go toward the generation of messages requesting the missing items, clarifying the erroneous ones and ignoring the optional ones. The performance of the BI process and its derived concept classification showed an $81,00\%$ F-measure at [9].
- The BNs enable **a true mixed initiative dialogue modelling**. Flexibility is probably the main asset of the proposed solution, and the most significant difference with regard to conventional approaches. In particular, **the user is not constrained to any predetermined goal or data sequence**. This flexibility is twofold, since it not only allows the user to decide the goals at the beginning of interaction, but also lets him/her jump to other goals without having completed the previous ones. Moreover, the user can respond with more data than those requested in a query, or even respond to a fact not asked by the system with regard to the inferred dialogue goals.
- Thanks to the **negotiation process** enabled between the users and the system, based on the FI and BI procedures, the system is capable of responding to complex issues (e.g. when the users provide inaccurate or insufficient information to meet the required dialogue goals) and **to assist or guide the users** toward the achievement of their dialogue goals driving the dialogue in an efficient manner, minimizing the number of questions and making maximum use of the context of dialogue.

3.2. Regarding the dynamic response of the system

We have laid the basis for enabling a dynamic response [10]:

- We have introduced the notion of **"relevance"** as the **remaining evidence** of a concept in the dialogue history.
- **Attenuation mechanisms** have been introduced that lower the relevance or the latency of information stored in past phases of the evolution of dialogue. Hence, the relevance of those elements can evolve to a level below a predefined threshold, so that they finally disappear definitively from the dialogue history. Due to this mechanism, it is possible to **maintain the dialogue history permanently updated** by assigning higher weight to more recent information, and lower to the older.

3.3. Regarding the use of a new BN based inference engine

In relation to the inference engine used by the DM [9]:

- We have presented a new alternative to traditional solutions based on multiple BN models (i.e. individually developed for each specific goal). In particular, we have proposed to rethink the inference problem from a **single global BN model** including all the defined concepts and goals. A "fusion" algorithm has been defined to obtain that BN model from the baseline multiple BN models.
- Unlike the baseline strategy, the proposed fusion method provides a single BN that ensures that both the FI and the BI processes are consistent with the dialogue context. Moreover, **the result of the analysis of congruence is also unique for each concept** and is obtained by considering a **whole goal evidence context**, thus avoiding potential mixed results for the same concept derived from analyzing each goal separately.

- This new solution offers **greatly improved performance in terms of BI**. By contrast, it offers a slightly lower performance in terms of FI. However, the fusion BN provides **a better overall performance** (i.e. combined F-measure is approximately 13% better [9]).
- We have designed solutions to **optimize the computational cost** of the models resulting from the fusion process (e.g. an information gain study from which we can select the most indicative concepts of each goal).

3.4. Regarding the use of contextual information

The DM is also provided with a set of contextual information handling strategies [3][11]. Regarding the benefits of applying those strategies we emphasize:

- **The robustness and the consistency of the system responses are improved** as the system is able to deal with dialogue phenomena such as "anaphora" (i.e. elements that refer to other previous parts of the dialogue) and "ellipsis"(i.e. omission of certain essential elements of the dialogue that may be derived from given context), the main dialogue phenomena that can mean a **loss of crucial information**. These strategies are based on:
  - the available confidence measures (from speech recognition and language understanding modules),
  - the history of the ongoing dialogue (∼short term),
  - the history of dialogue (∼long term, i.e. the dialogue concepts referred so far during the dialogue).
  - the status of the system (i.e. current values of the different parameters of the system, e.g. volume),
  - the task model (e.g. a semantic frame containing all the information needed to meet a specific goal),
  - and the application domain model (e.g. information on the number of tracks of a particular CD).
- To assess the relevance and appropriateness of the designed strategies, [1][8][12] we measured the **percentage of contextual turns** as the fraction of dialogue turns in which some of the strategies were successfully applied. In connection with that metric, we also measured the **percentage of system requests** which should be limited by the contextual capabilities of the system. The results for both metrics confirmed the valuable role of the contextual information handling strategies (i.e. more than half of the turns relied on this type of information) improving both dialogue efficiency and fluency.

3.5. Regarding the use of concept confidence measures

Regarding the proper consideration of concept confidence measures in terms of dialogue [4]:

- We have proposed the use of the confidence measures to weigh the evidence of the concepts. Thus, it is possible to incorporate these measures directly to both FI and BI processes, adopting the available confidences as the evidences from which to conduct the inference. As a consequence, both inference results are **directly weighted by the available concept confidences**.
- We have also expanded the analysis aimed at the correct classification of the concepts. This extension is based on the **incorporation of the confidence measures to the referred analysis** to exploit them by defining specific dialogue proceeding strategies for each confidence level.

3.6. Regarding the evaluations with real users

Most important results derived from these evaluations are:

- The results obtained from the defined set of metrics, collected automatically during the evaluation of different scenarios [1][8][12] (i.e. "basic", "advanced" and "free" scenarios, designed according to different initiative styles and task complexity levels), showed that a **suitable turn-taking algorithm** is essential to ensure a lively and effective dialogue.
- Those results also clearly showed the **learning process** that the user experiences while interacting with the system. Indeed, **"experience" proved to be a key factor regarding dialogue performance**. As the learning stage proceeds, the user is able to exploit the acquired experience leading to more fluent and efficient dialogues. **The user-system interaction improves as the user "learns" how to address the system**. This was supported by the fact that "free" scenarios, though allowing the user the highest degree of initiative and, therefore, favouring much more open and complex expressions, were precisely those that, objectively, performed the best.
- **Expert users are more efficient than novices** (i.e. need less dialogue turns to achieve the same thing). At the same time, novices rely more on contextual information resources. However, both types of users were able to establish productive dialogues with the system since the beginning. The negotiation that the system is able to establish with the users plays a key role in that regard. This negotiation allows users not only to achieve their goals, but also to accelerate the development of their dialogue skills, thereby improving the performance and the quality of the interaction with the system.
- Users tend to need significantly less feedback as they become more familiar with the system. Therefore, the behavior and **the response of the system must be tailored to different skill or experience levels**.
- We have defined a new actuation algorithm that provides **the proper sequence of execution** for those actions corresponding to the positively inferred goals, by combining the **prevalence relations** between those goals (i.e. priority information), and **the order that they appear in the sentence** (i.e. position information). This new solution ensures optimal usability since:
  - the system is acting as soon as it is possible resulting in a much more natural interaction,
  - it is acting even in the case that there are incompleted goals, thus resulting in a flexible interaction,
  - system's actuation is suitable and tidy since it respects both priority and position information, thus resulting in a more robust interaction as well.
- The new actuation algorithm allows optimal usability **solving the problem of potential "blockings"**. Blockings are produced by the observation of active but incomplete goals. This could prevent the execution of every other action corresponding to any goal that, though ready to be executed, either has a lower priority or appears later in the sentence with the same priority. As blockings are avoided, **dialogue performance (i.e. turn efficiency) improves**.

- To better understand how dialogue systems work, we also made **a correlation analysis** between different dialogue metrics. Main results are summarized below:
  - **High values of contextual turns** tend to be associated with **low values of system requests**.
  - **High values of null-efficiency turns** (i.e. out-of-domain sentences or recognition rejections) tend to be associated with a **poor contextuality**.
  - **High values of system requests** tend to be associated with **high values of null-efficiency turns**.
  - **A high turn efficiency** is usually associated with **high contextuality levels**. Yet a high contextuality helps, it does not guarantee a good turn efficiency.
  - **Low values of system requests** mean **good turn efficiencies**.
  - Logically, the lower the null-efficiency turns, the greater the turn efficiency.
- We almost reached the number of **two goals satisfied per turn**. This is a good outcome, especially bearing in mind that users were not given any specification regarding the number of turns in which they had to try to overcome the different scenarios. Therefore, the possibilities of the system in this regard are yet to be fully exploited.
- A better efficiency has led to a more flexible and fluent dialogue which, in turn, has **improved the system's response**. This improvement has been assessed very positively by users. Particularly, **"free" scenarios were ranked as the highest-rated**. This is a result of particular importance since free scenarios lacked of any restriction (i.e. complexity was maximum) and, indeed, were the nearest scenarios to the actual use of the interface.

3.7. Regarding the design methodology

Finally, this thesis delves into the analysis and implementation of efficient mechanisms and techniques that minimize the effort required to generate a new dialogue system (change of semantic context). We proposed the use of strategies for characterizing the application domain and that enable the automatic learning of dialogue models. This methodology allows to obtain a full dialogue model for any application based on the analysis of suitably labeled real situations and a description of the data model along with a semantic description of the application (ontology).

4. General conclusions

The intention of this doctoral thesis was to introduce new ideas whose application to dialogue modelling could prove to be useful. The scientific and technological results obtained enable the design of better devices and intelligent interfaces that fully integrate features that facilitate portability across domains and languages, and improve all aspects of interaction with the end user.

Generally, user satisfaction in relation to a particular system crucially depends on its "usability" and "functionality". To be "useful", a system must be "usable" first (i.e. providing services for which it is designed efficiently) and also "functional" (i.e. the services provided are of interest to users).

One of the keys for the usability of a system, and by extension for its usefulness, is its simplicity of use. The greater or lesser ease of use that a system is able to offer (and also the offered functionality), definitely conditions the final acceptance by users. Easiness of use was the best appreciated feature by users, so it can be considered one of the most important results.

In order to get a fluent and efficient dialogue, the user-sytem interaction should be: **natural, flexible and robust**. It is difficult to attribute each of the above features to a single aspect of the various dialogue solutions proposed. Rather, it is thanks to the synergy of these solutions, to the joint operation of all of them, how those characteristics become true.

In short, a more natural, flexible and robust dialogue is possible thanks to the solutions for dialogue modelling based on BNs that have been suggested This is supported by a good user satisfaction rate and by the results corresponding to the metrics that were automatically collected [1][8][12], which have shown the usefulness and benefits provided by the proposed solutions.

5. Acknowledgements

The authors would especially like to thank all the members of the EDECAN project consortium (TIN2005-08660-C04) who kindly collaborated in the evaluations.

6. References

[1] F. Fernández, PhD Thesis: "Análisis, diseño y aplicación de modelos de diálogo flexibles, contextuales y dinámicos basados en Redes Bayesianas", E.T.S.I.T., Universidad Politécnica de Madrid, Spain, 2008, "http://oa.upm.es/1810/".

[2] H.M. Meng, C.Wai and R.Pieraccini, "The use of belief networks for mixed-initiative dialog modelling", IEEE Trans. on Speech and Audio Processing, 2003, vol.11, n.6, pp.757-773.

[3] F. Fernández et al., "Speech interface for controlling an Hi-fi audio system based on a bayesian belief networks approach for dialog modelling", Eurospeech 2005, Lisboa (Portugal).

[4] J. Ferreiros, F. Fernández et al., "New Word-Level and Sentence-Level Confidence Scoring Using Graph Theory Calculus and its Evaluation on Speech Understanding", Eurospeech 2005, Lisboa.

[5] F. Fernández et al., "Demostración de una interfaz vocal para el control de un sistema de alta fidelidad", Revista Procesamiento del Lenguaje Natural (ISSN 1135-5948), Ed. SEPLN, N 35, Septiembre 2005, pp. 451-452.

[6] F. Fernández et al., "Human spontaneity and linguistic coverage: two related factors relevant to the performance of automatic understanding of ATC speech", IEEE Aerospace & Electronic Systems Magazine (ISSN 0885-8985, JCR 2006: 0, 423), Ed. IEEE-INST (USA), Vol. 21, No. 10, pp. 12-17, October 2006.

[7] R. San-Segundo, F. Fernández et al., "Speech technology at home: enhanced interfaces for people with disabilities", Journal of Intelligent Automation & Soft Computing (ISSN 1079-8587, JCR 2009: 0, 349), Ed. TSI (USA), Vol. 15, No. 4, pp. 645-664, 2009.

[8] F. Fernández et al., "Evaluation of a spoken dialogue system for controlling a Hifi audio system", IEEE SLT 2008, Goa (India).

[9] F. Fernández et al., "A Bayesian Networks approach for dialog modelling: The fusion BN", IEEE ICASSP09, Taipei (Taiwan).

[10] J.M.Lucas, F. Fernández, J.Ferreiros, "Using Dialogue-Based Dynamic Language Models for Improving Speech Recognition", Interspeech 2009 (ISSN 1990-9772), Brighton (UK).

[11] F. Fernández et al., "Flexible, Robust and Dynamic Dialogue Modeling with a Speech Dialogue Interface for Controlling a Hi-Fi Audio System", IEEE DEXA 2010, (ISBN 978-3-642-03572-2, ISSN 1529-4188), Bilbao (Spain).

[12] F. Fernández et al., "HIFI-AV: An Audio-visual Corpus for Spoken Language Human-Machine Dialog Research in Spanish", LREC-ELRA 2010 (ISBN 2-9517408-4-0), Valletta (Malta).

[13] J.M.Pardo, J.Ferreiros, F. Fernández et al., "Automatic Understanding of ATC Speech: Study of Prospectives and Field Experiments for Several Controller Positions", IEEE Transactions on Aerospace & Electronic Systems (ISSN 0018-9251, JCR 2009: 1, 230), In press, 2010.

Extended Abstract for Best Ph.D. Thesis Award: *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*

Daniel Ramos Castro

ATVS - Biometric Recognition Group, Escuela Politécnica Superior
C./ Francisco Tomás y Valiente 11, Universidad Autónoma de Madrid E-28049 Madrid, Spain
daniel.ramos@uam.es

Abstract

This Ph.D. Thesis proposes a complete methodology for the adaptation of automatic speaker recognition technology to forensic evaluation of the evidence. The nature of the problem of forensic automatic speaker recognition is deeply analyzed in the context of the current debate about scientific procedures in forensic disciplines worldwide. Then, a solution for this problem is proposed in the form of a hyerarchical methodology which integrates current standards and state of the art of automatic speaker recognition technology and the requirements and needs of the so-called *coming paradigm shift* in forensic science. The Thesis contributions are supported by numeorus peer-reviwed publications in national and international conferences and jorunals included in ISI-JCR. Also, this Thesis and its constributions have been the recipient of several awards in different national and international contexts. Moreover, the applicability of the Thesis is evidence by the multiple public and private contracts and projects which consider the framework presented here, as well as the impact of the proposed methodologies in important fora such as working groups of the European Network of Forensic Science Institutes.

Index Terms: Forensic speaker recognition, likelihood ratio, calibration, empirical cross-entropy, coming paradigm shift.

1. Nature of the Problem

This Thesis is focused on the use of automatic speaker recognition systems for forensic identification, in what is called forensic automatic speaker recognition [1, 2]. More generally, forensic identification aims at individualization, defined as the certainty of distinguishing an object or person from any other in a given population [3]. This objective is followed by the analysis of the forensic evidence [4], understood as the comparison between two samples of material, such as glass, blood, speech, etc. An automatic speaker recognition system can be used in order to perform such comparison between some *recovered* speech material of questioned origin (e.g., an incriminating wire-tapping) and some *control* speech material coming from a suspect (e.g., recordings acquired in police facilities).

However, the evaluation of such evidence is not a trivial issue at all. In fact, the debate about the presentation of forensic evidence in a court of law is currently a hot topic in many scientific and legal fora [5, 6]. The American Daubert rules for the admissibility of the scientific evidence in trials and the evidence of critical errors in positive identification reports for disciplines assumed as error-free have fostered the discussion. From this debate, DNA profiling arises as a model for a scientifically defensible approach in forensic identification, as it meets the most stringent Court admissibility requirements demanding scientific evaluation of the evidence, and testability of procedures [6]. In this Thesis we take into account such requirements in order to adapt forensic automatic speaker recognition to what has been dubbed *the coming paradigm shift* in forensic identification science.

2. Hypotheses to be Tested and Objectives

The Thesis presented, which summarizes the hypothesis to be tested, can be stated as follows:

The emerging requirements for evidence evaluation and reporting in forensic science can be satisfied for forensic automatic speaker recognition by the use of accurate Likelihood Ratios (LR) within a hierarchical methodology consisting of 3 levels: discrimination, presentation, and forensic.

The main objectives of this PhD Thesis are:

1. Reviewing and studying the problem of automatic speaker recognition for forensic evidence evaluation.
2. Identifying all the steps which are needed for the use of an automatic speaker recognition system for forensic identification.
3. Analyzing the requirements of each of the steps and their relationship in order to give a coherent methodology.
4. Defining a methodology for the LR-based evaluation of the evidence using automatic speaker recognition systems based on the DNA paradigm.
5. Establishing a definition and assessment framework of the LR accuracy, aiming at clear interpretation of results.
6. Improving the discrimination of automatic speaker recognition technology.
7. Improving the accuracy and robustness of the LR computation process.
8. Applying the proposed evaluation, interpretation and assessment methodology to forensic speaker recognition problems, either simulating real cases or using databases coming from real police investigations.

3. Methodology and Relevant Theory

In Chapters 1 and 2, we begin by reviewing related works in the literature concerning automatic speaker recognition and foren-

This Thesis has been funded by a FPI Scholarship from Comunidad Autónoma de Madrid. The Thesis has obtained the Best Student Thesis Award from the Spanish Official College of Telecommunication Engineers (COIT) in 2009. The full Dissertation can be downloaded at http://atvs.ii.uam.es/files/2007_11_28_thesis_daniel_ramos_preprint_v1.pdf

sic evaluation of the evidence. Then, the experimental framework to be used in this Thesis is described in detail. The widely accepted Speaker Recognition Evaluations (SRE) conducted by the American National Institute of Standards and Technology (NIST) are adopted as the experimental set-up for this Thesis [7]. The databases used for such protocols constitute challenging corpora presenting many different variability factors, simulating the typical conditions of lawful recordings in telephonic networks.

As a contribution in this Thesis, a hierarchical methodology for forensic automatic speaker recognition is proposed in Chapter 4. This methodology constitutes a powerful tool for practitioners, as it allows transparent and testable forensic identification using the typical score-based automatic speaker recognition systems. We then identify the main factors affecting the methodology proposed in this Thesis. First the elements of the *coming paradigm shift* are analyzed [6]. Then, the common procedures accepted in automatic forensic speaker recognition are also identified. Taking into account all factors, we define the hierarchical methodology, consisting of three different levels of abstraction, namely the discrimination level, the presentation level and the forensic level.

The Dissertation then focuses on the description of the levels which compose the proposed hierarchical methodology. First, the discrimination level is addressed in Chapter 5. The aim at this level is yielding a discriminating score, as a way of distinguishing whether the speech coming from the suspect and the questioned recording come from the same source or not. Since discrimination has been the aim of automatic speaker recognition in the last decades, we give a definition of the performance of the score derived from the literature in the field. Moreover, we overview and experimentally compare several widely used techniques found in the literature in order to improve the discriminating power of a score set, namely score normalization [8], session variability compensation [9] and fusion of systems [10]. A novel score normalization technique, namely KL-T-Norm, is presented as a contribution [11]. We experimentally demonstrate that KL-T-Norm increases the discriminating power of other popular score normalization techniques such as T-Norm [8], as well as it improves its computational efficiency.

Next, the presentation level is introduced in Chapter 6. The aim at this level is transforming the input score into a *likelihood ratio* (LR) as a measure of the weight of the evidence, with a meaning of degree of support of the evidence to any of the hypotheses present in the case. This methodology, popularized by DNA profiling, is probabilistic, data-driven and allows to include in a logical way the weight of the evidence into the inferential process in a forensic case. A definition of the *accuracy* of the evidence evaluation process is then given, introducing the important concept of calibration. Then, a novel assessment methodology based on information theory is reported, where the accuracy of the LR values is expressed in the form of information-theoretical magnitudes, namely empirical cross-entropy (ECE).

Also in the presentation level, a comparative study of different LR computation techniques is presented. Among them, we propose a novel method of generative suspect-adapted LR computation. The study shows that the proposed technique improves the discrimination and the calibration of the input scores, by means of the exploitation of the specificities of a given suspect. The proposed technique is also robust to scarcity in the control speech material, a problem which is often found in forensic casework. The presentation level is concluded with an alternative configuration of the proposed methodology in order to consider non-score-based LR computation techniques, common in other forensic areas and recently proposed for automatic speaker recognition.

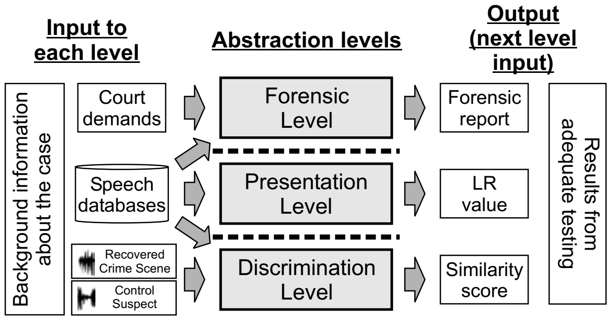


Figure 1: Proposed methodology for forensic evaluation of the evidence using speaker recognition systems, with inputs and outputs of each level in the hierarchy.

Finally, the last level in the hierarchy is described in Chapter 7, namely the forensic level. The aim at this level is considering the court demands and the requirements of the *coming paradigm shift* in forensic science in order to properly report the weight of the evidence and its accuracy. Two experimental examples illustrate the reporting and presentation of the results from evidence evaluation by means of the proposed information-theoretical assessment methodology. One of these examples has been built making use of the database and systems employed by the Spanish Guardia Civil in real forensic casework. The chapter ends with the demonstration of the adequacy of the proposed methodology for other forensic disciplines, by means of an experimental example of LR-based evidence evaluation using glass and paint analysis.

Figure 1 shows the proposed methodology for the use of automatic speaker recognition systems for forensic evidence evaluation. The described hierarchy of levels is shown, as well as the inputs and outputs of each level.

4. Contributions of the Thesis with Indicative References

The Thesis has generated a significant amount of research contributions, evidenced by the number of articles in conferences and journals with ISI-JCR impact factor. As a highlight, the results of the Thesis have received several awards and distinctions, namely:

- Best Ph.D. Thesis Award of the Official College of Telecommunication Engineers (COIT) in 2009.
- IBM Research Best Student Paper Award at the IEEE/ISCA Odyssey 2006 conference, for the article in [12].
- Finalist of the Spanish Network of Speech Technologies (RTTH) Best Journal Article Award, for the work in [11].

The research contributions of this PhD Thesis are the following (some publications are repeated in different items of the list):

- **Literature reviews**. Forensic evidence evaluation techniques in automatic speaker recognition [2, 12, 13];. Assessment of forensic speaker recognition systems [13, 14, 12] (IBM Research best student paper award). New requirements in forensic science [13, 2, 12] (IBM Research best student paper award). Automatic speaker recognition [15]. Score normalization for robust speaker verification [16][11] (RTTH best article finalist).

- **Theoretical frameworks.** Theoretical framework for the use of speaker recognition for forensic purposes [2, 12] (IBM Research best student paper award). Theoretical framework for the use of information theory for the assessment of LR values [17, 18, 19, 20].
- **Novel methods.** Novel methods for the use of automatic speaker recognition for forensic identification [2, 21] [12] (IBM Research best student paper award). Novel methods in robust LR computation [21, 14, 22, 23, 24, 20]. Novel methods in the assessment of LR values [17, 18, 19]. Novel methods of score normalization in speaker verification [16][11] (RTTH best article finalist).
- **Improvements in speaker recognition discrimination.** Contribution to the improvement of ATVS-UAM automatic speaker recognition system [2, 25, 14][11] (RTTH best article finalist).
- **New techniques in speaker verification.** New methods for the improvement of automatic speaker recognition discriminating power [26, 16][11] (RTTH best article finalist).
- **New experimental studies.** Experimental studies of automatic speaker recognition systems in the proposed methodology for forensic automatic speaker recognition [13, 2, 19, 12] (IBM Research best student paper award). Robustness in LR-based evaluation of the evidence [21, 14, 27, 22, 23, 24]. Calibration loss effects in forensic speaker recognition [13, 2, 19, 12] (IBM Research best student paper award). Reports on the ATVS-UAM automatic speaker recognition system with forensic applications at NIST SRE and at the NFI/TNO Forensic SRE [13, 2, 25, 26, 14, 12, 27, 16, 28, 22, 23, 24] [11] (RTTH best article finalist). Robust score normalization in speaker verification [16, 28][11] (RTTH best article finalist).
- **Application to other forensic disciplines.** Robust evidence evaluation methods in biometrics [27, 29]. Information theoretical evaluation of LR values coming from glass and paint evidences [17, 18, 19].

5. Results and Analysis

This section presents an analysis of the main results of this Thesis. First of all, the global contribution of the Thesis is the hierarchical methodology for forensic automatic speaker recognition, containing three levels (Figure 1).

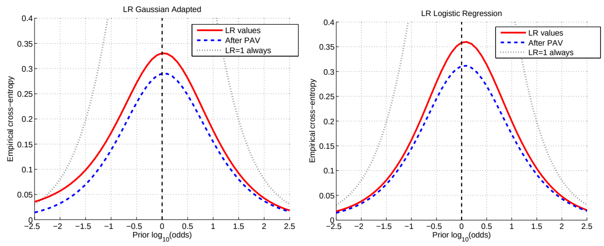
The Thesis also clearly defines the discrimination level, where a novel score normalization technique is proposed, namely KL-TNorm. This method improves the discriminating power of systems with respect to the state of the art in test-dependent score normalization, represented by T-Norm [8]. Table 1 illustrates the increase in discriminating power of KL-TNorm in one of the experimental set-ups in the Thesis, which can be seen as a reduction of the Equal Error Rate (EER).

At the defined presentation level, we present and compare several methodologies to compute this evidence weight in terms of likelihood ratios (LR), following the DNA standard. At this level, we define the precision of a set of LR values by means of the Empirical Cross-Entropy (ECE), and its decomposition into discrimination and calibration performance. This methodology allows the presentation of performance results in terms

GMM $K = 75$	1c-1c		8c-1c	
	male	female	male	female
EER T-Norm (Av.)	11.14	14.62	7.78	9.57
EER KL-T-Norm	10.76	13.88	7.25	9.12
EER Av. Improvement	3.4%	5.0%	6.8%	4.7%

Table 1: Comparison of EER for TNorm and the proposed KL-Tnorm for the for ATVS GMM system presented in NIST SRE 2005 [7].

of information theory, and can be illustrated in terms of ECE plots (Figure 2), which constitutes a step forward to make the understanding of the results presented by forensic practitioners easier to the court. Also at this level, we present a novel method for transforming scores from speaker recognition systems into LR values, namely suspect-adapted LR computation, which improves the state of the art of evidence evaluation methods when the amount of speech from the suspect is sparse (Figure 2).



(a) (b)

Figure 2: ECE plots to represent the accuracy of an experimental set of LR values. Red curve denotes information loss, and should be as low as possible. The difference among red and blue curves denote a calibration problem. Comparison of suspect-adapted (a) and logistic regression (b) LR computation.

Finally, we defined a forensic level in the proposed methodology, where the requirements of the *coming paradigm shift* in forensic science are taken into account. The whole methodology is tested by applying it to the procedures followed in the Acoustics Department of the Criminalistics Service of the Spanish Guardia Civil. Moreover, we demonstrate the applicability of the proposed methods to other forensic disciplines, such as Glass or Paint Analysis (Figure 3).

6. Applicability

The main results of this Thesis have been critical as research results and and technology transferred in the context of public and private research projects and contracts. In this sense, we have to highlight the stable collaboration agreement between the ATVS group and the Criminalistics Service of Spanish Guardia Civil. Nowadays, their Acoustics Department are implementing the methodology contributed in this Thesis for forensic evidence evaluation in real casework. Moreover, ATVS, and in particular the author and the advisor of this Ph.D. Thesis, are regular invited members of the Forensic Speech and Audio Analysis Working Group of the European Network of Forensic Science Institutes (ENFSI-FSAAWG), where the experience of Guardia Civil in the deployment of the proposed methods in casework serves as a driving standardization effort for other forensic lab-

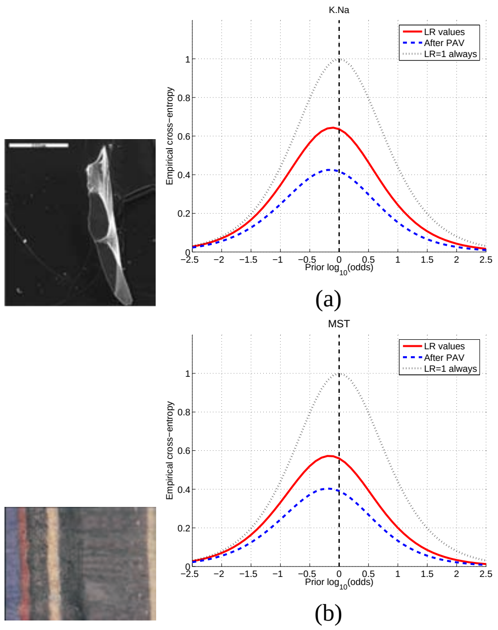


Figure 3: Application of the proposed methodology based on ECE plots to forensic glass (a) and paint (b) analysis.

oratories and police forces across Europe. Finally, part of the results in this Thesis have been used in technology transfer contracts with Agnitio S. L., and as part of the results of collaboration agreements with the Spanish Ministry of Defense.

7. References

[1] D. Meuwly, *Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique*, Ph.D. thesis, IPSC-Universite de Lausanne, 2001.

[2] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[3] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, pp. 193–203, 2000.

[4] C. G. G. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004.

[5] National Research Council, *Strengthening Forensic Science in the United States: A Path Forward*, National Academy of Sciences, 2009.

[6] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Science*, vol. 309, no. 5736, pp. 892–895, 2005.

[7] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the Mixer corpora-2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[10] N. Brümmer et al., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[11] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Speaker verification using speaker- and test-dependent fast score normalization," *Pattern Recognition Letters*, vol. 28, no. 1, pp. 90–98, 2007.

[12] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Likelihood ratio calibration in transparent and testable forensic speaker recognition," in *Proc. of Odyssey*, 2006.

[13] J. Gonzalez-Rodriguez and D. Ramos, "Forensic automatic speaker classification in the coming paradigm shift," in *Speaker Classification*, Christian Müller, Ed., vol. 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, 2007.

[14] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 331–355, 2006.

[15] J. Gonzalez-Rodriguez, D. Ramos-Castro, D. Garcia-Romero, and J. Ortega-Garcia, *Reconocimiento del Locutor*, chapter 7, pp. 174–206, RA-MA Editorial, 2005.

[16] D. Ramos-Castro, D. Garcia-Romero, I. Lopez-Moreno, and J. Gonzalez-Rodriguez, "Speaker verification using fast adaptive Tnorm based on Kullback-Leibler divergence," in *Proc. of 3rd COST 275 Workshop.*, 2005, pp. 49–52.

[17] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, J. Zieba-Palus, and C. G. G. Aitken, "Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation," in *Proceedings of International Workshop on Computational Forensics*, 2007, pp. 411–416.

[18] D. Ramos, "Evaluation of likelihood ratios based on information theory," 2007, Invited presentation at One Day One Topic Seminar and Workshop - Forensic Evidence Evaluation. Institute of Forensic Research, Cracow, Poland. 22nd-23rd June 2007.

[19] D. Ramos and J. Gonzalez-Rodriguez, "Information-theoretical evaluation of likelihood ratios," *Problems of Forensic Sciences*, 2007.

[20] D. Ramos and J. Gonzalez-Rodriguez, "Cross-entropy analysis of the information in forensic speaker recognition," in *Proc. of Odyssey*, Stellenbosch, South Africa, 2008.

[21] D. Ramos-Castro, J. Gonzalez-Rodriguez, A. Montero-Asenjo, and J. Ortega-Garcia, "Suspect-adapted MAP estimation of within-source distributions in generative likelihood ratio estimation," in *Proc. of Odyssey*, 2006.

[22] J. Gonzalez-Rodriguez, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "On robust estimation of likelihood ratios: The ATVS-UPM system at 2003 NFI/TNO forensic evaluation," in *Proc. of Odyssey*, 2004, pp. 83–90.

[23] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Forensic identification reporting using automatic speaker recognition systems," in *Proc. of ICASSP*, 2003, vol. 2, pp. 93–96.

[24] J. Gonzalez-Rodriguez, D. Ramos-Castro, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Analisis forense de voces dubitadas en la metodologia bayesiana," in *Actas del II congreso de la sociedad espaola de acustica forense (SEAF), Barcelona, Espaa*, 2003.

[25] J. Gonzalez-Rodriguez, D. Ramos-Castro et al., "On the use of high-level information for speaker recognition: the ATVS-UAM system at NIST SRE 2005," *IEEE Aerospace and Electronic Systems Magazine*, pp. 15–21, 2007.

[26] I. Lopez-Moreno, I. Mateos-Garcia, D. Ramos, and J. Gonzalez-Rodriguez, "Support vector regression for speaker verification," in *Proc. of Interspeech*, 2007, pp. 306–309.

[27] D. Ramos-Castro, J. Gonzalez-Rodriguez, C. Champod, J. Fierrez-Aguilar, and J. Ortega-Garcia, "Between-source modelling for likelihood ratio computation in forensic biometric recognition," in *Proc. of AVBPA*, 2005, vol. Springer LNCS-3546, pp. 1080–1089.

[28] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Reconocimiento de locutor multinivel y aplicaciones forenses. el sistema ATVS en NIST SRE 2005," in *Actas del III congreso de la sociedad espaola de acustica forense (SEAF), Santiago de Compostela, Espaa*, 2005.

[29] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, and J. Ortega-Garcia, "Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems," *Forensic Science International*, vol. 155, no. 2-3, pp. 126–140, 2005.

Speed Up Strategies for the Creation of Multimodal and Multilingual Dialogue Systems

Luis Fernando D'Haro, Ricardo de Cordoba

Speech Technology Group. Dept. of Electronic Engineering. Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040-Madrid, Spain
{lfdharo, cordoba}@die.upm.es

Abstract

In this paper we will summarize the work done in the PhD thesis that follows the same title as this document. In the thesis we propose different innovative, dynamic, and intelligent acceleration strategies applied to a development platform for reducing the design time of multimodal and multilingual dialogue systems and to improve the runtime modules. Throughout the paper we will describe the three different kinds of accelerations proposed, which are innovative with respect to current commercial and research platforms. The first kind of strategies was applied to the design platform in order to allow the prediction of the information required to complete the different aspects of the service. These strategies are mainly based on using the data model structure and database contents, as well as cumulative information obtained from the previous and sequential steps in the design. Thanks to them, the design is reduced, most of the times, to simple confirmations from the designer to the "proposals" that the platform automatically provides. The second kind of strategies is the incorporation of a new adaptation algorithm to the language models used by a machine translation system that automatically translates system's prompts (in audio or text) into an animated sequence in Sign Language for providing the designed service to deaf users using an avatar. Finally, the third kind is an innovative LID technique based on using a discriminative ranking of n-grams that allows the incorporation of contextual longer-span information into the language models used to identify the system needs to use to interact with the users of the service.

Index Terms: Dialogue Systems, Language Identification, Language Model Adaptation, Machine Translation.

1. Introduction

Currently, the growing demand of automatic dialogue services for different domains, user profiles, and languages has led to the development of a large number of sophisticated commercial and research platforms that provide all the necessary components for designing, executing, deploying and maintaining such services with minimum effort and with innovative functions that make them interesting for developers and final users. In general, both commercial and research platforms rely more or less in the same kind of acceleration strategies. For instance, the incorporation of state-of-the-art modules such as language identification, speech recognizers and synthesizers, etc., user-friendly graphical interfaces, inclusion of built-in libraries for typical dialogues, or additional assistants for debugging the service, as well as support for widespread standards such as VoiceXML or CCXML in order to increase the portability and reduce costs. However, surprisingly these platforms do not include any kind of acceleration strategies based on the contents or in the structure of the backend database that, as we will show, can provide important information to the design. The results of both a subjective as an objective evaluation demonstrate the usefulness and high acceptance of the proposed accelerations, while showing that the design time can be reduced on average by more than 56% when compared to a system without them.

Regarding the Language Identification system (LID), in the thesis we have focused on searching solutions for increasing the classification rates of detecting the user's language in the real-time system. In the thesis we proposed to use a discriminative ranking of occurrences of each n-gram with higher n-grams as language model. Our proposed ranking overcomes the state-of-the-art technique Parallel phone recognition followed by language modeling (PPRLM) (13% relative improvement) due to the inclusion of 4-gram and 5-gram in the classifier. Besides, our technique was also combined with acoustic information obtaining better results.

Finally, regarding the machine translation system, it is notorious the significant advances that these systems have reached in the last years, making it possible to face new challenges such as speech-to-speech or speech-to-sign-language translation. The later is especially useful to help deaf people to communicate with hearing people since many of them have problems when reading lips or written texts as they are used to the sign language grammar [1]. In this way, any kind of dialogue service that we can develop for hearing users has to be adapted to deaf users by using an avatar that gestures the system prompts on sign-language. On the other hand, it is well know that an efficient training of any statistical machine translation system requires a big parallel corpus in order to obtain reliable language and translation models. In the thesis we explored solutions for improving the language models (LMs) used to ensure correct grammatical sentences during the translation process. Our technique is based on adapting the original n-gram counts on the target side, using "translated" n-gram counts from the source side retrieved from the web. Our results show relative translation error reductions close to half the maximum performance obtainable when only the LM is optimized (i.e. without optimizing the translation model).

2. Description of the Accelerations

In this section we will describe briefly the main accelerations included in the platform. For further details please refer to the full thesis document[1] or corresponding papers at each section.

2.1. Accelerations to the Dialogue Design

The development platform that we have used in the thesis was the result of the European project GEMINI. The platform consists of three main layers integrated into a common graphical user interface (GUI) that guides the designer step-by-step and lets him go back and forth. In the first layer, the designer specifies global aspects related to the service, the data model structure, and the runtime functions to access the backend database. The next layer includes an assistant to define the dialogue flow at an abstract level by specifying the high-level states of the dialogue, plus the slots to ask to the user and the transitions among states, as well an assistant to

[1] http://www-gth.die.upm.es/~lfdharo/index_en.php?status=publications

define, in detail, all the actions to be done in each state (e.g., variables, loops, if-conditions, math or string operations, conditions for making transitions between states, calls to dialogs to provide/obtain information to/from the user). Finally, the third layer contains the assistants that complete the general flow specifying for each dialogue the details that are modality and language dependent. For instance, the prompts and grammars, the definition of user profiles, the error recovery logic for speech or Internet access, the presentation of information on screen or using speech, etc. Finally, the VoiceXML and xHTML scripts used by the real-time system are automatically generated in this layer too.

2.1.1. Accelerations to define the Data Model and Database access

One of the first steps in the design is the definition of the data model and the functions for accessing the backend database at runtime. Regarding the data model, the designer defines it through a visual and object-oriented representation (using classes and attributes) that provides information to following assistants about which fields in the database are relevant for the service (i.e. to provide or request information to/from the users), as well as the relationships between tables and fields.

During this step the assistant extracts information from the database contents such as the name and number of the tables, fields, and records. In addition, the following heuristic information for each field is calculated: a) field type, b) average length, c) number of empty records, d) language dependent fields, and e) proportion of records that are different. This information is used later to simplify the design or to improve the presentation of information in posterior assistants. For instance, we use them to propose which slots can be unify in order to be requested at the same time to the user, for creating automatic dialogue proposals, or to sort by relevance the information displayed in the assistants.

Finally, we have also incorporated an innovative acceleration strategy that simplifies the process of creating the prototypes (API) of the database access functions used by the runtime system, this way reducing the necessity of learning SQL and simplifying the process of adding the query into the real-time modules and scripts. The wizard semi-automatically creates the SQL statement for the given prototype and provides a pre-view of the results that the system would retrieve at runtime. Currently few development platforms include such kind of assistance forcing the designer to use third party software; however, none of these platforms provide such kind of automatic query proposals.

2.1.2. Accelerations to define the Dialogue Flow

The next step in the design is to define clearly the states, data to ask the users (slots), transitions between states, and the actions that make up each state. Since this process is the most complex one, in this layer we have incorporated the most important accelerations. Below we will describe the most interesting ones. For further information about these or other accelerations please refer to [2] or [3].

The first one is that the system automatically suggests the designer when two or more slots must be requested one by one (using directed forms) or at the same time (using mixed initiative forms) according to the VoiceXML standard. The proposal is based on the heuristic information extracted from the database contents related with the corresponding slots to ask to the user and on a set of predefined, but editable, rules. This way, for instance, if we need to ask two numeric data with a proportion of different values close to one, and the total number of records of both fields is high (configurable value), then the system determines that these slots have a large vocabulary and a high probability of misrecognition, therefore it is better to ask one slot at a time (i.e. system initiative). In case there are more than two slots in a state, the system checks different slots combinations in order to find those that can be requested together and those that need to be requested alone.

Another relevant acceleration is the creation of different configurable state/dialogue and action proposals based on the information of the data model and database access functions. This way, for instance, the assistant can propose a complete state for requesting the credit and debit account numbers and the amount of money required to perform a transaction in a banking application. In addition, thanks to the heuristics data and the information from previous assistants, the system is able to propose the designer the most probable actions to be done at each state (see *Figure 1*). In the example, the assistant proposes the following actions to complete the state: a configurable template for requesting the account numbers using mixed-initiative (the account numbers correspond with short-length user-defined aliases), then the dialogue to ask for the amount, the database function to perform the transaction, and finally a built-in dialogue to notify the user with the available balance.

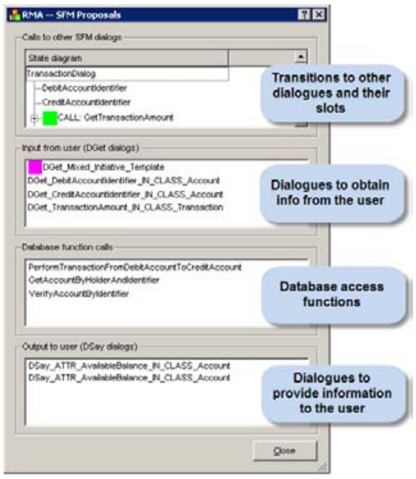


Figure 1: Example of action proposals

Another important contribution is that the platform allows the designer to create over-answering dialogues which are not currently provided by any other platform, since the VoiceXML standard only requires mixed-initiative dialogues. In order to do this, we have designed a special flow using standard elements that overcomes this limitation in the final script. In addition, the creation of these dialogues is very easy since the platform automatically proposes the slots that can be used for over-answering and automatically creates the flow.

Finally, several other accelerations are available such as a mechanism to automate the process of passing information among actions/dialogues by proposing the variables that best match the connections. This is a critical aspect of dialogue design since several actions and states have to be 'connected' as they use the information from the preceding dialogues. In addition, the platform supports the creation of different kind of dialogues, easy definition of dialogue variables, calls to other dialogs, variable assignments, a mathematical and strings assistant for including procedures, among others.

2.1.3. Other accelerations

The following step in the design is to complete the general flow specifying for each dialogue the details that are modality and language dependent. In this case, we have incorporated a wizard window that semi-automatically generates the dialogue flow for showing the lists of results after querying the database and to confirm user's answers. Another assistant allows the designer to specify the prompts and grammars used at runtime. Here, we have incorporated an assistant that helps in the creation of stochastic language models and debugging of JSGF grammars, and that automatically creates the pronunciation dictionaries used by the speech recognizer. Finally, the platform automatically generates the runtime VoiceXML script that can be run using any voice browser or using our own runtime modules. In the last case, the runtime system uses a distributed running platform similar to Galaxy [4] and the script is interpreted using OpenVXI [5].

2.1.4. Reported Results

In order to evaluate the platform and the accelerations, we carried out a subjective and objective evaluation where several developers, with different experience levels, were requested to fulfill typical design tasks covering each assistant and the proposed accelerations. For the subjective evaluation, the participants were asked to answer several questions about the platform and the strategies. The results confirm the usability of the accelerations and designer-friendliness of the platform since all of them were marked over 8.0. For the objective evaluation, we collected the following metrics: elapsed time, number of clicks, number of keystrokes, and number of keystroke errors. We compared these metrics obtained when using our assistants with a low-level accelerated editor included in the platform. The results confirm that the design time can be reduced, in average for all the assistants and tasks, in more than 45%, the number of keystrokes in 81%, and the number of clicks in 40%.

2.2. Improvements to Language Identification

Currently one of the most used technique for LID is PPRLM [6]. In this technique, the language is classified based on statistical characteristics extracted from the sequence of recognized allophones. In spite of the good results obtained by PPRLM, one of its main problems is that the accuracy is reduced due to an unreliable estimation of the LM. In order to reduce this problem, in the thesis we proposed a new algorithm for creating and using as LM a ranking of discriminative n-grams for each language to identify. Our proposed ranking resulted in a 15% relative improvement over PPRLM due to the inclusion of 4-gram and 5-gram in the classifier. Additional improvements were also obtained by including acoustic information into the GMM classifier.

In [7] the original ranking algorithm for a text-categorization task is described. In our system, we have incorporated innovative modifications to this technique. In summary the most relevant were: a) A new definition of the rank position following what we call "golf score" i.e. all n-grams that have the same number of occurrences share the same position in the rank, b) the creation of specific rankings for each n-gram order in order to avoid to take only the n-grams in the top positions that are always devoted to the unigrams, bigrams, etc., which are less discriminative, and c) the definition of a new training procedure for ranking first those n-grams that appear most in a particular language than in the others (i.e. discriminative). Finally, in the thesis we also investigated the performance of the new ranking-based system when incorporating additional information into the classifier. In detail, we tested the following features: a) Sentence acoustic score provided by the ASR, b) Phoneme acoustic score, and c) Duration for each phoneme.

2.2.1. Reported Results

Figure 2 and *Figure 3* show the cumulative results in LID error rate for the Invoca database obtained with the previous mentioned rank modifications and acoustic information. As we can see, in both cases the n-gram ranking outperforms the PPRLM system. Our final system is the integration of both systems (PPRLM and Ranking) and all the acoustic information which resulted in a significant reduction from 3.69% to 2.52% (31,7%). Further details in [8] and [9].

Figure 2: LID error rate results for the different changes in the original ranking algorithm in comparison with PPRLM

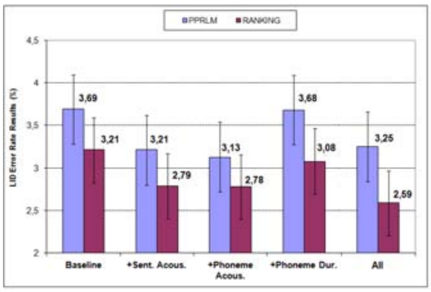


Figure 3: Comparative results between PPRLM and Ranking for adding acoustic scores into the classifier

2.3. Improvements to Machine Translation

Nowadays, most machine translation systems are trained using statistical-based algorithms that require big parallel corpora in order to guarantee a correct estimation of the translation and language models. In our case, this requirement could not be fulfilled since our target language is Sign Language (SL) and most of the currently available SL corpora are very small. For instance, [10] considers a corpus of about 2000 sentences while [11] uses a corpus of only few hundred sentences (in our case we had 266 sentences for training and 150 for test and dev). In addition, there is not any available corpus from online content as is usual in spoken languages. However, our proposed technique takes advantage of using the "source-side" language (in our case, Spanish) and from the phrase-based translation table created during the training of the MT model, in order to collect web frequency counts for the "source-side" language, using information retrieval techniques as reported in [12], and then "translating" them into "target-side" counts. The proposed technique is done in the following three steps (more details can be found in [13]):

Backward: The system uses the phrase pairs table created during the training of the translation probability $\Pr(f_1^J \mid e_1^I)$. This table consists of a list of n-gram pairs that are consistent translations between the source and target language, with their probabilities $p(\bar{f}_i|\bar{e}_i)$ and $p(\bar{e}_i|\bar{f}_i)$, and lexical weights [14]. Using this table, the system creates a list of source-side n-grams that satisfy $p(\bar{f}_i|\bar{e}_i) \geq \theta$. Here the threshold θ reduces the number of n-gram pairs to be queried in the web, so that they are more reliable. In our experiments, θ was set as a function of the number of reverse translations for $\bar{f}_i$.

Information Retrieval (IR): Using the list of previous selected n-gram, the system queries the internet to obtain web frequency counts using the Google-API[2].

[2] http://code.google.com/apis/ajaxsearch/

3.) Forward: Finally, the translation table is applied on the opposite direction to obtain the "translated" n-gram frequency counts on the target side. The conversion is done taking each n-gram pair in the list, $\bar{f}_i$, multiplying the retrieved web count, $N^{web}(\bar{f}_i)$, by the phrase translation probability, $p(\bar{e}_i|\bar{f}_i)$, and summing up all the contributions that satisfy $p(\bar{e}_i|\bar{f}_i) \geq \delta$, with $\delta = 1/n_i$, to obtain the counts for the target n-gram, $N(\bar{e}_i)$ (see Eq. 1). Then, MAP is applied to merge the counts from the original sign corpus with the converted counts. Finally, a new target LM is created form the linear interpolation of the original LM and the adapted one.

$$N(\bar{e}_i) = \frac{\sum_{\forall \bar{e}_i : p(\bar{e}_i|\bar{f}_i) \geq \delta} N^{Web}(\bar{f}_i) * p(\bar{e}_i \mid \bar{f}_i)}{\sum_{\forall \bar{e}_i : p(\bar{e}_i|\bar{f}_i) \geq \delta} p(\bar{e}_i \mid \bar{f}_i)} \qquad \textit{Eq. 1}$$

2.3.1. *Reported Results*

Table 1 shows the perplexities results provided by the baseline LMs and the adapted ones on train, dev, and test sets. The results for the test and dev sets correspond to the averaged perplexities of a three-fold cross validation test. The baseline LM is a backoff trigram with Good-Turing discount. The perplexities on both sides correspond to the adapted LMs. Values in parenthesis are relative improvements over the baseline perplexities. As we can see, the proposed technique provides a 15.5% relative improvement in the test set.

Table 1: *Perplexity results*

	Train	**Dev**	**Test**
Baseline	5.02	10.8	10.7
Adapted	3.16 (37.1%)	8.75 (18.7%)	9.04 (15.5%)

Table 2 shows the averaged MT results for text-to-sign translation on the test set. For the oracle experiment the LM is trained considering all sentences (train, development, and test sets). Since this model has all the available information, it corresponds to the top performance that it is possible to obtain only due to the LM component. As we can see, the results show that the proposed technique is able to reach approximately half (2.73%) of the maximum improvement (6.1%) in WER that it is possible to obtain when only the LM is improved (i.e. without improving the translation model).

Table 2: Machine translation results

		WER	PER	BLEU	NIST
Text-to-Sign	Baseline	34.74	29.59	0.50	6.30
	Adapted	**33.79** (2.73%)	**29.1** (1.68%)	**0.51** (2.61%)	**6.36** (1.06%)
	Oracle	32.62 (6.1%)	28.06 (5.48%)	0.55 (9.91%)	6.57 (4.23%)

3. Conclusions

In this paper, we have summarized the most important contributions of the PhD thesis that consist of different kinds of acceleration strategies applied to a complete development platform for designing and running dialogue applications.

The first kind of strategies are based on using heuristic information extracted from the backend database and on cumulative information obtained from the previous and sequential steps in the design. Our proposals include the unification of slots to be requested using mixed-initiative dialogues, the semi-automatic creation and debugging of SQL statements, as well as automatic action proposals for each dialogue. Subjective and objective evaluations confirm that the proposed strategies are useful and contribute to simplify and accelerate the design.

The second kind of strategy was applied to a language identification system that allows the dialogue system at runtime to detect the language to interact with the user. In this topic we have proposed a novel algorithm to create ranking templates with the most discriminative language-dependent n-grams which are then integrated into a state-of-the-art LID system for recognizing the user's language. The results show that this technique when unified with PPRLM and acoustic information results in a relative improvement of 31,7%.

Finally, the third kind of strategies was a LM adaptation technique successfully applied to a machine translation system that allows designers to translate automatically the prompts created for a traditional speech-based dialogue system into a visual sign language representation that can be used to offer the same dialogue service to deaf users.

4. Acknowledgements

In first place I want to thank God for His continuous direction and to my family for their support along my career. Special thanks to my advisor and members of my research group for the opportunity of doing the thesis, and to the different students that I have supervised in the development of the accelerations described above. And thanks to my friends for staying close to help me to never give up.

5. References

[1] Zhao, L., Kipper, K., et al. "A machine translation system from English to American Sign Language". AMTA, 2000. pp. 54-67.

[2] L. F. D'Haro, R. Cordoba, et al. 2006. "An advanced platform to speed up the design of multilingual dialog applications for multiple modalities". Speech Communication Vol. 48, Issue 8, pp. 863-887. July 2006.

[3] L .F. D'Haro, R. Cordoba, et al. 2009. "Speeding up the design of dialogue applications by using database contents and structure information". SigDial, pp. 160-169. London, UK.

[4] Seneff, S., Hurley, E., et al. 1998. "Galaxy-II: A reference architecture for conversational system development". ICSLP 1998, 931-934.

[5] Cordoba, R., Fernández, F., Sama, V., D'Haro, L. F., et al. 2004. "Implementation of Dialogue Applications in an Open-Source VoiceXML Platform". ICSLP 2004, pp. I-257-260.

[6] Zissman, M.A., 1996. "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech & Audio Proc., v. 4, pp. 31-44.

[7] Cavnar, W. B. and Trenkle, J. M., 1994. "N-Gram-Based Text Categorization". 3rd Symposium on Document Analysis & Information Retrieval, pp. 161-175.

[8] Cordoba, R., D'Haro, L. F., et al. 2007. "Language Identification based on n-gram Frequency Ranking". Interspeech 2007, pp. 354-357.

[9] Cordoba, R., D'Haro, L. F., et al. 2007. "Language Identification using several sources of information with a multiple-Gaussian classifier". Interspeech 2007, pp. 2137-2140.

[10] Chiu, Y.-H., Wu, C.-H., et al. 2007. "Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis", IEEE Trans. Pattern Analysis and Machine Intelligence, 29(1):28–39.

[11] Stein, D., Dreuw, P., Ney, H., et al. 2007. "Hand in hand: Automatic Sign Language to English Translation". TMI 2007, pp. 214-220.

[12] Keller, F., and Lapata, M. 2003. "Using the Web to Obtain Frequencies for Unseen Bigrams", Computational Linguistics, 29(3):459–484.

[13] D'Haro, L. F., Ney., H. et al. 2008. "Language Model Adaptation for a Speech to Sign Language Translation System Using Web Frequencies and a MAP framework". Interspeech 2008, pp. 2119-2202.

[14] Koehn, P., Och, F. J., and Marcu, D. 2003. "Statistical Phrase-Based Translation", HLT/NAACL 2003, pp. 48-54, Canada.

Albayzín'10 Evaluation: Oral Session

Albayzin 2010 Evaluation Campaign: Speaker Diarization

Martin Zelenák, Henrik Schulz, and Javier Hernando

TALP Research Center
Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya, Barcelona, Spain
{martin.zelenak,henrik.schulz,javier.hernando}@upc.edu

Abstract

In this paper we present the evaluation results for the task of speaker diarization in broadcast news domain as part of the Albayzin 2010 evaluation campaign of language and speech technologies. The evaluation data was a subset of the Catalan broadcast news database recorded from the 3/24 TV channel. Six competing systems from five different universities were submitted for the Albayzin 2010: Speaker diarization session and the lowest diarization error rate obtained was 30.4%.

Index Terms: speaker diarization, evaluation

1. Introduction

Objective evaluations became a valuable part of research and development in the field of spoken language processing. The comparison of performance of different approaches (systems) to a specific task helps setting new trends and stimulates the progress in a particular line of research. The Albayzin 2010 is the third in the series of evaluation campaigns (2006, 2008) organized by RTTH[1] and held under the FALA 2010 workshop. Largely inspired by the NIST Rich Transcription evaluations [1], the Albayzin 2010 campaign focuses among others on the task of speaker diarization of broadcast news.

Speaker diarization addresses the issue of segmenting a given audio stream according to different speakers and linking the speech regions which originate from the same person. In general, no kind of a priori speaker information is provided. In a broader sense, diarization also categorizes audio data according to music, background or channel conditions. Speaker diarization in broadcast news domain offers a strong application potential in many areas, in particular for transcription, indexing, searching and retrieval of audiovisual information.

In this paper we present an overview of the Albayzin 2010: Speaker diarization evaluation and report the results achieved by six submitted systems. The evaluation was performed on Catalan broadcast news data. Although the presented systems have several features in common (e.g. MFCCs, agglomerative clustering), there are also many differences among them (e.g. Poission-driven change rejection, online optimized processing, speaker factor analysis, dot-scoring similarity, or acoustic fingerprinting).

The rest of this paper is organized as follows. The conditions and database used for the evaluation are explained in Section 2. The participants are listed in Section 3 together with brief descriptions of their systems. The results are discussed in Section 4, followed by conclusions in Section 5.

This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first and second author are supported by a grant from the Catalan autonomous government.

[1]RTTH is the Spanish acronym for "Red Temática en Tecnologías del Habla" (the Spanish Speech Technologies Thematic Network)

2. Speaker diarization evaluation

2.1. Task and conditions

The organized evaluation campaign aims at evaluating the performance of automatic computer-based algorithms for speaker diarization, which can be also characterized as the "Who spoke when?" task. The participants could submit more than one system output, but only the primary hypothesis is considered here.

The minimum duration for a pause separating two utterances was set to 0.5 s, since pauses smaller than this value were not considered to be segmentation breaks in a speaker's speech (it is also complementary to the scoring collar discussed later).

The diarization error rate[2] (DER) defined by NIST [1] is the primary metric. DER is the ratio of incorrectly attributed speech time, (missed detections of speech, falsely detected speech, and speech assigned to the wrong speaker) to the total amount of speech time. Since there is no a priori relation between the system and reference speaker clusters, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is computed separately for each audio file. A scoring "forgiveness collar" of 0.25 s around each reference segment boundary is used. This accounts for both the inconsistent annotation of segment times by humans and the philosophical argument of when does speech begin for word-initial stop consonants.

2.2. Database

The database contains broadcast news channel recordings, i.e., announcements, reports, interviews, discussions and short statements recorded from Catalan 3/24 TV channel throughout the program. Its original video recordings were supplied by a stationary digital video broadcasting (DVB-T) receiver. Their original audio tracks were extracted being available at 32 kHz sample rate, 16 bit resolution, but were downsampled to 16 kHz sample rate.

The annotated recordings encompass a total duration of 88 hours, but for the Albayzin 2010 speaker diarization evaluation a subset of 8 recording totaling approximately 30 hours was selected. Although TV3 is primarily a Catalan television channel, the recorded broadcasts contain a proportion of roughly $\frac{1}{6}$ of Spanish speech segments.

Catalan (mainly spoken in Catalonia) exhibits substantial dialectical differences, dividing the language into an eastern and

[2]NIST scoring tool available at: http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl

western group. The majority of recorded Catalan speakers features the central Catalan dialect being part of the eastern dialect group [2].

A first annotation pass segmented the recordings with respect to background sounds, channel conditions, and speakers as well as speaking modes. Table 1 shows the speaker distribution. Since segments of overlapping speakers did not receive a gender tag, they form also a subset of the "unknown" gender account. The gender conditioned distribution indicates a clear misbalance in favor of male speech data. The number of speakers per recording ranges from 30 to 250. Some speakers appear in several recordings (newscaster, journalists), however, the majority of the speakers account to only a short duration, since they are connected to particular news.

Table 1: *Distribution of speakers*

Gender	# Speakers	Duration [h]	# Segments
male	1239	44:23:41	12869
female	507	25:43:54	7559
unknown	270	07:50:38	2579
overlapped	68	00:12:38	241

Besides the tabulated total durations of audio segments of specific conditions in Table 2, there are a few conditions featuring an overlap of all noted background sounds with minor duration. Few segments are indicated to originate from telephone speech. The recorded speech within these segments can be considered band-limited to frequencies from 300 Hz to 3.4 kHz.

A second annotation pass provided literal transcriptions and acoustic events of segments that feature planned and spontaneous speech, but no long term background noises. The non-speech acoustic events were furthermore tagged with time stamps indicating their beginning and end.

Because of the fact that silences were not manually annotated, the transcriptions were extended by passing the signal through the hierarchical audio segmentation described in [3]. This involved a simple low-energy silence detector to estimate regions with non-speech (silence). Furthermore, to avoid too short segments, a smoothing constraining the minimal non-speech duration to 0.5 s was applied.

Table 2: *Duration breakdown regarding recording environment and background conditions of speech segments (number of segments in parenthesis)*

Channel	Background [h]			
	None	Speech	Music	Noise
None	04:27:10 (2451)	00:18:54 (131)	04:36:06 (1945)	01:15:30 (1113)
Studio	15:04:24 (4752)	01:36:16 (594)	08:40:47 (1407)	00:57:12 (2067)
Telephone	00:00:40 (11)	00:00:10 (2)		00:06:47 (10)
Outside	14:49:44 (6558)	03:55:29 (1319)	01:52:52 (557)	18:55:19 (4342)

Table 3: *Participating teams in the Albayzin 2010: Speaker diarization section*

Team ID	Research institution
AhoLab	University of the Basque Country (EHU)
GSI	University of Coimbra (UC)
GTM	University of Vigo (UVigo)
GTC-VIVOLAB	University of Zaragoza (UZ)
GTTS	University of the Basque Country (EHU)
ATVS-UAM	Autonomous University of Madrid (UAM)

3. Evaluation participants

3.1. Teams

Six teams from five universities submitted their systems to the Albayzin 2010 speaker diarization evaluation. The list of participants is given in Table 3.

3.2. System descriptions

Several teams participated also in the Albayzin 2010: Audio segmentation section, where five acoustic classes were defined to segment the audio data [4]. The classes were as follows: music, clean speech, speech with music, speech with noise and other (e.g. noise, silence). Since audio segmentation normally constitutes a part of speaker diarization systems, we are referring in latter system descriptions to these five acoustic classes.

3.2.1. AhoLab system

The system from Aholab team was built to run online and thus the whole process is performed in a single iteration. A more detailed description of the selected algorithms and modifications is given in [5]. The speech activity detection (SAD) is based on Viterbi segmentation of the audio signal into five acoustic classes. Each class is modeled with a Gaussian mixture model (GMM) and signal parameterization involves MFCCs with first and second derivatives.

For speaker change detection, growing window architecture and the Bayesian information criteria (BIC) metric is applied. Though the growing window has higher computational cost, the authors report its better performance compared to fixed-size sliding window approach and implemented a number of adjustments in order to decrease the computation time. At this stage of the process, only MFCC features with no feature derivatives are used. Furthermore, only voiced frames are included in the speaker change detection.

During the online clustering algorithm, every time a speaker change is detected, the BIC value of the recent speech segment against all known clusters is computed. If the lowest BIC value falls below a certain threshold the segment is assigned to the given cluster. Otherwise, a new cluster is created.

3.2.2. GSI system

The diarization system proposed by team GSI [6] includes an audio segmentation system to determine speaker turns and discard non-speech segments like silence and music. It uses a set of 16 MFCCs, 8 other features (e.g. energy, zero-crossing rate, spectral measures) and their derivatives. Segmentation is based on a hybrid ANN/HMM Viterbi decoder and discriminates between five acoustic classes.

To classify speakers, the algorithm begins with training a background GMM with data of the entire audio file. Then, a decoder that outputs the most probable mixture sequence is used (with high mixture transition penalization) to detect speaker turns. Homogeneous segments with speech of only one speaker tend to produce sequences with few mixtures turns.

Two passes of verification are then applied to the labeled speaker segments to test whether every pair of segments is homogeneous or not. The first pass involves an audio fingerprint system and the other is based on BIC. If two segments are classified as similar, then the corresponding speaker labels are equated.

Acoustic or audio fingerprinting refers to a condensed representation of an audio signal that can be used to identify an audio sample or quickly locate similar items in audio streams. A binary representation of spectral patterns computed by the convolution of spectrogram with a mask is used. This technique is convenient to discover repeated segments with high confidence. Labels are determined according to a majority voting scheme in order to deal with classification inconsistencies in repeated segments.

3.2.3. GTM system

The GTM system [7] starts by making a coarse segmentation with the distance changing trend segmentation (DCTS) algorithm. Then, a refinement or rejection of detected audio change-points by an adaptive threshold-based BIC algorithm follows in order to reduce the false alarm rate. The change-point rejection approach assumes that the occurrence times of change-points can be modeled by a Poisson process (cumulative density function). Initially, a change is accepted with a very high probability, but as the number of accepted changes increases and is close or over the expected number, they are more likely to be rejected.

After this segmentation stage, the system successively decides whether a particular segment is speech, whether the speech is male or female, and, based on the cross likelihood ratio (CLR) test, whether the two latest speech segments are spoken by the same speaker. In that case both speech segments are merged.

Finally, an agglomerative hierarchical clustering step is performed to classify the speech segments by speaker identity. Similarity between speech segments is evaluated with a cosine distance measure which uses information about the likelihood score. Specifically, each speech segment is characterized with a collection of scores against a set of GMMs adapted for every segment from an universal background model (UBM).

The audio signal is characterized by 12 MFCCs augmented with the log-energy. The speech/non-speech and gender classification modules also consider the first and second derivatives.

3.2.4. GTC-VIVOLAB system

The speaker diarization systems submitted by the GTC-VIVOLAB team for the Albayzin 2010 speaker diarization evaluation [8] combines recent improvements in the field of speaker segmentation of two-speaker telephone conversations, using eigenvoice modeling, with the traditional BIC-based agglomerative hierarchical clustering approach.

The JFA-based (JFA stands for joint factor analysis) speaker segmentation system works with a given number of speakers (since it was designed for two-speaker dialogues). Because of that, after running speech activity detection, every recording is split into 5 minute slices and every slice is processed separately. The segmentation system is forced to find 10 speakers in every slice.

Once there are 10 clusters for every 5-minute slice, clustering over the whole recording is performed to merge those clusters belonging to the same speakers. For this purpose, BIC is considered as both a clustering metric and a stopping criterion. Clusters are modeled with a single full-covariance Gaussian function using 18 MFCCs.

3.2.5. GTTS system

The GTTS system detailed in [9] consists of three decoupled elements: speech/non-speech segmentation, acoustic change detection and clustering of speech segments. All of them rely on 13 MFCC features, which are augmented for clustering with first and second-order deltas.

Speech/non-speech segmentation is based on an ergodic continuous HMM with 5 states (one per acoustic class). With the aim to detect speaker changes, speech segments are further segmented by means of a naive XBIC-metric-based approach, which locates the most likely spectral change points. The authors state that almost all the speaker changes and many other additional changes were detected.

The third element is based on a dot-scoring speaker verification system, where speech segments are represented by MAP-adapted GMM zero- and first-order statistics. The dot scoring is then applied to compute a similarity measure between segments (or clusters) and finally an agglomerative clustering algorithm is used until no pair of clusters exceeds a similarity threshold.

3.2.6. ATVS-UAM system

The front-end parameterization of the ATVS-UAM speaker diarization involves the extraction of 19 MFCCs concatenated to their deltas, followed by cepstral mean normalization (CMN), RASTA filtering and feature warping. All speech data detected by a preceding audio segmentation step is used to train an UBM. Given this UBM, sufficient statistics are extracted for every segment. The next steps involve a factor analysis to model the total variability subspace resulting in so-called iVectors.

The MFCC feature stream is divided into 90-second audio slices. Compensated iVectors in each slice are clusterized based on their cosine distance. Cluster centroids are representing candidate speakers. Candidate speaker models are accumulated over all the slices in the test session together with the frequency of appearance of their clusters.

Speakers are expected to appear in several slices and thus a secondary clustering is used to merge the initial centroids, obtaining an enhanced set of candidate speakers. A prior probability is assigned to each of the candidate speakers according to its presence in the entire session. Likelihoods for each candidate speakers are estimated in a second pass over the iVector stream using the cosine distance and the prior probability of each candidate speaker. The final diarization labels are obtained with a Viterbi decoding of these scores. A more detailed description of the system can be found in [10].

4. Results

The DER results for six submitted systems in Albayzin 2010 are given in Table 4. In addition, the DER composition is also depicted in Figure 1. The best result of 30.4% DER was obtained by the AhoLab system, followed by similar performances of GTTS, GTC-VIVOLAB and ATVS-UAM systems. The performance rankings are closed with the DERs of GSI and GTM teams.

Note, that the most significant portion of DER is caused

Table 4: *Speaker diarization results for all participants in terms of Missed speech rate (*MS*), False alarm speech rate (*FA*), Speaker error rate (*SPKE*) and Diarization error rate (*DER*). All values are in given in (%).*

Team	MS	FA	SPKE	DER
AhoLab (EHU)	4.9	1.5	23.9	**30.4**
GSI (UC)	1.1	2.3	52.4	**55.8**
GTM (UVigo)	8.8	4.1	45.1	**58.0**
GTC-VIVOLAB (UZ)	3.7	1.5	28.6	**33.8**
GTTS (EHU)	2.2	2.2	28.8	**33.2**
ATVS-UAM	1.1	10.8	22.9	**34.7**

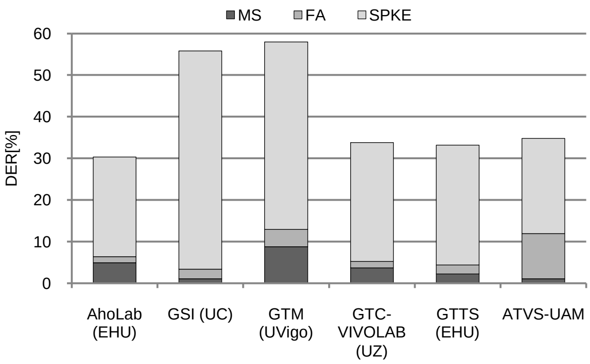


Figure 1: DER *distribution of missed-speech detections (*MS*), false-alarm detections (*FA*) and speaker error (*SPKE*).*

by incorrectly assigned speaker labels. This is very likely due to the high number of speakers in the evaluation corpus and variable background conditions. Lowest speaker error was achieved by the ATVS-UAM system with Viterbi decoding of iVector-stream scores over candidate clusters. Interesting question would be the impact of the score normalization according to cluster appearance probability on the error rates. Noteworthy is also the speaker error achieved by AhoLab, where the clustering happens in only a single iteration. The lowest error accounting to speech/non-speech detection produced the GSI system with a hybrid ANN/HMM approach.

The operation of the systems in terms of detected speaker count is shown in Figure 2. Here, the ATVS-UAM and GTTS systems exhibit the highest number of true detected speakers, but at the same time suffer from even higher counts of false speakers. The AhoLab system for instance, though detecting less correct speakers, maintains a significantly lower number of false speakers. Similarly the GTC-VIVOLAB system.

5. Conclusions

The Albayzin 2010 speaker diarization evaluation results were presented for six teams from four Spanish (EHU, UVigo, UZ, UAM) and one Portuguese (UC) university. The system which obtained the best result was also designed to run online and relies on modified growing-window BIC-based speaker-change detection and on a BIC-based clustering algorithm.

The evaluation data turned out to be relatively challenging, since the DER results in other comparable evaluations, e.g., the NIST RT'04 evaluation [11] or the ESTER evaluation on French broadcast news [12], were considerably lower than in this case. The high number of speakers in Catalan TV 3/24

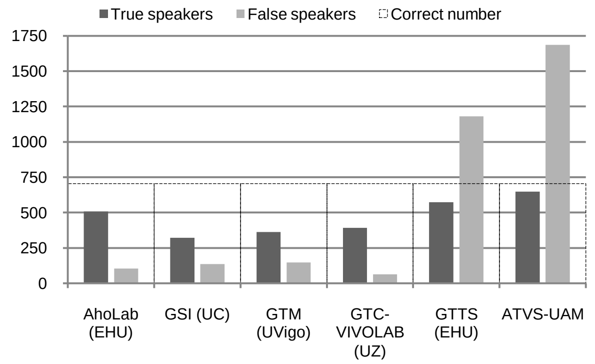


Figure 2: *Correctly detected (True) and falsely introduced (False) number of speakers by evaluated systems.*

broadcast news corpus was perhaps also the reason why no system managed to determine the correct speaker count in neither recording.

6. Acknowledgements

The authors would like to thank the evaluation organizers for their effort and also the participants for the help with the system descriptions.

7. References

[1] NIST. (2009) The NIST Rich Transcription evaluation project website. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/

[2] M. W. Wheeler, *The Phonology of Catalan*. Oxford, UK: Oxford University Press, 2005.

[3] M. Aguilo, T. Butko, A. Temko, and C. Nadeu, "A hierarchical architecture for audio segmentation in a broadcast news task," pp. 17–20, September 2009.

[4] T. Butko and C. Nadeu, "A Hierarchical Architecture with Feature Selection for Audio Segmentation in a Broadcast News Domain," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[5] I. Luengo, E. Navas, I. Saratxaga, I. Hernáez, and D. Erro, "AhoLab Speaker Diarisation System for Albayzin 2010," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[6] A. Veiga, C. Lopes, and F. Perdigão, "Speaker Diarization Using Gaussian Mixture Turns and Segment Matching," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[7] L. Docio-Fernandez, P. Lopez-Otero, and C. Garcia-Mateo, "The UVigo-GTM Speaker Diarization System for the Albayzin'10 Evaluation," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[8] C. Vaquero, A. Ortega, and E. Lleida, "VIVOLAB-UZ Speaker Diarization System for the Albayzin 2010 Evaluation Champaign," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[9] M. Diez, M. Penagarikano, A. Varona, L. J. Rodriguez-Fuentes, and G. Bordel, "GTTS System for the Albayzin 2010 Speaker Diarization Evaluation," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[10] J. Franco-Pedroso, I. Lopez-Moreno, D. T. Toledano, and J. Gonzalez-Rodriguez, "ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation," in *Proc. FALA 2010*, Vigo, Spain, 2010.

[11] J. Fiscus, A. Le, and G. Sanders. (2004) MDE Tasks and Results. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/rt04f-mde-nist.pdf

[12] S. Galliano *et al.*, "The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proc. Interspeech '05*, Lisbon, Portugal, 2005, pp. 1149–1152.

Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results

Taras Butko, Climent Nadeu and Henrik Schulz

TALP Research Center, Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya, Barcelona, Spain

taras.butko@upc.edu, climent.nadeu@upc.edu, henrik.schulz@upc.edu

Abstract

In this paper, we present the audio segmentation task from the Albayzín-2010 evaluation, and the results obtained by the eight participants from Spanish and Portuguese universities. The evaluation task consisted of the segmentation of audio files from the Catalan 3/24 TV channel into 5 acoustic classes: music, speech, speech over music, speech over noise and other. The final results from all participants show that the problem of segmenting broadcast news is still challenging. We also present an analysis of the segmentation errors of the submitted systems. Additionally, the evaluation setup, including the database and the segmentation metric, is also described.

Index Terms: audio segmentation, broadcast news, international evaluation, evaluation setup, database

1. Introduction

The Albayzín evaluation campaign is an internationally-open set of evaluations organized by the Spanish network of speech technologies every 2 years. In the context of Albayzín-2010, an audio segmentation task was proposed by the authors for the first time. It is motivated by the fast increase of audio data, which demands for efficient content-based automatic audio segmentation methods. Recently, researchers have put much effort on this problem due to its applications to tasks as audio indexing and retrieval [1], or automatic transcription of audio recordings [2]. Also, a previous identification of speech segments facilitates speech processing tasks as speech recognition or speaker diarization. Furthermore, audio segmentation is used to make online adaptation of ASR models, or to generate a set of acoustic cues for speech recognition in order to improve the overall system performance [1]. Additionally, multimedia surveillance and monitoring applications can benefit significantly from audio-based event detection [3].

Many research works address the problem of audio segmentation in different scenarios. In [4], the authors propose a method for robust speech, music, environment noise and silence segmentation of audio recorded in different conditions such as TV studio, telephone etc. In [5], the audio stream from broadcast news domain is segmented into 5 different classes, including speech, commercials, environmental sound, physical violence and silence. Content-based retrieval from TV programs is considered in [6], where 7 similar classes are defined.

The final results from 8 participants as well as the evaluation setup, including the database (which is freely available) and the segmentation metric, are described in this paper.

2. Albayzin 2010 audio segmentation evaluation

2.1. The database

The database used for evaluations consists of a Catalan broadcast news database from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was manually annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project. The database includes around 87 hours of annotated audio (24 files of approximately 4 hours long).[1]

The manual annotation of the database was performed in 2 passes. A first annotation pass segmented the recordings with respect to background sounds (speech, music, noise or none), channel conditions (studio, telephone, outside and none), and speakers as well as speaking modes. A second annotation pass provided literal transcriptions and acoustic events of segments (such as throat, breath, voice, laugh, artic, pause, sound, rustle or noise). For the proposed evaluation we took into account only the first pass of annotation. According to this material, five different audio classes were defined (Table 1).

Table 1: *The five acoustic classes defined for evaluation.*

Class	*Description*
Speech [sp]	Clean speech in studio from a close microphone
Music [mu]	Music is understood in a general sense
Speech over music [sm]	Overlapping of speech and music classes or speech with noise in background and music classes
Speech over noise [sn]	Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation)
Other [ot]*	This class refers to any type of audio signal (including noises) that doesn't correspond to the other four classes

* Not evaluated in final tests

The distribution of the classes within the database is the following: Clean speech: 37%; Music: 5%; Speech over music: 15%; Speech over noise: 40%; Other: 3%. Although 3/24 TV is primarily a Catalan television channel, the recorded broadcasts contain a proportion of roughly 17% of Spanish speech segments. The gender conditioned distribution indicates a clear unbalance in favor of male speech data (63% versus 37%).

[1] The Corporació Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development.

The database for evaluation was splitted into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

2.2. Metric

The metric is defined as a relative error averaged over all acoustic classes (ACs):

$$Error = \underset{i}{average}\left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}\right) \qquad (1)$$

where

$dur(miss_i)$ is the total duration of all deletion errors (misses) for the i*th* AC,

$dur(fa_i)$ is the total duration of all insertion errors (false alarms) for the i*th* AC, and

$dur(ref_i)$ is the total duration of all the i*th* AC instances according to the reference file.

The incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A forgiveness collar of 1 sec (both + and -) is not scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an AC begins/ends.

The proposed metric is slightly different from the conventional NIST metric for speaker diarization, where only the total error time is taken into account independently of the acoustic class. Since the distribution of the classes in the database is not uniform, the errors from different classes are weighed differently (depending on the total duration of the class in the database). This way we stimulate the participants to detect well not only the best-represented classes (“speech” and “speech over noise”, 77% of total duration), but also the minor classes (like music, 5%).

2.3. Evaluation organization

Ten research groups registered for participation, but only eight submitted segmentation results: **GTTS** (Universidad del País Basco), **GTC-**VIVOLAB (Universidad de Zaragoza), **GSI** (Instituto de Telecomunicações, Universidade de Coimbra, Portugal), **TALP** (Universitat Politècnica de Catalunya) **CEPHIS** (Universitat Autònoma de Barcelona), **ATVS** (Universidad Autónoma de Madrid), **GTM** (Universidade de Vigo), **GTH** (Universidad Politécnica de Madrid / Universidad Carlos III de Madrid).

The database was splitted into 2 parts: for training/development (2/3 of the total amount of data, 16 sessions), and testing (the remaining 1/3, 8 sessions). The training/development audio data together with ground truth labels and evaluation tool were distributed among all the participants by the date of release.

About 3 months were given to all the participants to design their own audio segmentation system. After that period, the testing data was released and 2 weeks were given to perform testing.

Though the evaluation was carried out with the outputs from the primary system submitted by each participant, each site could also submit a contrast (alternative) system. Each evaluated system had to be applied to the whole test database. Each participant site was asked to provide also the total time required to run the set of tests for each submitted system (specifying the used computational resources). The evaluated systems could only use audio signals. Any publicly available data was allowed to be used together with the provided data to train the audio segmentation system. When additional training material is used, the participant was obliged to provide the reference regarding it. Indeed, listening to the test data, or any other human interaction with data, was not allowed before all test results had been submitted.

3. Final results

Table 2 presents the final evaluation scores from the eight participants.

Table 2. *Results of the audio segmentation evaluation.*

Participant	Error rate				
	mu	sp	sm	sn	**Average**
GTH	19.21	39.52	24.97	37.19	**30.22**
GTM	22.41	41.80	27.47	40.93	**33.15**
ATVS	31.01	40.42	33.39	39.80	**36.15**
TALP	26.40	44.20	33.88	41.52	**36.50**
CEPHIS	23.65	45.07	36.95	45.21	**37.72**
GSI	21.43	48.03	51.66	48.49	**42.40**
GTC-VIVOLAB	28.14	51.06	48.78	51.51	**44.87**
GTTS	26.94	52.76	47.75	52.93	**45.09**

Note that the winner of the evaluation (GTH), which obtained the highest average score, also got the highest scores individually for each class. According to the presented results, the “Music” class is the easiest for detection, while “speech” and “speech over noise” are the most difficult.

The distribution of miss and false alarm errors for all participants is presented in Figure 1. According to it, these two types of errors are balanced for “music” and “speech over noise” class, while for “speech” class the false alarm errors are more dominant, and for “speech over music” class the dominant errors are misses.

In Table 3 we present the confusion matrix, which shows the percentage of hypothesized AEs (rows) that are associated to the reference AEs (columns). Data represent averages across the eight audio segmentation systems.

Table 3. *The confusion matrix of acoustic classe*s.

	mu	sp	sm	sn
mu	89.4	0.1	8.0	2.5
sp	0.0	70.6	2.9	26.5
sm	1.8	1.2	87.0	10.0
sn	0.3	10.2	8.3	81.2

According to that confusion matrix, the most common errors are confusions between “Music” and “Speech over music”, between “Speech over music” and “Speech over noise” and also between “Speech” and “Speech over noise” classes. Indeed, these classes have very similar acoustic content. Another interesting observation is the low proportion (almost 0%) of confusions between “Speech” and “Music” classes.

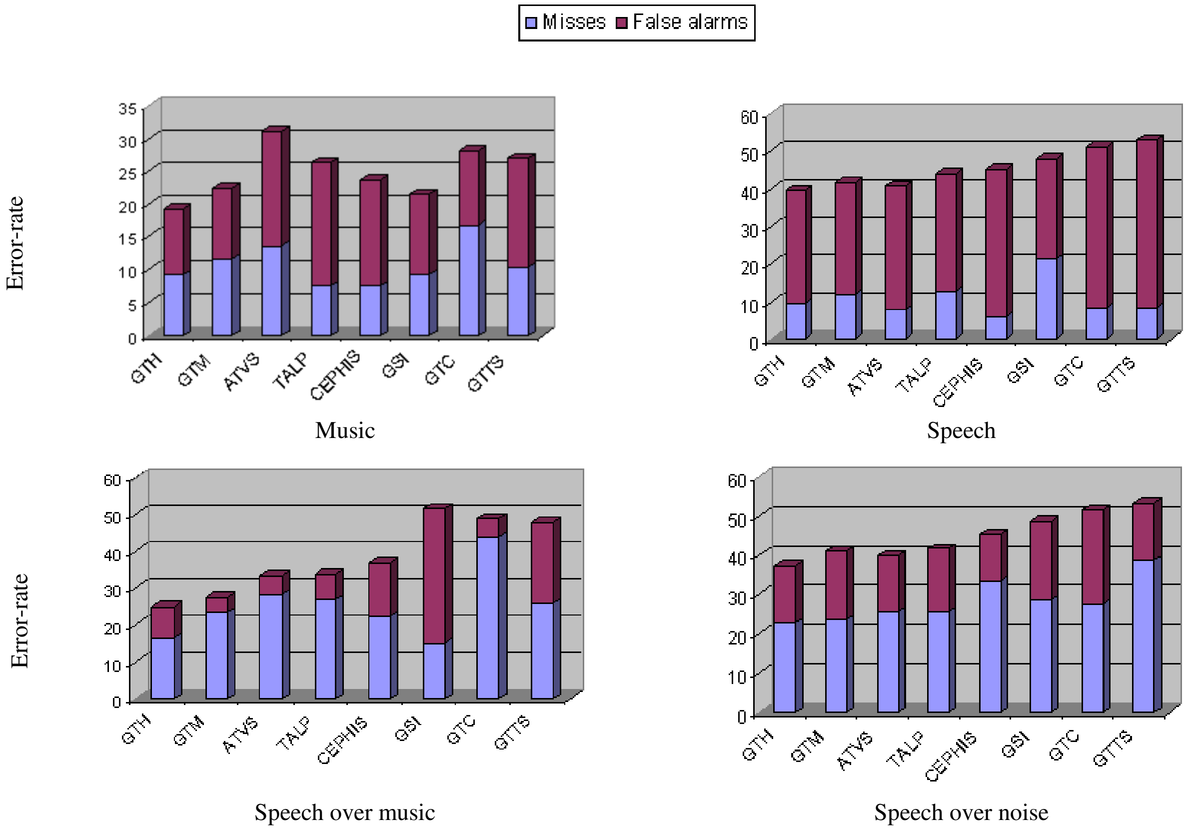


Figure 1: *Distribution of errors for the eight participants, and for each acoustic class.*

In Figure 2 we present the cumulative distribution of errors in terms of duration. Each point *(x, y)* of this plot shows the percentage *y* of total amount of errors with duration less than *x* seconds.

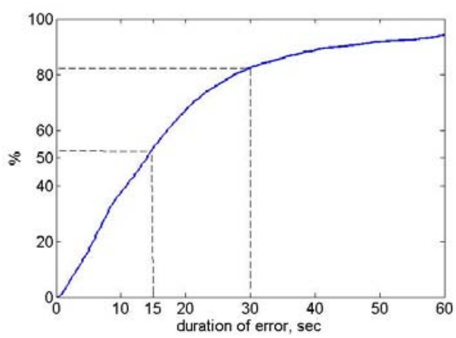


Figure 2: *The cumulative distribution of errors in terms of duration*.

According to that plot, more than half of the errors are shorter than 15 sec and more than 80% of the errors are shorter than 30 sec. In other words, almost each long segment is detected correctly.

In order to measure the difficulty of the proposed audio segmentation task, in Figure 3 we display the proportion of 3 different types of segments in the testing database: *very difficult*, *difficult* and *from winner*. *Very difficult* are those segments on which all 8 audio segmentation systems produced errors (misses or false alarms). *Difficult* segments are those where 7 out of 8 systems produced errors. Finally, *from winner* are those segments with errors produced by the winner system.

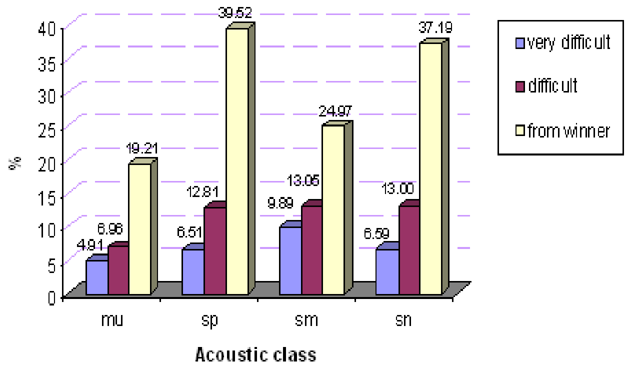


Figure 3: *Comparison of different types of segments in terms of segmentation difficulty.*

According to that plot, only 6.22% of segments are *very difficult*, while the rest of the segments were labeled correctly at least by one participant.

Table 4 and Figure 4 show a grouping of the errors which are common to all the 8 participant segmentation systems. The groups were defined after listening to all the error segments which are *very difficult* and longer than 5 seconds. Seven different semantic groups were distinguished, and the rest were included in *Other*.

Table 4. *Different semantic types of errors which are common to all eigth systems.*

Type of error	*Description*
Type 1	Low level of background sound
Type 2	Speech in background
Type 3	Annotation error
Type 4	The microphone is affected by the wind
Type 5	Singing in background
Type 6	Noise in background is more dominant than music for the [sm] class
Type 7	The quality of music in background is low
Type 8	Other

The percentages of distribution of the above mentioned errors are depicted in Figure 4.

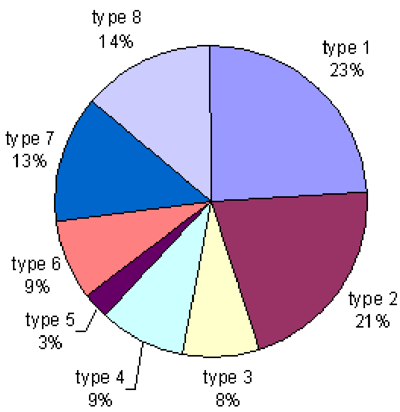


Figure 4: *Percentages of distribution of the different types of common errors.*

According to the plot, a large percentage of common errors are provoked by the presence of either a low level of sound in the background (23%) or overlapped speech (21%), while the annotators' mistakes cause only 8% of the total amount of common errors. Thus the audio segmentation task is still challenging.

4. Conclusions

In this paper, we have presented the submitted results of the Albayzín-2010 audio segmentation evaluation and the evaluation setup, including the database and the segmentation metric.

8 participants from different universities of Spain and Portugal submitted the evaluation results, and the winner got 30.22 % of error-rate in terms of proposed metric. By analyzing the submitted results from all participants we conclude that most of the errors are confusions between "Music" and "Speech over music", "Speech over music" and "Speech over noise", and also between "Speech" and "Speech over noise". Besides, more than half of the total amount of errors is shorter than 15 seconds.

By analyzing the semantic content of errors produced by all submitted systems we found that most of the errors are provoked by the presence of either a low level of sound in the background (23%) or overlapped speech (21%), while the annotators' mistakes cause only 8% of the total amount of common errors.

5. Acknowledgements

The authors wish to thank our colleagues at **GTTS**, **GTC-VIVOLAB**, **GSI**, **CEPHIS**, **ATVS**, **GTM**, **GTH** for their participation in the evaluation. The UPC's work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is partially supported by a grant from the Catalan autonomous government.

6. References

[1] H. Meinedo, J. Neto, "Audio Segmentation, Classification And Clustering in a Broadcast News Task", Proc. *ICASSP*, vol. 2, pp. 5-8, 2003

[2] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young, "The development of the 1996 HTK broadcast news transcription system," in DARPA Speech Recognition Workshop, 1997

[3] P. K. Atrey, N. C. Maddage and M. S. Kankanhalli, " Audio based event detection for multimedia surveillance", in Proc. ICASSP, vol. 5, pp. 813-816, 2006

[4] L. Lie, J. Hao and Z. HongJiang, "A robust audio classification and segmentation method", Proc. *9th ACM conference on Multimedia*, p. 203-211, 2001

[5] T. L. Nwe H. Li, "Broadcast news segmentation by audio type analysis", in Proc. *ICASSP*, vol. 2, pp. 1065-1068, 2005

[6] D. Li, I.K. Sethi, N. Dimitrova, T. McGee, "Classification of general audio data for content-based retrieval", in *Pattern Recognition Letters*, v. 22, pp. 533-544, 2001

Overview of the Albayzin 2010 Language Recognition Evaluation: database design, evaluation plan and preliminary analysis of results

Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
luisjavier.rodriguez@ehu.es

Abstract

This paper presents an overview of the Albayzin 2010 Language Recognition Evaluation, carried out from June to October 2010, organized by the Spanish Thematic Network on Speech Technology and coordinated by the Speech Technology Working Group of the University of the Basque Country. The evaluation was designed according to the test procedures, protocols and performance measures used in the last NIST Language Recognition Evaluations. Development and evaluation data were extracted from KALAKA-2, a database including clean and noisy speech in various languages, recorded from TV broadcasts and stored in single-channel 16-bit 16 kHz audio files. The task consisted in deciding whether or not a target language was spoken in a test utterance. Four different conditions were defined: closed-set/clean-speech, closed-set/noisy speech, open-set/clean-speech and open-set/noisy speech. Evaluation was performed on three subsets of test segments, with nominal durations of 30, 10 and 3 seconds, respectively. The task involved 6 target languages: English, Portuguese and the four official languages spoken in Spain (Basque, Catalan, Galician and Spanish), other (*unknown*) languages being also recorded to allow open-set verification tests. Four teams (2 from Spanish universities, one from a Portuguese research center and one from a Finnish university) presented their systems to this evaluation. The best primary system in the closed-set/clean-speech condition on the subset of 30-second segments yielded $C_{avg} = 0.0184$ (around 2% EER).

Index Terms: Language Recognition Evaluation, KALAKA-2, Spanish Thematic Network on Speech Technology

1. Introduction

The Albayzin 2010 Language Recognition Evaluation (Albayzin 2010 LRE), coordinated by the Software Technologies Working Group of the University of the Basque Country, with the support of the Spanish Network on Speech Technology [1], aimed to promote creativity, discussion and collaboration between research groups (specially from Spain and Portugal, though worldwide participation was welcome) working on automatic language identification and verification, to explore the limits of state-of-the-art technology and eventually to foster research progress and technological developments in this area.

Regarding the task, test conditions and performance measures, the Albayzin 2010 LRE was defined in almost the same terms as the last NIST Language Recognition Evaluations [2, 3], but considering a reduced set of target languages (Spanish, Catalan, Basque, Galician, Portuguese and English) and dealing with speech extracted from multi-speaker TV broadcast recordings. Note that a test segment could contain speech from various speakers. This is a relevant difference with regard to NIST evaluations, whose data were extracted from telephone-channel two-speaker conversations, test segments containing speech from a single speaker.

Test conditions for this evaluation were almost identical to those applied for the Albayzin 2008 LRE [4], with three important changes:

- Portuguese and English were added as target languages,
- The so-called *restricted development* condition was not considered anymore, and
- a new test condition involving noisy and/or overlapped speech was introduced.

Therefore, four different test conditions, depending on the operation mode (closed-set vs. open-set) and the background conditions (clean vs. noisy), were defined. Also, 3 nominal segment durations (30, 10 and 3 seconds) were considered, leading to 12 different tracks. An award was presented to the system yielding best performance in the CC-30 track (closed-set verification of 30-second segments containing clean-speech), which was mandatory.

The rest of the paper is organized as follows. The language detection task is briefly defined in Section 2. The test conditions and the measures used to evaluate system performance are described in Sections 3 and 4, respectively. Section 5 describes the database and Section 6 addresses issues related to the organization of Albayzin 2010 LRE. Results are presented and briefly discussed in Section 7, with special attention to the closed-set clean-speech condition (which was mandatory), and devoting some space to a special activity carried out after evaluation results were submitted. Finally, conclusions and future work are outlined in Section 8.

2. The language detection task

The language detection task was defined in the same terms as for NIST evaluations [2, 3]: *given a segment of speech and a language of interest (target language), determine whether or not that language is spoken in the segment, based on an automated analysis of the data contained in the segment*. Performance was computed by presenting the system a set of trials

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

and comparing system decisions with the right ones (stored in a keyfile).

Each trial comprises the following elements:

- a segment of audio containing speech in a single language,
- the identity of the target language of interest, and
- the identities of the languages that might be spoken in the segment (which we will call *non-target* languages).

For each trial, the system must output:

- a hard decision (yes/no) about whether or not the target language is spoken in the segment, and
- a score indicating how likely is for the system that the target language is spoken in the segment, the higher the score the greater the confidence that the segment contains the target language.

3. Test conditions

3.1. Closed-set vs. open-set verification

Depending on the restrictions imposed to the set languages that might be spoken in the segment, two types of verification tests were defined:

- In *closed-set verification*, the set of trials is limited to segments containing speech in one of the target languages, and scores are computed based on those trials. This means that, for each trial, non-target languages are limited to all the target languages except for the target language of interest in that trial.
- In *open-set verification*, scores are computed based on the whole set of trials, including those corresponding to segments containing speech in an *unknown* language. This means that, for each trial, non-target languages are all the possible languages except for the target language of interest in that trial.

This way, systems could be designed specifically for closed-set or open-set verification, and research groups were given the opportunity to submit separate results for each condition. The set of *unknown* languages were not disclosed to participants.

3.2. Clean vs. noisy speech

The development and evaluation datasets consisted of two subsets:

- *clean* segments, featuring high SNR speech signals, maybe with short fragments of noisy and/or overlapped speech (in a single language), and
- *noisy* segments, featuring noisy and/or overlapped speech (in a single language), maybe with short fragments of clean speech.

The subset of noisy segments might contain different and variable types of noise: street, music, cocktail party, laughs, clapping, etc. Telephone-channel speech signals were not be used in any case. Segments containing overlapped speech were extracted from informal debates in late night shows, magazines, etc. which, on the other hand, might feature clean-channel and quiet-background (studio) conditions. As noted above, each segment contains speech in a single language, which also applies to overlaps and fragments with background speech, except for the case of segments in unknown languages, which might contain speech in two or more languages, provided that none of them are target languages.

This condition was introduced with two objectives:

- to measure the performance of language verification systems designed to deal with clean speech, when dealing with noisy and/or overlapped speech, and
- to measure the performance of language verification systems specifically designed to deal with noisy and/or overlapped speech.

3.3. Duration of speech segments

With the aim to measure performance as a function of the available amount of speech, the development and evaluation sets were each divided into three subsets, containing segments of three nominal durations: 30, 10 and 3 seconds, respectively. Segments were defined to begin and end at times of non-speech as determined by an automatic speech activity detection algorithm. So, actual segment durations may be slightly longer (but not shorter) than nominal durations. Note that each segment was extracted from an original TV broadcast recording, containing speech in a single language (from one or more speakers) mixed with fragments of non-speech (silence or background noise), so the actual amount of speech was smaller than segment duration. Nominal segment durations were not disclosed to participants (though they could be guessed very easily).

4. Performance measures

The language verification task defined for this evaluation considers two types of errors: (1) *misses*, those for which the correct answer is *yes* but the system says *no*; and (2) *false alarms*, those for which the correct answer is *no* but the system says *yes*. Therefore, for any test condition the corresponding error rates can be computed as the fraction of target trials that are rejected (*miss rate*, P_{miss}) and the fraction of impostor trials that are accepted (*false alarm rate*, P_{fa}), and suitable cost functions can be defined as combinations of these basic error rates.

4.1. Average cost across target languages

Let assume that there are L target languages. Let $P_{miss}(i)$ be the miss rate computed on trials corresponding to target language i ($i \in [1, L]$), and $P_{fa}(i,j)$ the false alarm rate computed on trials corresponding to other language j (the index 0 representing *unknown* languages), that is, the fraction of trials corresponding to language j that are erroneously accepted as containing language i. The *pairwise cost* $C(i,j)$ is defined as follows:

$$\begin{aligned} C(i,j) &= C_{miss} \cdot P_{target} \cdot P_{miss}(i) + \\ &\quad C_{fa} \cdot (1 - P_{target}) \cdot P_{fa}(i,j) \end{aligned} \tag{1}$$

The cost model depends on three application parameters: C_{miss}, C_{fa} and P_{target}. For this evaluation, the same values used in the Albayzin 2008 LRE (which are also the same used in NIST 2007 and 2009 LRE) were applied:

$$C_{miss} = C_{fa} = 1$$
$$P_{target} = 0.5$$

Finally, an average cost is defined by adding the contributions for all the combinations of target and non-target languages, as follows:

$$\begin{aligned} C_{avg} &= \frac{1}{L}\sum_{i=1}^{L}\{C_{miss}\cdot P_{target}\cdot P_{miss}(i) \\ &+ \sum_{\substack{j=1\\ j\neq i}}^{L} C_{fa}\cdot P_{non-target}\cdot P_{fa}(i,j) \\ &+ C_{fa}\cdot P_{OOS}\cdot P_{fa}(i,0)\} \end{aligned} \qquad (2)$$

where $P_{non-target}$ is the prior probability of non-target languages (assuming for them a uniform distribution) and P_{OOS} the prior probability of *unknown* (*Out-Of-Set*) languages. In this evaluation, the following values were applied:

$$\begin{aligned} P_{OOS} &= \begin{cases} 0.0 & \text{closed-set condition} \\ 0.2 & \text{open-set condition} \end{cases} \\ P_{non-target} &= \frac{1 - P_{target} - P_{OOS}}{L-1} \end{aligned}$$

The average cost C_{avg} was computed separately for each of the four test conditions and for each of the three segment duration categories, and served as the main system performance measure in this evaluation.

4.2. Log-Likelihood Ratio (LLR) average cost

Sites may specify that their scores could be interpreted as log-likelihood ratios. In such cases, detection results were also evaluated in terms of the so called C_{LLR} [5], which is commonly used as an alternative performance measure in NIST evaluations. C_{LLR} shows two important features: (1) it allows us to evaluate system performance globally by means of a single numerical value; and (2) it does not depend on application costs.

Let $LR(X, i)$ be the *likelihood ratio* corresponding to segment X and target language i. The likelihood ratio can be expressed in terms of the conditional probabilities of X with regard to the alternative target and non-target hypotheses, as follows:

$$LR(X, i) = \frac{prob(X|i)}{prob(X|\neg i)} \qquad (3)$$

Let consider an evaluation set E, consisting of the union of $L+1$ disjoint subsets: E_j ($j \in [1, L]$) containing segments in the target language j, and E_0 containing segments in *unknown* languages. Pairwise costs $C_{LLR}(i, j)$, for $i \in [1, L]$ and $j \in [0, L]$, are defined as follows:

$$C_{LLR}(i,j) = \begin{cases} \frac{1}{|E_i|}\sum\limits_{X\in E_i}\log_2(1 + LR(X,i)^{-1}) & j = i \\ \frac{1}{|E_j|}\sum\limits_{X\in E_j}\log_2(1 + LR(X,i)) & j \neq i \end{cases} \qquad (4)$$

Finally, the average cost C_{LLR} is computed by adding the pairwise costs for all the combinations of target and non-target (including Out-Of-Set) languages, as follows:

$$\begin{aligned} C_{LLR} &= \frac{1}{L}\sum_{i=1}^{L}\{P_{target}\cdot C_{LLR}(i,i) \\ &+ \sum_{\substack{j=1\\ j\neq i}}^{L} P_{non-target}\cdot C_{LLR}(i,j) \\ &+ P_{OOS}\cdot C_{LLR}(i,0)\} \end{aligned} \qquad (5)$$

The cost function C_{LLR} returns an unbounded non-negative value which can be interpreted as information bits, with lower values representing better performance, the value 0 corresponding to a perfect system and the value $\log_2(L)$ corresponding to a system which just relies on (uniform) priors, thus providing no information to decide a trial. Further details about the reasons for using and the interpretation of C_{LLR} can be found in [5, 6].

4.3. Graphical evaluation: DET curves

Detection Error Tradeoff (DET) curves [7] provide a straightforward way of comparing global performance of different systems for a given test condition. A DET curve is generated by computing P_{miss} and P_{fa} for a wide range of operation points (thresholds), based on the scores yielded by the analyzed system for a given test set. Besides C_{avg} and C_{LLR}, DET curves are used in NIST evaluations to support system performance comparisons. In this evaluation, NIST software [8] was used to generate DET curves, including marks for the operation point given by system decisions and the operation point corresponding to the minimum C_{avg}.

5. Data

The database used for this evaluation, called KALAKA-2, was organized in three subsets: train, development and evaluation. Speech signals were extracted from TV broadcast recordings (news, documentaries, debates, interviews, reportages, magazines, late night shows, etc.), featuring various dialects and/or linguistic competence levels, speech modalities (planned speech, formal conversations, spontaneous speech, etc.), and diverse environment conditions. Broadcasts were digitally recorded using a Roland Edirol R-09 recorder, audio signals being stored in WAV files (PCM, 16 kHz, single channel, 16 bits/sample). We strongly recommended to prepare systems starting from the materials provided for this evaluation, but participants were allowed to use any available data and subsystems. The sets of TV shows posted to each subset were forced to be disjoint, meaning that any show appearing in one subset did not appear in the other two. This restriction was imposed as an attempt to guarantee speaker independence.

The database was designed as an extension of KALAKA, the database created ad-hoc for the Albayzin 2008 LRE [9]. To reduce development costs, all the materials of KALAKA were re-used for KALAKA-2, as follows:

- The train and development datasets of KALAKA were used to build the train dataset of KALAKA-2.
- The evaluation dataset of KALAKA was used to build the development dataset of KALAKA-2.

To complete the datasets of KALAKA-2, new TV broadcasts were recorded, selected and classified, specially for the two new target languages (Portuguese and English) and for the *unknown* (Out-Of-Set) languages. In particular, the evaluation dataset was completely new and independent of KALAKA.

5.1. Training data

The train dataset consisted of more than 10 hours of clean speech per target language. Its contents (fragments of variable length) did not all strictly consist of clean speech. Besides some portions of silence, they also featured short fragments containing noisy and/or overlapped speech. In a separate folder, more than 2 hours of noisy/overlapped speech were also provided for each target language. No data were provided containing *un-*

known (Out-Of-Set) languages. The distribution of training data is shown in Table 1.

Table 1: Distribution of training segments per target language for clean and noisy speech: number of segments (#) and total duration (T, in minutes).

	Clean speech		Noisy speech	
	#	T (minutes)	#	T (minutes)
Basque	406	644	112	135
Catalan	341	687	107	131
English	249	731	136	152
Galician	464	644	125	134
Portuguese	387	665	160	197
Spanish	342	625	133	222

5.2. Development and evaluation data

The development and evaluation datasets had the same size and characteristics, except for the distribution of unknown languages and the proportion of clean and noisy speech. Both datasets contained segments with nominal durations of 30, 10 and 3 seconds, with at least 150 speech segments per target language and nominal duration. Each segment contained speech (from one or more speakers) in one of the 6 target languages or in an *unknown* (Out-Of-Set) language. Speech segments were given random names, so that languages and durations appeared in a random sequence.

The development set consisted of 4950 speech segments, 3492 containing clean speech and 1458 containing noisy speech, their total duration being 21.24 hours (70% of the time corresponding to clean speech and 30% to noisy speech). The evaluation set consisted of 4992 speech segments, 3345 containing clean speech and 1647 containing noisy speech, their total duration being 21.43 hours (67% of the time corresponding to clean speech and 33% to noisy speech). The distribution of segments per language is shown in Table 2.

In the case of clean speech, speech segments of 30, 10 and 3 seconds were automatically extracted from fragments of clean speech according to the following criteria:

1. Speech segments must be enclosed by a certain amount of silence (i.e. low-energy frames), which is included as part of the segments. This way, it is expected to catch natural segments and to avoid cutting words.

2. A 30-second segment is validated if and only if it contains a valid 10-second segment. Similarly, a 10-second segment is validated if and only if it contains a 3-second segment.

3. Segments can be slightly longer (but not shorter) than their nominal duration: 3-second segments are allowed to last up to 5 seconds; 10-second segments are allowed to last up to 12 seconds; and 30-second segments are allowed to last up to 33 seconds.

In the case of noisy speech, segments from 30 to 35 seconds were manually extracted from recordings, and then segments with nominal durations of 10 and 3 seconds were automatically extracted from the former, according to the same criteria applied for clean speech.

Table 2: Distribution of segments per language (the same for each duration) in the development and evaluation datasets.

		Devel		Eval	
		clean	noisy	clean	noisy
Target languages	**Basque**	146	29	130	74
	Catalan	120	47	149	55
	English	133	60	135	69
	Galician	137	60	121	83
	Portuguese	164	77	146	58
	Spanish	136	83	125	79
Unknown languages	**Arabic**	100	25	115	22
	French	120	32	70	34
	German	108	73	13	32
	Romanian	0	0	111	43

6. Rules and schedule

All the registered participants received three DVD containing train speech for the six target languages, plus an additional DVD with development data, a keyfile and a scoring script which allowed to tune system parameters (such as verification thresholds, fusion weights, etc.). The scoring script was based on that used for the NIST 2007 and 2009 LRE, with minor changes needed to match the task and to add the identifiers of the 6 target languages considered in this evaluation. The evaluation dataset was released via web (restricted to registered participants) and eventually distributed in a single DVD at FALA 2010.

Registration involved the commitment to use data exclusively for research purposes, distribution being allowed only with explicit permission. After the evaluation, registered participants were allowed to use the data to develop or evaluate their own systems, provided that they acknowledged that use by means of a suitable reference:

> KALAKA-2. Speech database created for the Albayzin 2010 Language Recognition Evaluation, organized by the Spanish Network on Speech Technology. Produced by the Software Technologies Working Group (GTTS, http://gtts.ehu.es), University of the Basque Country.

Four test conditions were defined: CC (closed-set, clean-speech), CN (closed-set, noisy-speech), OC (open-set, clean-speech) and ON (open-set, noisy-speech). The ranking of systems in all conditions and for the three nominal segment durations was determined by taking into account the average cost C_{avg}, as defined in Section 4.1. Participants could send results for as many systems as they want, but only one primary system per test condition, the remaining systems being *contrastive*. Only primary systems were taken into account for rankings. The CC condition was mandatory: an award was presented for the best system (i.e. that yielding the least C_{avg}) in the CC condition on the subset of 30-second segments.

Detection results had to be sent in a format similar to that used for NIST evaluations: a text file with a trial per line, each trial consisting of 6 blank-separated fields: background condition (clean/noisy), target language, operation mode (closed-set/open-set), test file, decision and score. Since multiple systems could be submitted, a naming protocol was established, consisting of a site identifier, a test condition identifier and a system identifier (primary, contrastive1, contrastive2, etc.). Each participant committed to send a complete description of their systems, with the aim to give readers a clear sense of what each system was about (methods, references, training data, processing speed, etc.).

The Evaluation Schedule was as follows:

- *May 18, 2010*
  - The evaluation plan is released through the website of FALA 2010.
  - Registration for Albayzin 2010 LRE opens.
  - An online registration form is made available through the website of FALA 2010.
- *June 22, 2010.*
  - Train and development datasets are sent to registered sites via courier.
  - A *wiki* is activated to improve communication and collaboration between the registered participants and the organizing team.
- *July 15, 2010.*
  - Registration for Albayzin 2010 LRE closes.
- *September 27, 2010.*
  - The evaluation dataset is released via web (restricted to registered participants).
  - System submission (via e-mail) opens.
- *October 17, 2010.*
  - System submission deadline (24:00, GMT+1).
- *October 25, 2010.*
  - Preliminary results in all conditions and the keyfile for the evaluation dataset are released to participants through the wiki.
- *November 2, 2010.*
  - Deadline for submitting final system descriptions (including analysis of results).
- *November 10-12, 2010 (FALA 2010, Vigo, Spain).*
  - Albayzin 2010 LRE Workshop: delivery of a DVD including documentation and evaluation data to registered participants, poster presentations and discussion.
  - Plenary session: summary of results and awards.

7. Results

Four teams, two from Spain, one from Portugal and one from Finland, submitted their systems to the Albayzin 2010 LRE (see Table 3). Full descriptions of the submitted systems can be found as regular papers in the proceedings of FALA 2010. Results (in terms of C_{avg}) in the four test conditions and for the three segment durations are shown in Tables 4, 5, 6 and 7.

Regarding the mandatory condition, the best primary system on the subset of 30-second segments was submitted by GTC-VIVOLAB (thus the award winner), yielding $C_{avg} = 0.0184$ (around 2% EER). Note, however, that the best system in this condition was the second contrastive system submitted by L^2F, with $C_{avg} = 0.0181$. The postkey submissions by L^2F outperformed the replaced systems, but not their second contrastive system nor the primary system by GTC-VIVOLAB. The DET curves for all the primary (thick lines) and contrastive (thin lines) systems submitted to this condition (CC-30) are shown in Figure 1. The actual and the minimum achievable costs (marked in the DET curves of primary systems with X and O, respectively) are shown in Figure 2, revealing calibration losses for some systems.

Table 3: Teams participating in the Albayzin 2010 Language Recognition Evaluation.

Team ID	**Research institution**	**Submitted conditions**
GTC-VIVOLAB	University of Zaragoza	CC, OC
L2F	L^2F INESC-ID Lisboa	All
UEF_NTNU	Univ. of Eastern Finland	CC
GTM	University of Vigo	CC, CN

Table 4: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the CC test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	CC-30	**CC-10**	**CC-3**
VIVOLAB_UZ_CC_pri	**0.0184**	0.0418	0.0943
VIVOLAB_UZ_CC_alt1	0.0238	0.0498	0.1087
L2F_CC_pri	0.0320	0.0513	0.1034
L2F_CC_pri_postkey	0.0223	0.0359	0.0853
L2F_CC_alt1	0.0910	0.0540	0.1065
L2F_CC_alt1_postkey	0.0219	0.0363	0.0844
L2F_CC_alt2	0.0181	0.0459	0.1055
UEF-NTNU_CC_pri	0.1636	0.3035	0.3799
UVIGO-GTM_CC_pri	0.1916	0.2934	0.4447
UVIGO-GTM_CC_alt1	0.2888	0.3181	0.3956

Table 5: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the CN test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	CN-30	**CN-10**	**CN-3**
L2F_CN_pri	0.0316	0.0767	0.1503
L2F_CN_pri_postkey	0.0416	0.0810	0.1273
L2F_CN_alt1	0.3556	0.0892	0.2080
L2F_CN_alt1_postkey	0.0403	0.0754	0.1217
L2F_CN_alt2	0.0253	0.0636	0.1342
UVIGO-GTM_CN_pri	0.2744	0.3534	0.4476
UVIGO-GTM_CN_alt1	0.2978	0.3412	0.4309

Regarding the dependence on the available amount of speech, for the most competitive systems the C_{avg} obtained on the subset of 10-second segments doubled that obtained on the subset of 30-second segments. The same trend was observed for 3-second segments with regard to 10-second segments (e.g. see results for the best primary system in the CC condition). This was consistent with previous results for other evaluations. The following analyses will focus on the subset of 30-second segments.

As may be expected, the performance degraded in open-

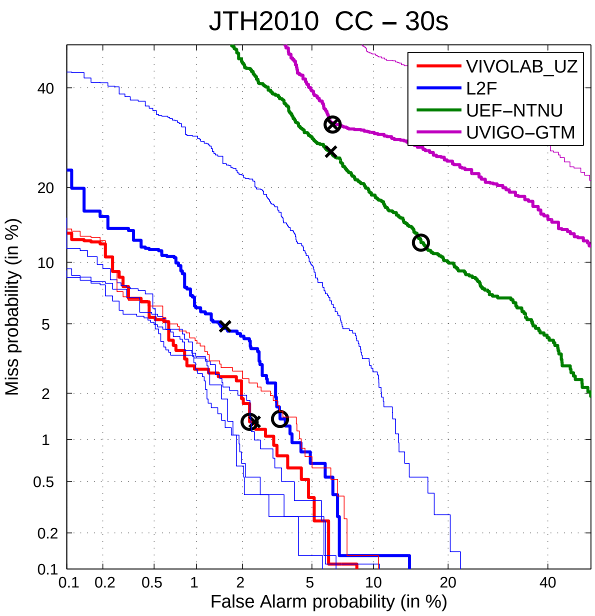


Figure 1: Pooled DET curves of systems submitted to the Albayzin 2010 LRE in the CC condition for the subset of 30-second segments.

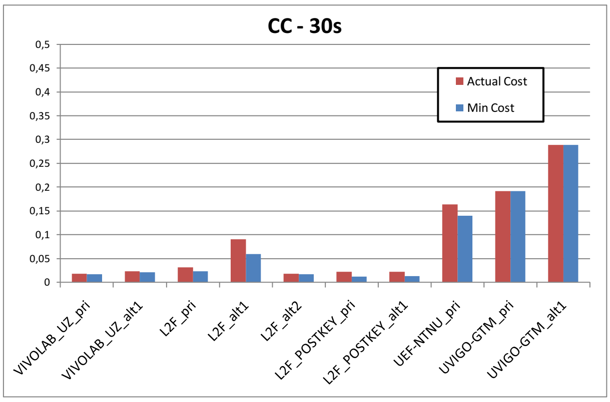


Figure 2: Actual and minimum achievable costs of systems submitted to the Albayzin 2010 LRE in the CC condition for the subset of 30-second segments.

set tests, due to the increase of false alarms in trials involving speech signals in *unknown* languages (e.g. see Tables 4 and 6). For instance, the primary system of GTC-VIVOLAB yielded $C_{avg} = 0.0307$ in the OC-30 condition, which means around 67% increase in cost with regard to the CC-30 condition. Similar figures were observed for other systems in the same conditions: 49% and 88% cost increases for the primary and second contrastive L^2F systems, respectively. A detailed study of the confusion among languages is not included here for a lack of space.

Finally, a new condition was introduced in this evaluation with the aim to test how much the performance of language recognition systems degraded when dealing with noisy and/or overlapped speech. Not all the sites submitted results for the CN and ON conditions. In fact, only L^2F submitted results for all the conditions, so the analysis will focus on the primary and second contrastive L^2F systems (which are competitive systems with a consistent behavior across all conditions). Performance degradation was not so catastrophic as we expected. In fact, when comparing CN-30 with CC-30 results, the primary L^2F system surprisingly showed a slight improvement, whereas the second contrastive L^2F system showed *only* a 40% cost increase. The latter result is quite representative, since the increase in cost ranges from 30% to 50%, depending on the system and condition. In any case, it seems that good performance can be attained even on noisy speech if data are provided to train and calibrate systems.

Table 6: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the OC test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	OC-30	**OC-10**	**OC-3**
VIVOLAB_UZ_OC_pri	0.0307	0.0644	0.1202
VIVOLAB_UZ_OC_alt1	0.0373	0.0635	0.1309
L2F_OC_pri	0.0478	0.0750	0.1297
L2F_OC_pri_postkey	0.0296	0.0468	0.1073
L2F_OC_alt1	0.1416	0.1225	0.1460
L2F_OC_alt1_postkey	0.0309	0.0445	0.1029
L2F_OC_alt2	0.0341	0.0611	0.1289

Table 7: Performance (C_{avg}) of primary and contrastive systems submitted to Albayzin 2010 LRE in the ON test condition. Systems submitted after the keyfile release are shown too (though they do not count for rankings).

	ON-30	**ON-10**	**ON-3**
L2F_ON_pri	0.0749	0.1092	0.1735
L2F_ON_pri_postkey	0.0700	0.0981	0.1551
L2F_ON_alt1	0.3778	0.1311	0.2328
L2F_ON_alt1_postkey	0.0839	0.0948	0.1609
L2F_ON_alt2	0.0475	0.0936	0.1654

7.1. Processing times

Processing times for the submitted systems, in terms of real-time factor (×RT), along with the CPU and memory specifications of the servers used to run the experiments, are shown in Table 8 (only data provided by the participating teams are shown). All the systems are reported to run under 1×RT, but on servers with very different computational power. The most competitive systems have reported processing times of 0.9 (GTC-VIVOLAB) and 0.51 (L^2F).

Table 8: Processing time (×RT) for the submitted systems.

Systems	**CPU-RAM**	**×RT**
GTC-VIVOLAB	–	0.9
L2F	2xQuad Xeon E5530 2.4GHz, 48 GB	0.51
UEF_NTNU	Xeon X5450 3.0GHz	0.051
GTM (p)	Xeon E5620 2.4 GHz,18 GB	0.0288
GTM (c)	Xeon E5620 2.4 GHz,18 GB	0.0533

7.2. Exploring cross-site fusions

We proposed to participants an interesting way of collaboration: to investigate which subsystems combined better under a FoCal-based fusion paradigm, which may help future developments of language recognition systems (and potential collaborations). We focused on the core condition (closed-set, clean speech, 30-second segments). To accomplish that objective, we asked for them to submit log-likelihoods for their subsystems, giving details of the applied methodology. This way, previously unexplored cross-site fusions may give valuable cues of which kind of systems would be worth developing and combining.

Three sites submitted the log-likelihoods for the six target languages produced by their subsystems on the CC-30 evaluation subset. Note that this information had not been previously disclosed, since each team studied and optimized the fusion of subsystem scores, and what they called *system* was in fact the fusion of various subsystems. Additionally, the organizing team (GTTS) included the log-likelihoods of its own subsystems: three phonotactic-SVM subsystems, for Czech, Hungarian and Russian BUT decoders, using expected n-gram counts (up to 3-grams) computed on phone-lattices.

All the information was uploaded and results presented through the wiki created for this evaluation. For a lack of space, we do not include results here. Only note that the best cross-site fusion (including 5 subsystems from GTTS, GTC-VIVOLAB and L^2F) yielded $C_{avg} = 0.0054$, almost three times lower than that obtained by the best system in the CC-30 condition.

8. Conclusions

In this paper, the main features of the Albayzin 2010 Language Recognition Evaluation have been described, and results obtained by the submitted systems have been presented and briefly discussed. The evaluation involved six target languages: the four official languages spoken in Spain (Basque, Catalan, Galician and Spanish) plus Portuguese and English. A new database, KALAKA-2, was created for the evaluation, including clean and noisy speech in various languages, recorded from TV broadcasts and stored in single-channel 16-bit 16 kHz audio files.

In closed-set clean-speech verification tests on the evaluation subset of 30-second segments, the best primary system, employing state-of-the-art technology, yielded $C_{avg} = 0,0184$. This reveals a remarkable technology improvement with regard to the previous Albayzin 2008 LRE, where the best system yielded $C_{avg} = 0,0552$ on a similar task.

A new condition has been introduced in this evaluation, with the aim to evaluate performance degradation when dealing with noisy and/or overlapped speech. The increase in cost observed in noisy-speech tests (with regard to clean-speech tests) ranged from 30% to 50%, depending on the system and condition. This reveals that reasonably good performance can be attained even on noisy speech if enough and suitable data are available to train and calibrate systems.

Finally, a post-eval activity was organized which tried to investigate which subsystems combined better under a FoCal-based fusion paradigm. Starting from the log-likelihoods for the six target languages produced by the subsystems of various teams, we discovered that cross-site fusion may provide great performance improvements (the best 5-subsystem fusion yielding $C_{avg} = 0.0054$).

9. Acknowledgements

We thank the Organizing Committee of FALA 2010, specially Laura Docio, coordinator of the Albayzin 2010 Evaluations, for their help and support. We also thank all the participants, for their work and feedback, which were essential for the success of this evaluation.

10. References

[1] *Spanish Network on Speech Technology*. Web (in Spanish): http://lorien.die.upm.es/∼lapiz/rtth/.

[2] A. F. Martin and A. N. Le, “NIST 2007 Language Recognition Evaluation,” in *Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008.

[3] A. Martin and C. Greenberg, “The 2009 NIST Language Recognition Evaluation,” in *Odyssey 2010 - The Speaker and Language Recognition Workshop, paper 030*, (Brno, Czech Republic), 28 June - 1 July 2010.

[4] L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, and A. Varona, “The Albayzin 2008 Language Recognition Evaluation,” in *Odyssey 2010: The Speaker and Language Recognition Workshop*, (Brno, Czech Republic), 28 June - 1 July 2010.

[5] N. Brümmer and J. du Preez, “Application-Independent Evaluation of Speaker Detection,” *Computer, Speech and Language*, vol. 20, pp. 230–275, April-July 2006.

[6] N. Brümmer and D. van Leeuwen, “On calibration of language recognition scores,” in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, pp. 1–8, 2006.

[7] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET Curve in Assessment of Detection Task Performance,” in *Proceedings of Eurospeech*, pp. 1985–1988, 1997.

[8] *NIST DET-Curve Plotting software for use with MATLAB*. http://www.itl.nist.gov/iad/mig/tools/ DETware_v2.1.targz.htm.

[9] L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, A. Varona, and M. Diez, “KALAKA: A TV broadcast speech database for the evaluation of language recognition systems,” in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, (Valleta, Malta), 17-23 May 2010.

The Albayzín 2010 Text–to–Speech Evaluation

Francisco Méndez Pazó[1], *Laura Docío Fernández*[1],
Montserrat Arza Rodríguez[2], *Francisco Campillo Díaz*[1]

[1]Group on Multimedia Technologies, Universidade de Vigo, Spain
[2]Centro Ramón Piñeiro para a Investigación en Humanidades. Xunta de Galicia.
{fmendez, ldocio, marza, campillo}@gts.uvigo.es

Abstract

This paper describes the Albayzín 2010 Text-to-Speech Evaluation Campaign. 6 teams submitted a total of 10 systems to the evaluation. Participants were asked to build their system from a common Spanish database and synthesize a set of test sentences. An online evaluation in three sections (similarity to the original speaker, naturalness and intelligibility) was conducted. Also a complete statistical analysis and results of the evaluation are presented.
Index Terms: speech synthesis, subjective evaluation

1. Introduction

The main goal of the Albayzín Text-to-Speech evaluation is to compare the different techniques employed by the research teams in building their TTS systems. With this purpose in mind, a contest has been proposed among speech synthesis systems, similar to the *Blizzard Challenge* [1] for English and Mandarin. The voice development was carried out on a Spanish common database. The participating teams had up to seven weeks to build their system from the development material that was supplied. Next, a set of test sentences was released and the teams had five days to synthesize them and send the audio files back. No manual intervention was allowed during synthesis (neither prompt sculpting, nor using different subsets of the database for different test sentences or sentence types unless this is a fully automatic part of the system). Then, a subset of these test sentences was evaluated online in three different listening tests: similarity with the original voice, naturalness and intelligibility. Each participant was expected to provide fifteen speech experts and Spanish native speakers as listeners of the evaluation.

2. Participants

The Albayzín 2008 Text-to-Speech Evaluation [2] had 7 participant groups, with a total of 8 systems submitted. This year there were 6 groups, who submitted a total of 10 different systems to the evaluation:

AhoHTS University of the Basque Country (AhoLab)

AhoTTS University of the Basque Country (AhoLab)

Cotovía University of Vigo, Group on Multimedia Technologies (GTM)

Cotovía-hts University of Vigo, Group on Multimedia Technologies (GTM)

GTHCSTR-2008 Technical University of Madrid - University of Edinburgh (GTH-CSTR)

GTHCSTR-2010 Technical University of Madrid - University of Edinburgh (GTH-CSTR)

GTHCSTR-2010-adaptation Technical University of Madrid - University of Edinburgh (GTH-CSTR)

MS-HTS-TTS Microsoft Language Development Center

Ogmios UPC-Barcelona Tech

SalleTTS La Salle – Ramon Llull University, Group on Multimedia Technologies Research (GTM)

See appendix A for a summary table of main characteristics of the systems submitted to the evaluation.

3. Evaluation

3.1. Speech Database

The Uvigo_esda Database contains speech recordings from an amateur male speaker that read prompted texts in “neutral” style.

The database contains approximately 2:00 hours of speech. Collection was performed at the recording studio of the Signal Theory Group of the University of Vigo. The owner of the database is the Signal Theory Group. It consists of 1217 phonetically balanced sentences, automatically extracted from journalistic texts by means of a greedy algorithm.

The following information is provided for each utterance:

- Audio data: the speech waveform. The signal was sampled at 16 kHz, with 16 bits resolution. Also provided are the original recordings, sampled at 44.1 kHz, with 16 bits resolution.
- Phone segmentation: phone labels were automatically extracted using the segmentation tools of HTK. SAMPA code is used for phones.
- Pitch marks files, extracted with Praat.
- Prompt text. Silences are always marked with commas or periods, and they are aligned with the phone segmentation files.
- Prompt text with information about intonation boundaries. Two markers are added:
  1. , $\#R - E\# \rightarrow$ Intonation break. Boundary of intonation group. Commas and periods are assumed to be always boundaries of intonation group, so in those cases this marker is not included.
  2. , $\#R - C\# \rightarrow$ Intonation break related to a comma. Boundary of intonation group related to a comma in the original sentence that was not realized by the speaker (this is, even though there is a comma in the text, in the corresponding phone segmentation file there is no silence).

	Sect. 1	Sect. 2	Sect. 3	Total
Completed	137	135	132	132
Partially completed	3	1	0	8
No response at all	7	11	15	7
Total registered	147			

Table 1: *Number of listeners.*

- Lexicon table derived from the texts. The lexicon includes all the words in the corpus and contains the different pronunciations found there.

3.2. Test Sentences

There were two different sets of test sentences:

- 350 held-out phonetically balanced sentences from the Spanish corpus Uvigo_esda, automatically extracted and belonging to four different broad types: declarative, interrogative, exclamatory and suspensive. Too long sentences were considered to be unsuitable for the test, so they were excluded.
- 82 semantically unpredictable sentences, manually designed for the intelligibility test. These sentences are seven words long, all with the same morphosyntactic structure: DETERMINER + NOUN + ADJECTIVE + VERB + DETERMINER + NOUN + ADJECTIVE.

Participants had five days to synthesize all 432 sentences and send back the synthetic audio files.

3.3. Listening test design

The evaluation was carried online, following the design developed for the Blizzard Challenge 2007 [1]. In the registration page, listeners had to fill in certain information (age group, gender, speech technologies expertise, use of speakers or headphones and whether they were Spanish language native or not). This initial page also presented a short overview and instructions to the tasks they were expected to complete.

The test was divided into three different sections which could be completed in any order, and in several sessions if desired. Around 30 minutes were needed to complete the whole test.

Speech Technology Expert	Yes	64
	No	83
Spanish Native	Yes	134
	No	13
Listening equipment	Headphones	119
	Loudspeakers	28
Gender	Male	98
	Female	49

Table 2: *Information about registered listeners.*

The online evaluation was open for approximately three weeks. The number and type of listeners that participated is summarized in tables 1 and 2.

3.3.1. Section 1 - Similarity to the original voice

In each part listeners could play four fixed reference samples of the original voice talent and one synthetic sample. Then, they had to score the synthetic sample taking only into account how similar to the original voice it was, on a scale ranging from 1 *[Sounds like a totally different person]* to 5 *[Sounds like exactly the same person]*. In this section each evaluator had to listen to 11 audio files, one from each system and another one from the original recording.

3.3.2. Section 2 - Naturalness MOS (Mean Opinion Scores)

In each part evaluators listened to one sample and chose a score which represented how natural or unnatural the sentence looked like on a scale between 1 *[Completely Unnatural]* and 5 *[Completely Natural]*. This was the main section of the listening test, and each evaluator had to listen to a total of 44 audio files, 4 for each participant plus 4 natural ones.

3.3.3. Section 3 - Semantically Unpredictable Sentences (SUS)

Semantically unpredictable sentences were designed to test the intelligibility of synthetic speech. Evaluators listened to one utterance in each part and typed in what they heard. In this part each listener had to transcribe two sentences from each system, 20 in total (no original voice in this section, since there were no recordings of these SUS sentences in the Uvigo_esda database).

3.3.4. Listener groups

Following the Blizzard 2007 design, listeners were grouped in 11 (10 submitted systems plus original voice) groups using a Latin square strategy. For each listener group, each section of the test had a different system ordering: all evaluators listened the same 75 sentences in the same order, but synthesized by different systems.

3.3.5. Listening test sentences selection

The sentences used in the online evaluation are a subset of the test sentences described in section 3.2. For sections one and two, sentences were randomly chosen to meet the following criteria:

- Interrogative: two sentences in section 1 and 6 in section 2.
- *Short* sentences (5-7 words): 2 sentences in section 1 and 6 in section 2.
- *Long* sentences (20-24 words): 2 sentences in section 1 and six in section 2.
- *Normal* sentences (10-15 words): 5 sentences in section 1 and 26 in section 2.
- Neither exclamatory nor suspensive sentences were selected.
- Sentences with foreign words or words which had to be text normalized were excluded.

SUS sentences to be included in the listening test were selected manually.

4. Analysis methodology

A complete statistical analysis was made using the programs and scripts available from the Blizzard Challenge [4]. System names were anonymised assigning letters from A to K, A representing natural speech. For each section in the listener test, a summary table with descriptive statistics is presented: median, median absolute deviation (mad), mean, standard deviation (sd), number of samples (n) and number of missing samples (na).

In these tables, systems are sorted in descending order of the mean scores in sections 1 and 2. This order cannot be interpreted as a ranking, it is intended only for readability, as it is more appropriate to compare medians than means in this MOS Likert-type scale (see [3]).

For word error rates, it makes sense to compare means, so ordering is in ascending order of the means.

4.1. Box-plots

For sections one and two (similarity to original voice and naturalness), standard box-plots are presented for the ordinal data where the median is represented by a solid bar across a box showing the quartiles; whiskers extend to 1.5 times the interquartile range and outliers beyond them are represented as circles. Also provided are tables with information about the data used in the box-plots: minimum value, first quartile (1C), the median or second quartile, the third quartile (3C), maximum value, and the lower and upper confidence limits for the median (LCL and UCL). Alphabetic ordering was used in these box-plots.

4.2. Wilcoxon test

To determine whether there are significant statistical differences between the MOS scores of systems a series of Bonferroni-corrected pairwise Wilcoxon signed rank significance tests with $\alpha = 0.01$ are used. It is represented with a symmetrical matrix in which significant differences between two systems are represented with a "•", while blank spaces denote no significant statistical differences.

4.3. Word error rates

For section 3, word error rates (WER) are calculated automatically, using the same methodology and scripts as in the Blizzard Challenge [1][4]. Capitalization and written accent marks were ignored, and certain common orthographic errors, such as confusion between *b* and *v* or erroneous misuse of *h*, were allowed. Also allowance was made for certain spelling variations in listener responses. Compounding or splitting words (e.g. *chico leo* instead of the correct word *chicoleo*) are also handled.

Bar charts instead of box-plots are presented for the word error rate interval data.

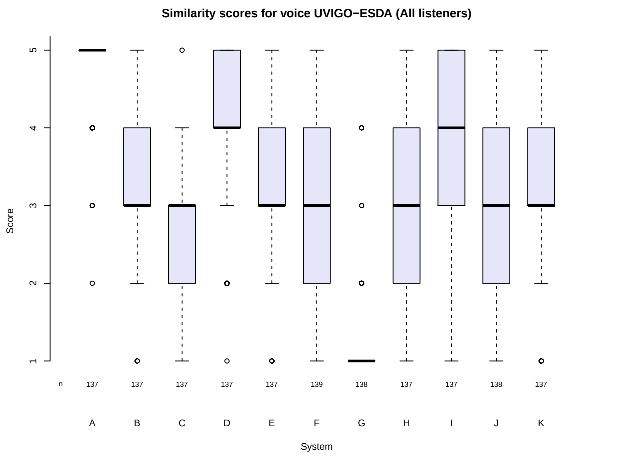


Figure 1: *Boxplot: Similarity scores for all listeners*

	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.87	3.13
C	1	2	3	3	5	2.87	3.13
D	1	4	4	5	5	3.87	4.13
E	1	3	3	4	5	2.87	3.13
F	1	2	3	4	5	2.73	3.27
G	1	1	1	1	4	1.00	1.00
H	1	2	3	4	5	2.73	3.27
I	1	3	4	5	5	3.73	4.27
J	1	2	3	4	5	2.73	3.27
K	1	3	3	4	5	2.87	3.13

Table 3: Boxplot data: Similarity scores for all listeners

5. Results

In this section the results obtained in each part of the listening test are presented. Unless notated, figures and tables display the results for all listener types combined, including those listeners who completed only partially any section of the test. A detailed report of the statistics for the different listener groupings shown in table 2 is gathered in the appendixes B, C and D.

5.1. Section 1: Similarity to the original speaker

	median	MAD	mean	sd	n	na
A	5	0.00	4.83	0.49	137	10
D	4	1.48	4.07	0.94	137	10
I	4	1.48	4.02	0.94	137	10
B	3	1.48	3.34	0.94	137	10
H	3	1.48	3.23	1.14	137	10
K	3	1.48	3.20	0.99	137	10
E	3	1.48	3.15	0.97	137	10
J	3	1.48	3.13	1.11	138	9
F	3	1.48	2.91	1.12	139	8
C	3	1.48	2.54	0.96	137	10
G	1	0.00	1.25	0.60	138	9

Table 4: Similarity scores for all listeners

Results for section one are presented in figure 1 and tables 3, 4 and 5. As expected, natural speech obtained the highest score, with a median of 5. Two systems, D and I, perform significantly better than the average, scoring a median of 4. Although comparing to the results obtained in the last Albayzín 2008 TTS Evaluation [2] (best system with a median of 3) is not completely fair since a very different speech database was used, these two systems show a great progress, although there is a lot of room for improvement yet. Then, a group of 6 systems (B, H, K, E, J, F) with a median of 3, shows no significant statistical differences. System C also scored a median of 3, but with a somewhat lesser mean (2.54). Finally, G was the worst performing system in this section, with a median of 1 which states as *Sounds like a totally different person*. As a conclusion, this test shows that concatenative systems (D, I, H, K) can deliver speech more similar to the original than the statistical

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•		•	•		•		
C	•	•		•	•		•	•	•	•	•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•	•		•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•		•	•
J	•		•	•			•		•		
K	•		•	•			•		•		

Table 5: Wilcoxon test: Similarity scores for all listeners

	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.93	3.07
C	1	2	2	3	5	1.93	2.07
D	1	3	4	4	5	3.93	4.07
E	1	2	3	4	5	2.86	3.14
F	1	2	3	4	5	2.86	3.14
G	1	1	1	1	4	1.00	1.00
H	1	2	3	3	5	2.93	3.07
I	1	3	4	4	5	3.93	4.07
J	1	2	3	4	5	2.86	3.14
K	1	2	3	3	5	2.93	3.07

Table 6: Boxplot data: Mean opinion scores for all listeners

parametric synthesizers.

The results for the different groups of listeners according to the types shown in table 2 are quite similar to those obtained for all listeners, and are shown in Appendix B.

5.2. Section 2: Mean opinion scores

Figure 2 and tables 6, 7 and 8 show the results for section 2. Again, only original speech achieved a median of 5. Then, two systems, D and I scored a median of 4 (*mostly natural* voice). These two systems again outperformed the best system in the last Albayzín 2008 TTS Evaluation. The rest of participants obtained a median of 3, except system C (median of 2) and finally system G was again the least scored system with a median of 1 (*Completely Unnatural* voice). By inspecting the significant statistical differences shown by Wilcoxon test (table 8), two groupings can be established: systems E and F (MOS of 3.15 and 3.10) and another group of three systems, K, H and C, with means around 2.50-2.60. The rest of positions are well defined.

Appendix C shows the results for the different groupings of listeners according to the types shown in table 2. System D was consistently the most natural system for all listeners but those who listened through loudspeakers, in which case mean score decreased from 3.86 (listeners who used headphones) to 3.40.

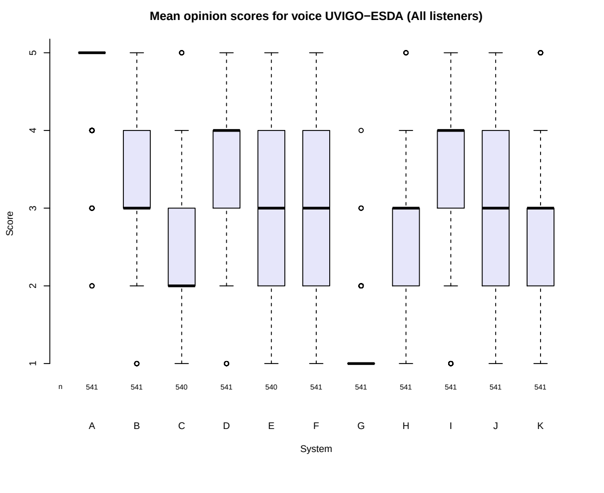


Figure 2: *Boxplot: Mean opinion scores for voice all listeners*

	median	MAD	mean	sd	n	na
A	5	0.00	4.75	0.57	541	47
D	4	1.48	3.78	0.98	541	47
I	4	1.48	3.50	1.02	541	47
B	3	1.48	3.33	0.95	541	47
F	3	1.48	3.15	1.00	541	47
E	3	1.48	3.10	0.96	540	48
J	3	1.48	2.91	1.00	541	47
K	3	1.48	2.62	0.98	541	47
H	3	1.48	2.60	0.97	541	47
C	2	1.48	2.51	0.90	540	48
G	1	0.00	1.10	0.34	541	47

Table 7: Mean opinion scores for all listeners

5.3. Section 3: Word error rates

The results for the intelligibility test are displayed in figure 3 and table 9. They show only the responses of native listeners (122 out of 132 who completed this section) due to the unusual and difficult kind of words that were used in the SUS sentences design. In fact, by inspecting the results obtained by non-native listeners (see appendix D), much higher error rates are observed, ranging from 24 to 49%, showing no significant differences between any system, thus suggesting that their results might not be taken into account. Therefore, system E was

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•	•	•	•	•	•	•	•
C	•	•		•	•	•	•		•	•	
D	•	•	•		•	•	•	•	•	•	•
E	•	•	•	•			•	•	•	•	•
F	•	•	•	•			•	•	•	•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•	•	•	•	•	•	•	•		•	•
J	•	•	•	•	•	•	•	•	•		•
K	•	•		•	•	•	•		•	•	

Table 8: Wilcoxon test: Mean opinion scores for all listeners

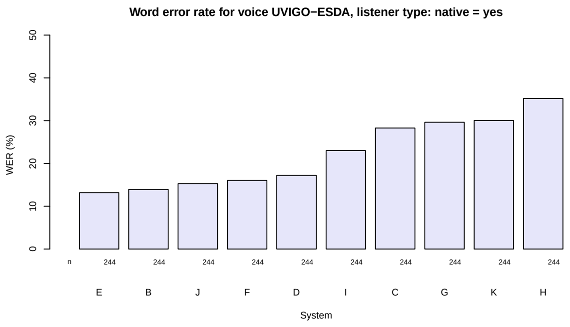


Figure 3: *Word error rates for native listeners*

the best system here, scoring a WER of 13%, although there are no significant statistical differences with systems B, D, F and J, all of them with word error rates ranging between 14 and 17%. System I achieved a WER of 23% and finally C, G, H and K got around 30%.

These figures clearly state that the SUS sentences used in the intelligibility task this year were considerably more difficult than 2008's, when error rates were around 5%. As far as no original recording of the SUS sentences was available, no comparison against the natural voice was done, so it is difficult to discern if these bad results are due to the difficult nature of the test sentences or a consequence of poor readings by the submitted systems.

	median	MAD	mean	sd	n	na
E	0.00	0.00	0.13	0.17	244	24
B	0.00	0.00	0.14	0.20	244	24
J	0.14	0.21	0.15	0.21	244	24
F	0.14	0.21	0.16	0.19	244	24
D	0.14	0.21	0.17	0.20	244	24
I	0.14	0.21	0.23	0.22	244	24
C	0.29	0.42	0.28	0.25	244	24
G	0.29	0.42	0.30	0.26	244	24
K	0.29	0.21	0.30	0.26	244	24
H	0.29	0.42	0.35	0.27	244	24

Table 9: *Word error rates for native listeners*

6. Listener feedback

The listeners who completed the evaluation could if desired fill in a questionnaire with comments and suggestions about the evaluation. Here is a short summary of some of this listener feedback.

- Almost all completed the test in the same quiet environment.
- Majority (106 out of 132) did the evaluation in one session.
- Similarity and naturalness tests were mostly considered easy (109 listeners), listening to each sample only 1–2 times.
- For the intelligibility task, 74 listeners usually understood all or most of the words, while the rest had troubles to understand it. The majority (91) had to listen each sentence 3–5 times.
- There were mixed opinions about the length of the test: short for some evaluators and too long and tiring for other ones, often pointing to the 44 samples which had to be listened to in section 2.
- SUS section also was a mixed bag for the likemost-likeleast opinions: difficult, hard, funny, amusing, interesting…
- Some people suggested to include some examples of the submitted systems in order to identify the scale prior to actually begin the test.

7. Conclusions

This paper is dedicated to the Albayzín 2010 TTS evaluation, from the description of the common Spanish database, to the listening tests and results.

On comparing with the Albayzín 2008 TTS evaluation, the first detail that stands out is the different distribution of system technologies. While in 2008 there were 7 concatenative systems and only 1 HMM-based, in 2010 there were 3 purely concatenative systems, 6 HMM-based, and a hybrid one. This agrees with the trend change that seems to be going on nowadays, being HMM-based a very promising approach.

Regarding the evaluation, three different tests were conducted: similarity with the original voice, naturalness, and intelligibility. Although comparing performances of systems in different tests, and with different voices, might be misleading, it can give an idea of the evolution of the systems. With respect to similarity with the original voice, results clearly outperform those obtained in Albayzín 2008. Moreover, there are two systems (D and I) that achieve a median of 4. About naturalness, these same two systems improve the results of Albayzín 2008, but both of them are still far from 4, which shows that there is still a lot of room for improvement. With regards to intelligibility, results were definitely worse than 2008, but the authors consider it to be a consequence of the test being much more difficult.

Finally, the general main characteristics of concatenative and HMM-based systems are also reflected in the results. While concatenative systems tend to get better scores in naturalness and similarity with the original voice, HMM-based systems tend to get better results in intelligibility.

8. Acknowledgments

The Albayzín 2010 TTS Evaluation organizing committee would like to thank the organizing committee of the Albayzín TTS 2008 and the Blizzard Challenge, whose work has served as a model for the Albayzín 2010 TTS Evaluation. Also thanks to the Blizzard Challenge for kindly providing the tools used to run the listening test and analyze the results [4]. Finally, thanks to all participants and volunteer listeners who completed the on-line evaluation.

The work reported here was partially supported by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04 "Búsqueda de Información en Contenidos Audiovisuales plurilingües", Xunta de Galicia "Isidro Parga Pondal" research programme and *Centro Ramón Piñeiro para a Investigación en Humanidades*.

9. References

[1] Mark Fraser and Simon King, "The Blizzard Challenge 2007" , In Proc. Blizzard Workshop (in Proc. SSW6), 2007

[2] Iñaki Sainz, Eva Navas, Inma Hernáez, Antonio Bonafonte, Francisco Campillo, "TTS evaluation campaign with a common Spanish database", Proceedings of LREC, 2010.

[3] R.A.J. Clark, M. Posiadlo, M. Fraser, C. Mayo and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results" , In Proc. Blizzard Workshop (in Proc. SSW6), 2007

[4] Online http://www.cstr.ed.ac.uk/projects/blizzard/tools.html

A. Participant overview

System	Team					System				Technical		
	Name	Albayzin 2008	People	Hours/ person	Native	Develop. years	Availability	Platforms	Development language	Memory footprint	Sub-words units	Type
AhoHTS	Aholab	yes	2	40	yes	15+	Still not	Linux	C/C++	50MB	context-dependent HMMs	Parametric synthesizer
AhoTTS	Aholab	yes	2	40	yes	15	Research	Linux	C/C++	400MB	Demiphone	Concatenative hybrid
Cotovía	Uvigo-GTM	yes	2	3	yes	10	Research purposes	Linux Windows	C/C++ perl, Matlab	Train: 1GB Synth:50MB	Demiphone	Unit selection
Cotovía-HTS	Uvigo-GTM	yes	2	3	yes	3 months	Research purposes	Linux	C/C++ , awk perl, bash	Train:9.5GB Synth:150MB	Context-dependent quin-phone HMMs	HMM based
GTHCSTR-2008	GTH-CSTR	yes	6	6	50%	8	Partially free	Linux macosx	Scheme, Linux shell scripts, C	100MB	quinphone	Statistical parametric
GTHCSTR-2010	GTH-CSTR	yes	6	6	50%	10	Partially free	Linux macosx	Scheme, Linux shell scripts, C	500MB	quinphone	Statistical parametric
GTHCSTR-2010-adapt	GTH-CSTR	yes	6	6	50%	7	Partially free	Linux macosx	Scheme, Linux shell scripts, C	100MB	quinphone	Statistical parametric
Ms-HTS-TTS	Microsoft LDC	no	3	5	yes	2	yes	desktop, telephony, server	C#/C++	N/A	N/A	HTS
Ogmios	UPC	yes	4	10	yes	12	Research	Linux Windows	C++	300MB	Demiphones	Unit selection
SalleTTS	GTM La Salle U. Ramon Llull	yes	6	500	yes	10	License agreement	Windows Linux	C++	200MB	Diphones	Unit selection

Table 10: *Albayzín 2010 TTS participant questionnaire — Part 1.*

System	**Voice Building**				**Components**				
	CPU Hours	Labelling system	Manual verif	Labels	Tools	Lexicon	Prosodic model	Target cost	Join cost
AhoHTS	10–15	HTK	30 min	Segmentation & Intonation break	HTK-HTS, new vocoder	Own	context-dependent MSD-HSMMs	-	-
AhoTTS	60	HTK	30 min transcrip. of foreign words	Segmentation & Intonation break	HTK, HTS, Praat, Snack	Own	f0 (Unit Selection + MSD-HSMMs), dur. & intonation break prediction (CART)	spectrum, phonetic context, f0, dur., stress,voiceness position & type of phrase word/syllable boundaries	spectrum, f0, intensity, penaliz. depend. concatenation type
Cotovía	2	Cotovia	None	Phone,syllable,word segment,phrase	Cotovia Matlab	Cotovía	Unit selection	f0, duration, power, phonetic	f0, spectral envelope, power
Cotovía-HTS	48	Cotovia	None	Phone,syllable,word segment,phrase	Cotovia,HTS,openFST SPTK,matlab,Straight	Cotovía	HTS	-	-
GTHCSTR-2008	60	Festival	None	-	Unix core-tools HTS, SPTK, Festival	grapheme to phoneme rules	HSMM context-dependent model	none	none
GTHCSTR-2010	60	Festival	None	re-aligment iterations	Unix core-tools HTS, SPTK, Festival	grapheme to phoneme rules	HSMM context-dependent model	none	none
GTHCSTR-2010-adapt	48	Festival	None	-	Unix core-tools HTS, SPTK, Festival	grapheme to phoneme rules	HSMM context-dependent model	none	none
Ms-HTS-TTS	10	N/A	N/A	N/A	HTK	Own	N/A	N/A	N/A
Ogmios	5	Automatic HMM-based ASR	No	-	Ramses Praat	UPC	Intonation: stylized contours of stress groups. Duration: syl + phoneme CARTs	phonetic, acoustic prosodic info.	segmental, prosodic spectral info.
SalleTTS	40	Praat + HTK + proprietary	No	Phoneme, pitch, duration, energy	Proprietary + Freeling + itpp + wagon	No	CBR	Acoustic + Linguistic	Acoustic

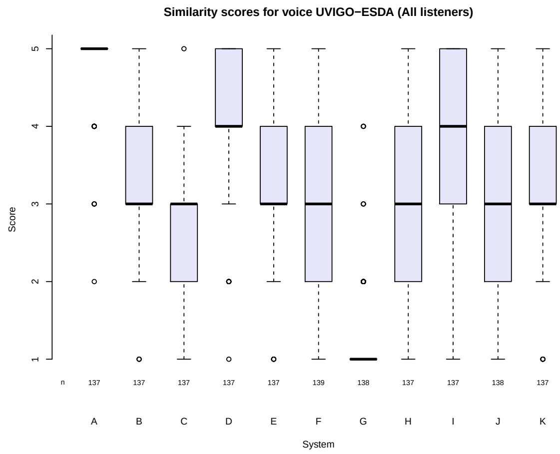
Table 11: *Albayzín 2010 TTS participant questionnaire — Part 2.*

System	Data				Signal Processing				Opinions		
	Extra Data	Pruning	First Spanish	Comments	Spectrum	Source	Pitch marks	Signal modification	Systems's best quality	Strongest components	Weakest components
AhoHTS	-	-	Yes	-	MFCCs	f0	No	HNM-based vocoder	No	-	phrase break insertion
AhoTTS	No	Outlier Penalization	2nd one	-	None	None	Yes, our method	TD-PSOLA	Perhaps	Unit Selection	phrase break insertion
Cotovía	None	MFCC	No	-	MFCC	-	Praat + post-processing	PSOLA	No	Prosody	Segmentation
Cotovía-HTS	None	None	yes	-	Straight	Straight	-	Straight	First voice	Robustness, stability	Muffled speech
GTHCSTR-2008	None	None	No	Albayzín 2008 Best system	Straight MCEP	Mixed:f0 & aperiodicity	voting between 3 F0 estimation	Vocoder	depends	Robustness	Buzzyness
GTHCSTR-2010	None	None	No	-	Straight MCEP	Mixed:f0 & aperiodicity	voting between 3 F0 estimation	Vocoder Postfiltering	depends	Robustness	Buzzyness
GTHCSTR-2010-adapt	Joaquin's voice from SEV Corpus	None	No	adapted to uvigo w. 50 sentences	Straight MCEP	Mixed:f0 & aperiodicity	voting between 3 F0 estimation	Vocoder	depends	Robustness, few amount of data	Buzzyness
Ms-HTS-TTS	LTS, POS tagger, voice font	-	No	-	N/A	N/A	N/A	N/A	No	-	-
Ogmios	other 10h to train prosodic model	segmentation errors 10%	No	-	MFCC	no	Praat	adjust f0 and unit duration	No	None	Concatenation
SalleTTS	No	Yes	No	-	No	No	Praat + Proprietary	PSOLA	No	perceptual weight-tuning	segmentation

Table 12: *Albayzín 2010 TTS participant questionnaire — Part 3.*

B. Results: Similarity to original speaker

- Similarity scores for voice UVIGO-ESDA (All listeners)

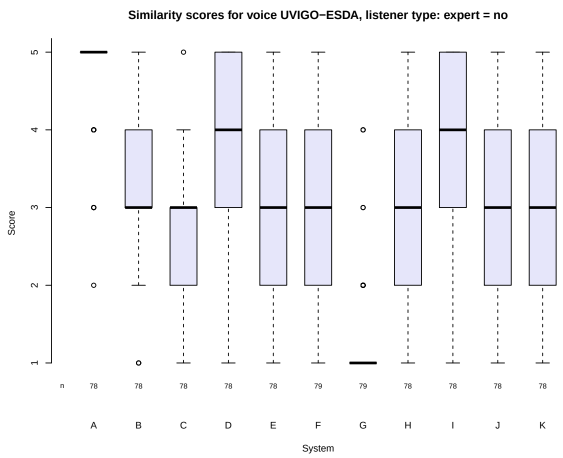


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.87	3.13
C	1	2	3	3	5	2.87	3.13
D	1	4	4	5	5	3.87	4.13
E	1	3	3	4	5	2.87	3.13
F	1	2	3	4	5	2.73	3.27
G	1	1	1	1	4	1.00	1.00
H	1	2	3	4	5	2.73	3.27
I	1	3	4	5	5	3.73	4.27
J	1	2	3	4	5	2.73	3.27
K	1	3	3	4	5	2.87	3.13

	median	MAD	mean	sd	n	na
A	5	0.00	4.83	0.49	137	10
D	4	1.48	4.07	0.94	137	10
I	4	1.48	4.02	0.94	137	10
B	3	1.48	3.34	0.94	137	10
H	3	1.48	3.23	1.14	137	10
K	3	1.48	3.20	0.99	137	10
E	3	1.48	3.15	0.97	137	10
J	3	1.48	3.13	1.11	138	9
F	3	1.48	2.91	1.12	139	8
C	3	1.48	2.54	0.96	137	10
G	1	0.00	1.25	0.60	138	9

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•		•	•		•		
C	•	•		•	•		•	•	•	•	•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•	•		•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•		•	•
J	•		•	•			•		•		
K	•		•	•			•		•		

- Similarity scores for voice UVIGO-ESDA, listener type: expert = no

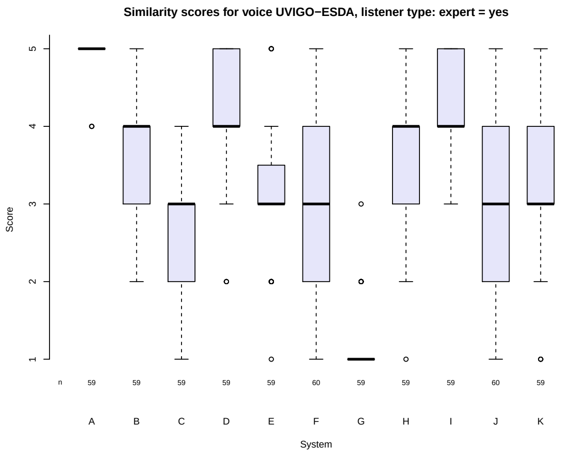


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.82	3.18
C	1	2	3	3	5	2.82	3.18
D	1	3	4	5	5	3.64	4.36
E	1	2	3	4	5	2.64	3.36
F	1	2	3	4	5	2.64	3.36
G	1	1	1	1	4	1.00	1.00
H	1	2	3	4	5	2.64	3.36
I	1	3	4	5	5	3.64	4.36
J	1	2	3	4	5	2.64	3.36
K	1	2	3	4	5	2.64	3.36

	median	MAD	mean	sd	n	na
A	5	0.00	4.74	0.61	78	5
D	4	1.48	3.99	1.03	78	5
I	4	1.48	3.83	1.00	78	5
B	3	1.48	3.24	0.98	78	5
K	3	1.48	3.21	1.02	78	5
E	3	1.48	3.18	1.04	78	5
J	3	1.48	3.14	1.17	78	5
H	3	1.48	3.14	1.27	78	5
F	3	1.48	2.90	1.17	79	4
C	3	1.48	2.55	1.04	78	5
G	1	0.00	1.30	0.70	79	4

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•			•		•		
C	•	•		•	•		•	•	•	•	•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•			•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•			•
J	•		•	•			•				
K	•		•	•			•		•		

- Similarity scores for voice UVIGO-ESDA, listener type: expert = yes

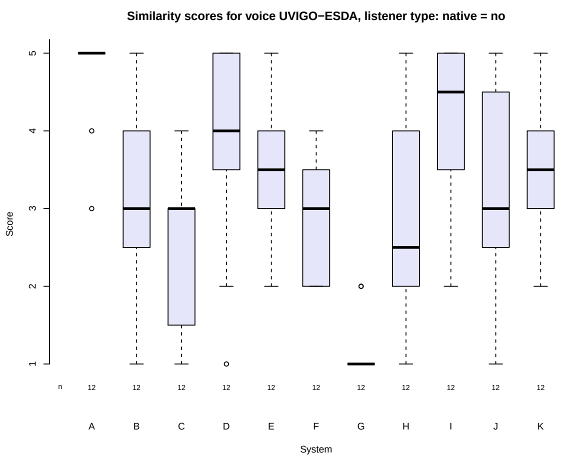


	Min	1Q	Med	3Q	Max	LCL	UCL
A	4	5	5	5	5	5.00	5.00
B	2	3	4	4	5	3.79	4.21
C	1	2	3	3	4	2.79	3.21
D	2	4	4	5	5	3.79	4.21
E	1	3	3	4	5	2.90	3.10
F	1	2	3	4	5	2.59	3.41
G	1	1	1	1	3	1.00	1.00
H	1	3	4	4	5	3.79	4.21
I	3	4	4	5	5	3.79	4.21
J	1	2	3	4	5	2.59	3.41
K	1	3	3	4	5	2.79	3.21

	median	MAD	mean	sd	n	na
A	5	0.00	4.95	0.22	59	5
I	4	1.48	4.27	0.78	59	5
D	4	1.48	4.19	0.82	59	5
B	4	1.48	3.47	0.88	59	5
H	4	1.48	3.34	0.94	59	5
K	3	1.48	3.20	0.96	59	5
J	3	1.48	3.12	1.03	60	4
E	3	0.00	3.10	0.88	59	5
F	3	1.48	2.92	1.05	60	4
C	3	1.48	2.53	0.86	59	5
G	1	0.00	1.19	0.43	59	5

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•			•		•		
C	•	•		•	•		•	•	•	•	•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•			•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•		•	•
J	•		•	•			•		•		
K	•		•	•			•		•		

- Similarity scores for voice UVIGO-ESDA, listener type: native = no

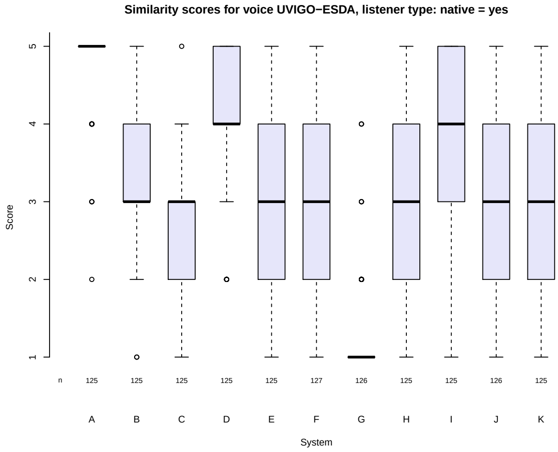


	Min	1Q	Med	3Q	Max	LCL	UCL
A	3	5	5	5	5	5.00	5.00
B	1	2	3	4	5	2.32	3.68
C	1	2	3	3	4	2.32	3.68
D	1	4	4	5	5	3.32	4.68
E	2	3	4	4	5	3.04	3.96
F	2	2	3	4	4	2.32	3.68
G	1	1	1	1	2	1.00	1.00
H	1	2	2	4	5	1.59	3.41
I	2	4	4	5	5	3.82	5.18
J	1	2	3	4	5	2.09	3.91
K	2	3	4	4	5	3.04	3.96

	median	MAD	mean	sd	n	na
A	5	0.00	4.75	0.62	12	1
I	4	0.74	4.08	1.16	12	1
D	4	1.48	3.92	1.31	12	1
K	4	0.74	3.50	0.80	12	1
E	4	0.74	3.42	0.90	12	1
J	3	1.48	3.33	1.30	12	1
B	3	1.48	3.17	1.11	12	1
F	3	1.48	2.92	0.79	12	1
H	2	1.48	2.75	1.29	12	1
C	3	1.48	2.50	1.09	12	1
G	1	0.00	1.17	0.39	12	1

	A	B	C	D	E	F	G	H	I	J	K
A											
B											
C											
D											
E											
F											
G											
H											
I											
J											
K											

- Similarity scores for voice UVIGO-ESDA, listener type: native = yes

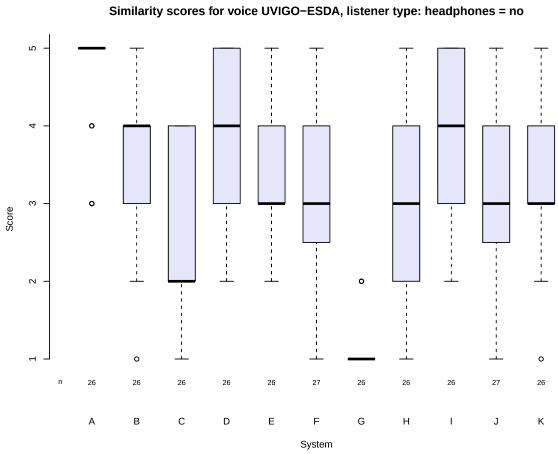


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.86	3.14
C	1	2	3	3	5	2.86	3.14
D	2	4	4	5	5	3.86	4.14
E	1	2	3	4	5	2.72	3.28
F	1	2	3	4	5	2.72	3.28
G	1	1	1	1	4	1.00	1.00
H	1	2	3	4	5	2.72	3.28
I	1	3	4	5	5	3.72	4.28
J	1	2	3	4	5	2.72	3.28
K	1	2	3	4	5	2.72	3.28

	median	MAD	mean	sd	n	na
A	5	0.00	4.84	0.48	125	9
D	4	1.48	4.09	0.91	125	9
I	4	1.48	4.02	0.92	125	9
B	3	1.48	3.36	0.93	125	9
H	3	1.48	3.27	1.12	125	9
K	3	1.48	3.18	1.01	125	9
E	3	1.48	3.12	0.98	125	9
J	3	1.48	3.11	1.09	126	8
F	3	1.48	2.91	1.14	127	7
C	3	1.48	2.54	0.95	125	9
G	1	0.00	1.26	0.62	126	8

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•			•		•		
C	•	•		•	•		•	•	•	•	•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•			•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•		•	•
J	•		•	•			•		•		
K	•		•	•			•		•		

- Similarity scores for voice UVIGO-ESDA, listener type: headphones = no

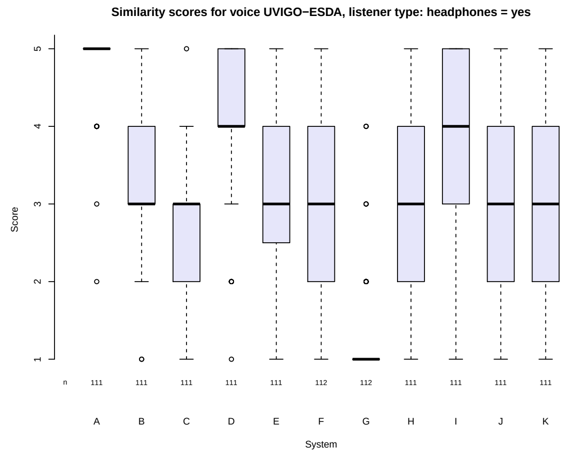


	Min	1Q	Med	3Q	Max	LCL	UCL
A	3	5	5	5	5	5.00	5.00
B	1	3	4	4	5	3.69	4.31
C	1	2	2	4	4	1.38	2.62
D	2	3	4	5	5	3.38	4.62
E	2	3	3	4	5	2.69	3.31
F	1	2	3	4	5	2.54	3.46
G	1	1	1	1	2	1.00	1.00
H	1	2	3	4	5	2.38	3.62
I	2	3	4	5	5	3.38	4.62
J	1	2	3	4	5	2.54	3.46
K	1	3	3	4	5	2.69	3.31

	median	MAD	mean	sd	n	na
A	5	0.00	4.65	0.69	26	2
D	4	1.48	3.96	1.00	26	2
I	4	1.48	3.92	0.98	26	2
B	4	1.48	3.50	0.99	26	2
E	3	1.48	3.31	1.01	26	2
J	3	1.48	3.26	1.16	27	1
K	3	0.74	3.23	0.91	26	2
H	3	1.48	3.23	1.07	26	2
F	3	1.48	3.15	1.06	27	1
C	2	1.48	2.54	1.07	26	2
G	1	0.00	1.15	0.37	26	2

	A	B	C	D	E	F	G	H	I	J	K
A		•	•		•	•	•	•			•
B	•						•				
C	•						•				
D							•				
E	•						•				
F	•						•				
G	•	•	•	•	•	•		•	•	•	•
H	•						•				
I							•				
J							•				
K	•						•				

- Similarity scores for voice UVIGO-ESDA, listener type: headphones = yes

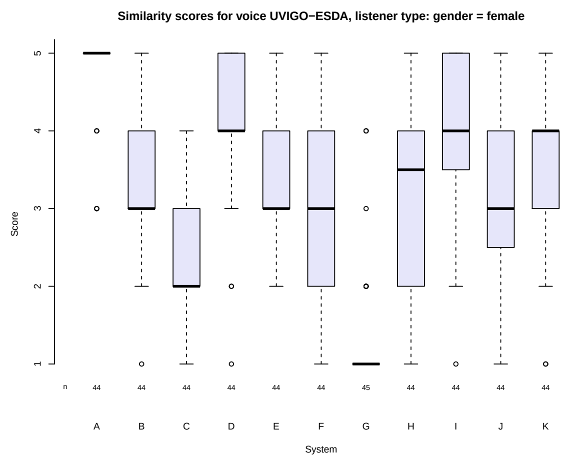


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.85	3.15
C	1	2	3	3	5	2.85	3.15
D	1	4	4	5	5	3.85	4.15
E	1	2	3	4	5	2.78	3.22
F	1	2	3	4	5	2.70	3.30
G	1	1	1	1	4	1.00	1.00
H	1	2	3	4	5	2.70	3.30
I	1	3	4	5	5	3.70	4.30
J	1	2	3	4	5	2.70	3.30
K	1	2	3	4	5	2.70	3.30

	median	MAD	mean	sd	n	na
A	5	0.00	4.87	0.43	111	8
D	4	1.48	4.10	0.93	111	8
I	4	1.48	4.05	0.93	111	8
B	3	1.48	3.31	0.93	111	8
H	3	1.48	3.23	1.16	111	8
K	3	1.48	3.20	1.02	111	8
E	3	1.48	3.11	0.97	111	8
J	3	1.48	3.10	1.10	111	8
F	3	1.48	2.85	1.12	112	7
C	3	1.48	2.54	0.94	111	8
G	1	0.00	1.28	0.65	112	7

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•			•		•		
C	•	•		•	•		•	•	•	•	•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•			•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•		•	•
J	•		•	•			•		•		
K	•		•	•			•		•		

- Similarity scores for voice UVIGO-ESDA, listener type: gender = female

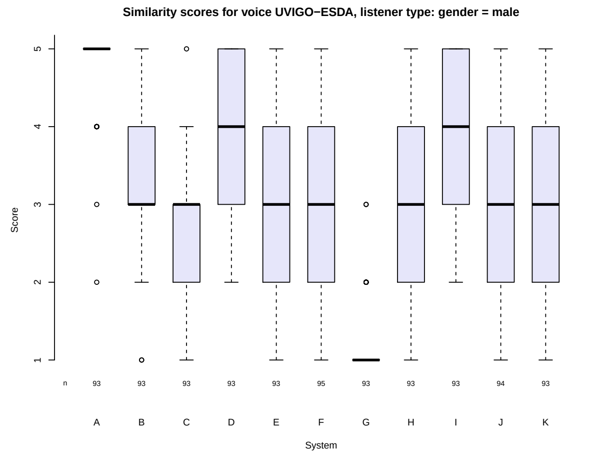


	Min	1Q	Med	3Q	Max	LCL	UCL
A	3	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.76	3.24
C	1	2	2	3	4	1.76	2.24
D	1	4	4	5	5	3.76	4.24
E	2	3	3	4	5	2.76	3.24
F	1	2	3	4	5	2.52	3.48
G	1	1	1	1	4	1.00	1.00
H	1	2	4	4	5	3.02	3.98
I	1	4	4	5	5	3.64	4.36
J	1	2	3	4	5	2.64	3.36
K	1	3	4	4	5	3.76	4.24

	median	MAD	mean	sd	n	na
A	5	0.00	4.80	0.55	44	5
D	4	1.48	4.20	0.98	44	5
I	4	1.48	4.07	1.00	44	5
K	4	1.48	3.52	0.95	44	5
J	3	1.48	3.41	1.19	44	5
E	3	1.48	3.36	0.92	44	5
B	3	1.48	3.25	0.94	44	5
H	4	1.48	3.25	1.30	44	5
F	3	1.48	3.11	1.20	44	5
C	2	1.48	2.50	1.05	44	5
G	1	0.00	1.38	0.83	45	4

	A	B	C	D	E	F	G	H	I	J	K
A		•	•		•	•	•	•		•	•
B	•			•			•				
C	•			•	•		•		•	•	•
D		•	•			•	•	•			
E	•		•				•				
F	•			•			•				
G	•	•	•	•	•	•		•	•	•	•
H	•			•			•				
I			•				•				
J	•		•				•				
K	•		•				•				

- Similarity scores for voice UVIGO-ESDA, listener type: gender = male



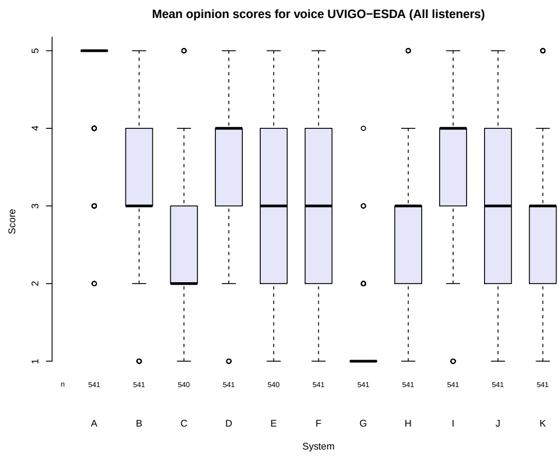
	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.84	3.16
C	1	2	3	3	5	2.84	3.16
D	2	3	4	5	5	3.67	4.33
E	1	2	3	4	5	2.67	3.33
F	1	2	3	4	5	2.68	3.32
G	1	1	1	1	3	1.00	1.00
H	1	2	3	4	5	2.67	3.33
I	2	3	4	5	5	3.67	4.33
J	1	2	3	4	5	2.67	3.33
K	1	2	3	4	5	2.67	3.33

	median	MAD	mean	sd	n	na
A	5	0.00	4.85	0.46	93	5
D	4	1.48	4.01	0.93	93	5
I	4	1.48	4.00	0.91	93	5
B	3	1.48	3.39	0.94	93	5
H	3	1.48	3.22	1.06	93	5
K	3	1.48	3.05	0.98	93	5
E	3	1.48	3.04	0.99	93	5
J	3	1.48	3.00	1.05	94	4
F	3	1.48	2.81	1.06	95	3
C	3	1.48	2.56	0.93	93	5
G	1	0.00	1.19	0.45	93	5

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•		•	•		•		
C	•	•		•	•		•	•	•		•
D	•	•	•		•	•	•	•		•	•
E	•		•	•			•		•		
F	•	•		•			•		•		
G	•	•	•	•	•	•		•	•	•	•
H	•		•	•			•		•		
I	•	•	•		•	•	•	•		•	•
J	•			•			•		•		
K	•		•	•			•		•		

C. Results: Mean opinion scores

- Mean opinion scores for voice UVIGO-ESDA (All listeners)

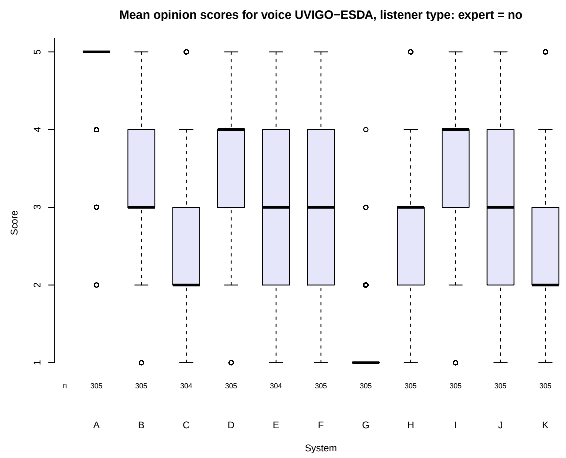


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.93	3.07
C	1	2	2	3	5	1.93	2.07
D	1	3	4	4	5	3.93	4.07
E	1	2	3	4	5	2.86	3.14
F	1	2	3	4	5	2.86	3.14
G	1	1	1	1	4	1.00	1.00
H	1	2	3	3	5	2.93	3.07
I	1	3	4	4	5	3.93	4.07
J	1	2	3	4	5	2.86	3.14
K	1	2	3	3	5	2.93	3.07

	median	MAD	mean	sd	n	na
A	5	0.00	4.75	0.57	541	47
D	4	1.48	3.78	0.98	541	47
I	4	1.48	3.50	1.02	541	47
B	3	1.48	3.33	0.95	541	47
F	3	1.48	3.15	1.00	541	47
E	3	1.48	3.10	0.96	540	48
J	3	1.48	2.91	1.00	541	47
K	3	1.48	2.62	0.98	541	47
H	3	1.48	2.60	0.97	541	47
C	2	1.48	2.51	0.90	540	48
G	1	0.00	1.10	0.34	541	47

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•	•	•	•	•	•	•	•
C	•	•		•	•	•	•		•	•	
D	•	•	•		•	•	•	•	•	•	•
E	•	•	•	•			•	•	•	•	•
F	•	•	•	•			•	•	•	•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•	•	•	•	•	•	•	•		•	•
J	•	•	•	•	•	•	•	•	•		•
K	•	•		•	•	•	•		•	•	

- Mean opinion scores for voice UVIGO-ESDA, listener type: expert = no

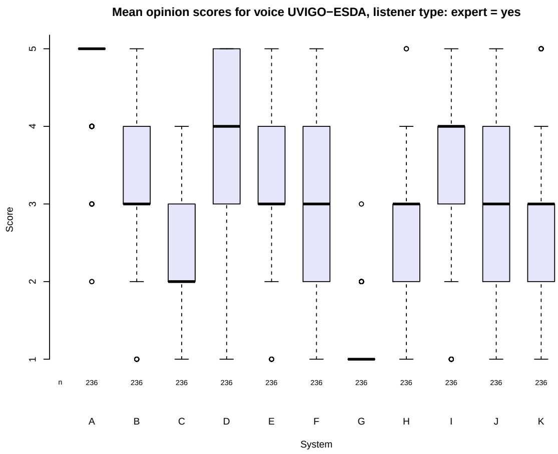


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.91	3.09
C	1	2	2	3	5	1.91	2.09
D	1	3	4	4	5	3.91	4.09
E	1	2	3	4	5	2.82	3.18
F	1	2	3	4	5	2.82	3.18
G	1	1	1	1	4	1.00	1.00
H	1	2	3	3	5	2.91	3.09
I	1	3	4	4	5	3.91	4.09
J	1	2	3	4	5	2.82	3.18
K	1	2	2	3	5	1.91	2.09

	median	MAD	mean	sd	n	na
A	5	0.00	4.70	0.59	305	27
D	4	1.48	3.69	0.97	305	27
I	4	1.48	3.40	1.00	305	27
B	3	1.48	3.25	0.95	305	27
F	3	1.48	3.16	1.00	305	27
E	3	1.48	2.98	0.98	304	28
J	3	1.48	2.92	1.02	305	27
H	3	1.48	2.55	1.01	305	27
C	2	1.48	2.53	0.98	304	28
K	2	1.48	2.52	0.97	305	27
G	1	0.00	1.12	0.39	305	27

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•	•		•	•		•	•
C	•	•		•	•	•	•		•	•	
D	•	•	•		•	•	•	•	•	•	•
E	•	•	•	•		•	•	•	•		•
F	•		•	•	•		•	•	•	•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•		•	•	•	•	•	•		•	•
J	•	•	•	•		•	•	•	•		•
K	•	•		•	•	•	•		•	•	

- Mean opinion scores for voice UVIGO-ESDA, listener type: expert = yes

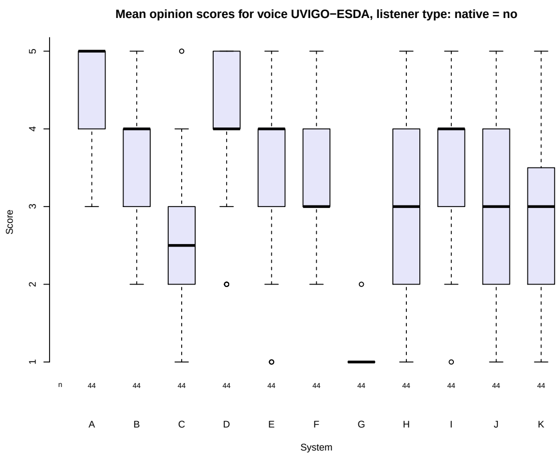


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.90	3.10
C	1	2	2	3	4	1.90	2.10
D	1	3	4	5	5	3.79	4.21
E	1	3	3	4	5	2.90	3.10
F	1	2	3	4	5	2.79	3.21
G	1	1	1	1	3	1.00	1.00
H	1	2	3	3	5	2.90	3.10
I	1	3	4	4	5	3.90	4.10
J	1	2	3	4	5	2.79	3.21
K	1	2	3	3	5	2.90	3.10

	median	MAD	mean	sd	n	na
A	5	0.00	4.81	0.52	236	20
D	4	1.48	3.89	0.98	236	20
I	4	1.48	3.62	1.05	236	20
B	3	1.48	3.42	0.94	236	20
E	3	1.48	3.26	0.91	236	20
F	3	1.48	3.13	1.00	236	20
J	3	1.48	2.90	0.97	236	20
K	3	1.48	2.75	0.97	236	20
H	3	1.48	2.66	0.92	236	20
C	2	1.48	2.49	0.79	236	20
G	1	0.00	1.08	0.28	236	20

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•		•	•	•		•	•
C	•	•		•	•	•	•		•	•	•
D	•	•	•		•	•	•	•	•	•	•
E	•		•	•			•	•	•	•	•
F	•	•	•	•			•	•	•	•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•		
I	•		•	•	•	•	•	•		•	•
J	•	•	•	•	•	•	•		•		
K	•	•	•	•	•	•	•		•		

- Mean opinion scores for voice UVIGO-ESDA, listener type: native = no

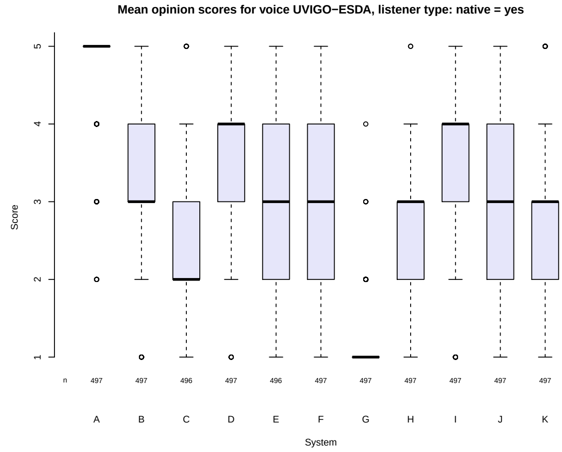


	Min	1Q	Med	3Q	Max	LCL	UCL
A	3	4	5	5	5	4.76	5.24
B	2	3	4	4	5	3.76	4.24
C	1	2	2	3	5	2.26	2.74
D	2	4	4	5	5	3.76	4.24
E	1	3	4	4	5	3.76	4.24
F	2	3	3	4	5	2.76	3.24
G	1	1	1	1	2	1.00	1.00
H	1	2	3	4	5	2.52	3.48
I	1	3	4	4	5	3.76	4.24
J	1	2	3	4	5	2.52	3.48
K	1	2	3	4	5	2.64	3.36

	median	MAD	mean	sd	n	na
A	5	0.00	4.64	0.61	44	8
D	4	1.48	4.14	0.93	44	8
I	4	1.48	3.66	0.99	44	8
B	4	1.48	3.55	0.87	44	8
E	4	1.48	3.36	1.12	44	8
F	3	1.48	3.25	0.92	44	8
J	3	1.48	3.02	0.95	44	8
H	3	1.48	2.82	1.26	44	8
K	3	1.48	2.73	1.11	44	8
C	2	0.74	2.66	0.89	44	8
G	1	0.00	1.02	0.15	44	8

	A	B	C	D	E	F	G	H	I	J	K
A		•	•		•	•	•	•	•	•	•
B	•		•	•			•	•			•
C	•	•		•	•	•	•		•		
D		•	•		•	•	•	•		•	•
E	•		•	•			•				•
F	•		•	•			•				
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•			•		•		
I	•		•				•	•			•
J	•			•			•				
K	•	•		•	•		•		•		

- Mean opinion scores for voice UVIGO-ESDA, listener type: native = yes

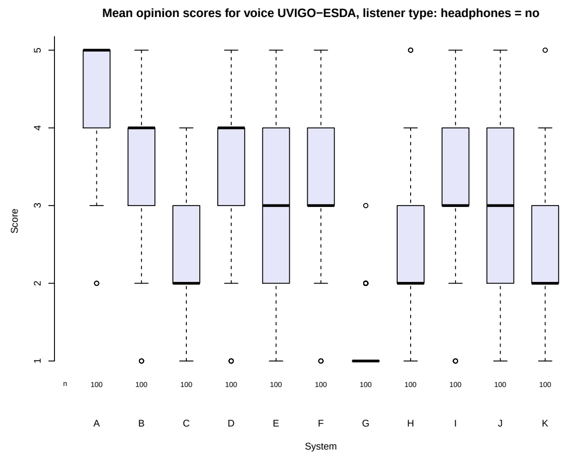


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.93	3.07
C	1	2	2	3	5	1.93	2.07
D	1	3	4	4	5	3.93	4.07
E	1	2	3	4	5	2.86	3.14
F	1	2	3	4	5	2.86	3.14
G	1	1	1	1	4	1.00	1.00
H	1	2	3	3	5	2.93	3.07
I	1	3	4	4	5	3.93	4.07
J	1	2	3	4	5	2.86	3.14
K	1	2	3	3	5	2.93	3.07

	median	MAD	mean	sd	n	na
A	5	0.00	4.76	0.56	497	39
D	4	1.48	3.75	0.98	497	39
I	4	1.48	3.48	1.03	497	39
B	3	1.48	3.31	0.95	497	39
F	3	1.48	3.14	1.01	497	39
E	3	1.48	3.08	0.94	496	40
J	3	1.48	2.90	1.01	497	39
K	3	1.48	2.61	0.97	497	39
H	3	1.48	2.58	0.94	497	39
C	2	1.48	2.50	0.90	496	40
G	1	0.00	1.11	0.36	497	39

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•	•	•	•	•	•	•	•
C	•	•		•	•	•	•		•	•	
D	•	•	•		•	•	•	•	•	•	•
E	•	•	•	•			•	•	•	•	•
F	•	•	•	•			•	•	•	•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•	•	•	•	•	•	•	•		•	•
J	•	•	•	•	•	•	•	•	•		•
K	•	•		•	•	•	•		•	•	

- Mean opinion scores for voice UVIGO-ESDA, listener type: headphones = no

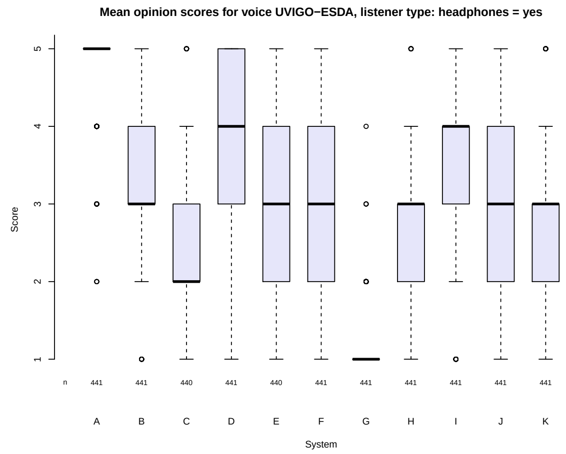


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	4	5	5	5	4.84	5.16
B	1	3	4	4	5	3.84	4.16
C	1	2	2	3	4	1.84	2.16
D	1	3	4	4	5	3.84	4.16
E	1	2	3	4	5	2.68	3.32
F	1	3	3	4	5	2.84	3.16
G	1	1	1	1	3	1.00	1.00
H	1	2	2	3	5	1.84	2.16
I	1	3	3	4	5	2.84	3.16
J	1	2	3	4	5	2.68	3.32
K	1	2	2	3	5	1.84	2.16

	median	MAD	mean	sd	n	na
A	5	0.00	4.59	0.73	100	12
B	4	1.48	3.40	0.93	100	12
D	4	1.48	3.40	1.06	100	12
F	3	1.48	3.31	0.95	100	12
I	3	1.48	3.23	0.97	100	12
E	3	1.48	3.09	1.09	100	12
J	3	1.48	2.88	1.05	100	12
C	2	1.48	2.41	0.88	100	12
K	2	1.48	2.36	0.95	100	12
H	2	1.48	2.31	0.97	100	12
G	1	0.00	1.15	0.39	100	12

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•				•	•		•	•
C	•	•		•	•	•	•		•	•	
D	•		•				•	•		•	•
E	•		•				•	•			•
F	•		•				•	•		•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•		•				•	•			•
J	•	•	•	•		•	•	•			•
K	•	•		•	•	•	•		•	•	

- Mean opinion scores for voice UVIGO-ESDA, listener type: headphones = yes

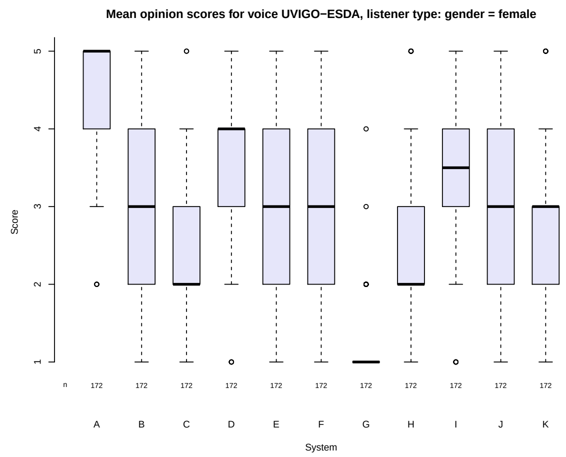


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.92	3.08
C	1	2	2	3	5	1.92	2.08
D	1	3	4	5	5	3.85	4.15
E	1	2	3	4	5	2.85	3.15
F	1	2	3	4	5	2.85	3.15
G	1	1	1	1	4	1.00	1.00
H	1	2	3	3	5	2.92	3.08
I	1	3	4	4	5	3.92	4.08
J	1	2	3	4	5	2.85	3.15
K	1	2	3	3	5	2.92	3.08

	median	MAD	mean	sd	n	na
A	5	0.00	4.79	0.52	441	35
D	4	1.48	3.86	0.94	441	35
I	4	1.48	3.56	1.03	441	35
B	3	1.48	3.31	0.95	441	35
F	3	1.48	3.11	1.01	441	35
E	3	1.48	3.11	0.93	440	36
J	3	1.48	2.92	0.99	441	35
K	3	1.48	2.68	0.98	441	35
H	3	1.48	2.66	0.96	441	35
C	2	1.48	2.54	0.90	440	36
G	1	0.00	1.09	0.33	441	35

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•	•	•	•	•	•	•	•	•
C	•	•		•	•	•	•		•	•	
D	•	•	•		•	•	•	•	•	•	•
E	•	•	•	•			•	•	•	•	•
F	•	•	•	•			•	•	•	•	•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•	•	•	•	•	•	•	•		•	•
J	•	•	•	•	•	•	•	•	•		•
K	•	•		•	•	•	•		•	•	

- Mean opinion scores for voice UVIGO-ESDA, listener type: gender = female

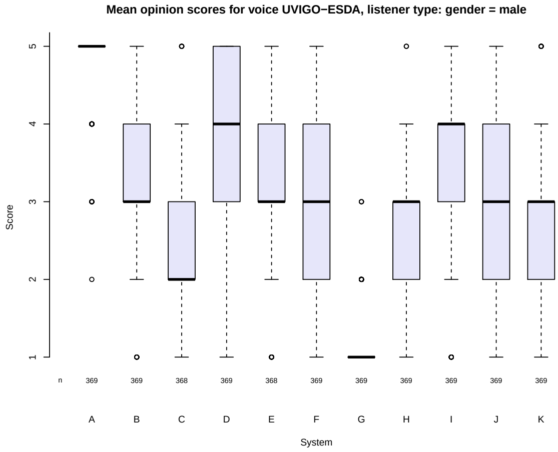


	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	4	5	5	5	4.88	5.12
B	1	2	3	4	5	2.76	3.24
C	1	2	2	3	5	1.88	2.12
D	1	3	4	4	5	3.88	4.12
E	1	2	3	4	5	2.76	3.24
F	1	2	3	4	5	2.76	3.24
G	1	1	1	1	4	1.00	1.00
H	1	2	2	3	5	1.88	2.12
I	1	3	4	4	5	3.38	3.62
J	1	2	3	4	5	2.76	3.24
K	1	2	3	3	5	2.88	3.12

	median	MAD	mean	sd	n	na
A	5	0.00	4.63	0.72	172	24
D	4	1.48	3.56	1.10	172	24
I	4	0.74	3.34	1.07	172	24
B	3	1.48	3.30	1.05	172	24
F	3	1.48	3.17	0.99	172	24
E	3	1.48	3.05	1.04	172	24
J	3	1.48	2.91	1.07	172	24
K	3	1.48	2.62	1.06	172	24
H	2	1.48	2.55	1.10	172	24
C	2	1.48	2.54	0.96	172	24
G	1	0.00	1.10	0.38	172	24

	A	B	C	D	E	F	G	H	I	J	K
A		•	•	•	•	•	•	•	•	•	•
B	•		•				•	•		•	•
C	•	•		•	•	•	•		•	•	
D	•		•		•	•	•	•		•	•
E	•		•	•			•	•			•
F	•		•	•			•	•			•
G	•	•	•	•	•	•		•	•	•	•
H	•	•		•	•	•	•		•	•	
I	•		•				•	•		•	•
J	•	•	•	•			•	•	•		
K	•	•		•	•	•	•		•		

- Mean opinion scores for voice UVIGO-ESDA, listener type: gender = male



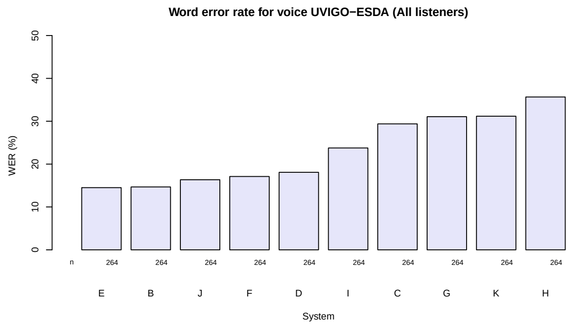
	Min	1Q	Med	3Q	Max	LCL	UCL
A	2	5	5	5	5	5.00	5.00
B	1	3	3	4	5	2.92	3.08
C	1	2	2	3	5	1.92	2.08
D	1	3	4	5	5	3.84	4.16
E	1	3	3	4	5	2.92	3.08
F	1	2	3	4	5	2.84	3.16
G	1	1	1	1	3	1.00	1.00
H	1	2	3	3	5	2.92	3.08
I	1	3	4	4	5	3.92	4.08
J	1	2	3	4	5	2.84	3.16
K	1	2	3	3	5	2.92	3.08

	median	MAD	mean	sd	n	na
A	5	0.00	4.81	0.47	369	23
D	4	1.48	3.88	0.90	369	23
I	4	1.48	3.57	1.00	369	23
B	3	1.48	3.34	0.90	369	23
F	3	1.48	3.14	1.01	369	23
E	3	1.48	3.13	0.91	368	24
J	3	1.48	2.91	0.97	369	23
H	3	1.48	2.62	0.90	369	23
K	3	1.48	2.62	0.94	369	23
C	2	1.48	2.50	0.87	368	24
G	1	0.00	1.10	0.33	369	23

	A	B	C	D	E	F	G	H	I	J	K
A		●	●	●	●	●	●	●	●	●	●
B	●		●	●	●	●	●	●	●	●	●
C	●	●		●	●	●	●		●	●	
D	●	●	●		●	●	●	●	●	●	●
E	●	●	●	●			●	●	●	●	●
F	●	●	●	●			●	●	●	●	●
G	●	●	●	●	●	●		●	●	●	●
H	●	●		●	●	●	●		●	●	
I	●	●	●	●	●	●	●	●		●	●
J	●	●	●	●	●	●	●	●	●		●
K	●	●		●	●	●	●		●	●	

D. Results: Word error rates for SUS test.

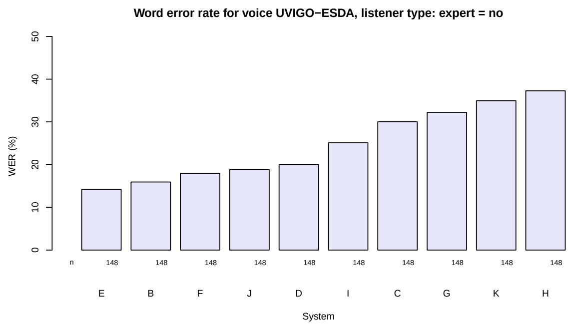
- Word error rate for voice UVIGO-ESDA (All listeners)



	median	MAD	mean	sd	n	na
E	0.14	0.21	0.15	0.18	264	30
B	0.00	0.00	0.15	0.21	264	30
J	0.14	0.21	0.16	0.21	264	30
F	0.14	0.21	0.17	0.19	264	30
D	0.14	0.21	0.18	0.20	264	30
I	0.14	0.21	0.24	0.23	264	30
C	0.29	0.32	0.29	0.25	264	30
G	0.29	0.42	0.31	0.27	264	30
K	0.29	0.21	0.31	0.26	264	30
H	0.29	0.42	0.36	0.27	264	30

	B	C	D	E	F	G	H	I	J	K
B		•				•	•	•		•
C	•		•	•	•		•		•	
D		•				•	•	•		•
E		•				•	•	•		•
F		•				•	•	•		•
G	•		•	•	•			•	•	
H	•	•	•	•	•			•	•	
I	•		•	•	•	•	•		•	•
J		•				•	•	•		•
K	•		•	•	•			•	•	

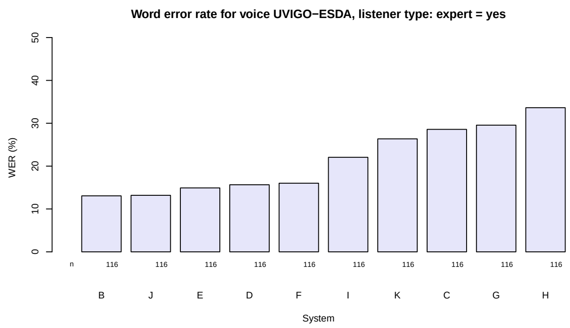
- Word error rate for voice UVIGO-ESDA, listener type: expert = no



	median	MAD	mean	sd	n	na
E	0.00	0.00	0.14	0.18	148	18
B	0.00	0.00	0.16	0.22	148	18
F	0.14	0.21	0.18	0.18	148	18
J	0.14	0.21	0.19	0.24	148	18
D	0.14	0.21	0.20	0.22	148	18
I	0.29	0.21	0.25	0.23	148	18
C	0.29	0.21	0.30	0.25	148	18
G	0.29	0.42	0.32	0.27	148	18
K	0.43	0.42	0.35	0.27	148	18
H	0.43	0.21	0.37	0.27	148	18

	B	C	D	E	F	G	H	I	J	K
B		•				•	•	•		•
C	•		•	•	•				•	
D		•				•	•			•
E		•				•	•	•		•
F		•				•	•	•		•
G	•		•	•	•				•	
H	•		•	•	•			•	•	
I	•			•	•		•			•
J		•				•	•			•
K	•		•	•	•			•	•	

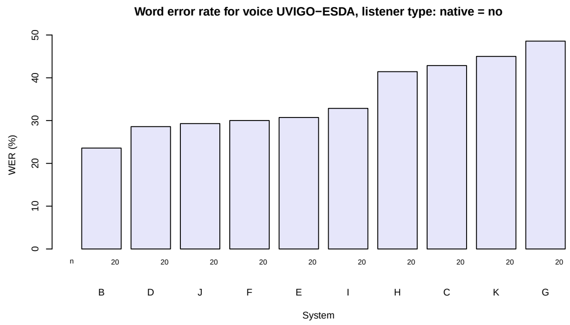
- Word error rate for voice UVIGO-ESDA, listener type: expert = yes



	median	MAD	mean	sd	n	na
B	0.00	0.00	0.13	0.19	116	12
J	0.14	0.21	0.13	0.17	116	12
E	0.14	0.21	0.15	0.18	116	12
D	0.14	0.21	0.16	0.17	116	12
F	0.14	0.21	0.16	0.21	116	12
I	0.14	0.21	0.22	0.22	116	12
K	0.14	0.21	0.26	0.24	116	12
C	0.29	0.42	0.29	0.25	116	12
G	0.29	0.42	0.30	0.27	116	12
H	0.29	0.42	0.34	0.28	116	12

	B	C	D	E	F	G	H	I	J	K
B		•				•	•	•		•
C	•		•	•	•				•	
D		•				•	•			•
E		•				•	•	•		•
F		•				•	•			•
G	•		•	•	•				•	
H	•		•	•	•			•	•	
I	•			•			•		•	
J		•				•	•	•		•
K	•		•	•	•				•	

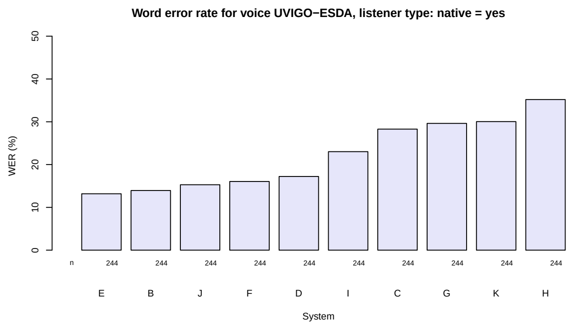
- Word error rate for voice UVIGO-ESDA, listener type: native = no



	median	MAD	mean	sd	n	na
B	0.14	0.21	0.24	0.29	20	6
D	0.29	0.21	0.29	0.28	20	6
J	0.21	0.32	0.29	0.23	20	6
F	0.29	0.21	0.30	0.21	20	6
E	0.29	0.21	0.31	0.24	20	6
I	0.43	0.21	0.33	0.27	20	6
H	0.43	0.42	0.41	0.30	20	6
C	0.43	0.21	0.43	0.21	20	6
K	0.43	0.21	0.45	0.26	20	6
G	0.57	0.32	0.49	0.29	20	6

	B	C	D	E	F	G	H	I	J	K
B										
C										
D										
E										
F										
G										
H										
I										
J										
K										

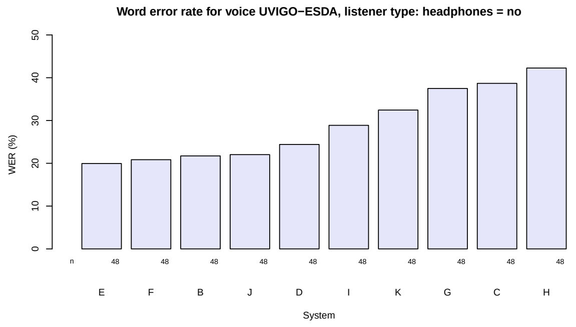
- Word error rate for voice UVIGO-ESDA, listener type: native = yes



	median	MAD	mean	sd	n	na
E	0.00	0.00	0.13	0.17	244	24
B	0.00	0.00	0.14	0.20	244	24
J	0.14	0.21	0.15	0.21	244	24
F	0.14	0.21	0.16	0.19	244	24
D	0.14	0.21	0.17	0.20	244	24
I	0.14	0.21	0.23	0.22	244	24
C	0.29	0.42	0.28	0.25	244	24
G	0.29	0.42	0.30	0.26	244	24
K	0.29	0.21	0.30	0.26	244	24
H	0.29	0.42	0.35	0.27	244	24

	B	C	D	E	F	G	H	I	J	K
B		•				•	•	•		•
C	•		•	•	•		•		•	
D		•				•	•	•		•
E		•				•	•	•		•
F		•				•	•	•		•
G	•		•	•	•			•	•	
H	•	•	•	•	•			•	•	
I	•		•	•	•	•	•		•	•
J		•				•	•	•		•
K	•		•	•	•			•	•	

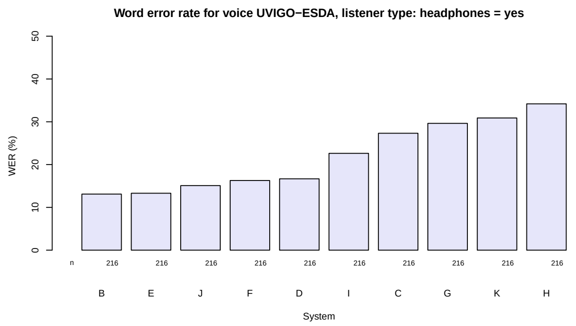
- Word error rate for voice UVIGO-ESDA, listener type: headphones = no



	median	MAD	mean	sd	n	na
E	0.14	0.21	0.20	0.22	48	8
F	0.14	0.21	0.21	0.20	48	8
B	0.14	0.21	0.22	0.22	48	8
J	0.14	0.21	0.22	0.25	48	8
D	0.14	0.21	0.24	0.24	48	8
I	0.29	0.42	0.29	0.25	48	8
K	0.43	0.42	0.32	0.27	48	8
G	0.36	0.53	0.37	0.31	48	8
C	0.43	0.42	0.39	0.28	48	8
H	0.43	0.21	0.42	0.28	48	8

	B	C	D	E	F	G	H	I	J	K
B		•					•			
C	•			•	•				•	
D							•			
E		•				•	•			
F		•				•	•			
G				•	•					
H	•		•	•	•				•	
I										
J		•					•			
K										

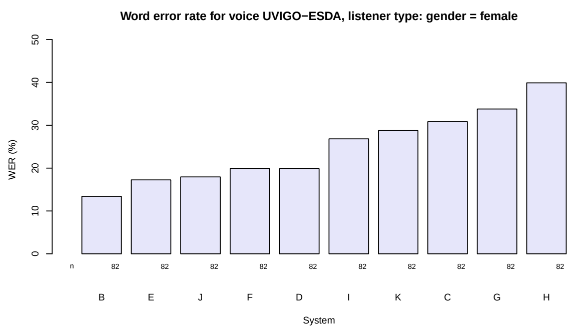
- Word error rate for voice UVIGO-ESDA, listener type: headphones = yes



	median	MAD	mean	sd	n	na
B	0.00	0.00	0.13	0.20	216	22
E	0.00	0.00	0.13	0.17	216	22
J	0.14	0.21	0.15	0.20	216	22
F	0.14	0.21	0.16	0.19	216	22
D	0.14	0.21	0.17	0.19	216	22
I	0.14	0.21	0.23	0.22	216	22
C	0.29	0.21	0.27	0.24	216	22
G	0.29	0.42	0.30	0.26	216	22
K	0.29	0.21	0.31	0.26	216	22
H	0.29	0.42	0.34	0.27	216	22

	B	C	D	E	F	G	H	I	J	K
B		•				•	•	•		•
C	•		•	•	•				•	
D		•				•	•			•
E		•				•	•	•		•
F		•				•	•	•		•
G	•		•	•	•			•	•	
H	•		•	•	•			•	•	
I	•			•	•	•	•		•	•
J		•				•	•	•		•
K	•		•	•	•			•	•	

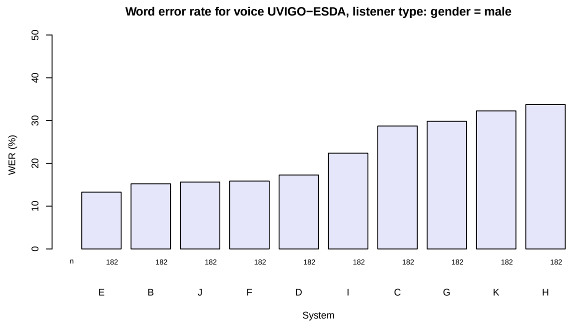
- Word error rate for voice UVIGO-ESDA, listener type: gender = female



	median	MAD	mean	sd	n	na
B	0.00	0.00	0.13	0.18	82	16
E	0.14	0.21	0.17	0.20	82	16
J	0.14	0.21	0.18	0.23	82	16
F	0.14	0.21	0.20	0.20	82	16
D	0.14	0.21	0.20	0.22	82	16
I	0.29	0.21	0.27	0.23	82	16
K	0.21	0.32	0.29	0.26	82	16
C	0.29	0.42	0.31	0.26	82	16
G	0.29	0.42	0.34	0.27	82	16
H	0.43	0.42	0.40	0.28	82	16

	B	C	D	E	F	G	H	I	J	K
B		•				•	•	•		•
C	•		•	•	•				•	
D		•				•	•			
E		•				•	•			•
F		•				•	•			
G	•		•	•	•				•	
H	•		•	•	•			•	•	
I	•						•		•	
J		•				•	•	•		
K	•			•						

- Word error rate for voice UVIGO-ESDA, listener type: gender = male



	median	MAD	mean	sd	n	na
E	0.00	0.00	0.13	0.18	182	14
B	0.00	0.00	0.15	0.22	182	14
J	0.14	0.21	0.16	0.20	182	14
F	0.14	0.21	0.16	0.19	182	14
D	0.14	0.21	0.17	0.20	182	14
I	0.14	0.21	0.22	0.22	182	14
C	0.29	0.21	0.29	0.25	182	14
G	0.29	0.42	0.30	0.27	182	14
K	0.29	0.21	0.32	0.26	182	14
H	0.29	0.42	0.34	0.27	182	14

	B	C	D	E	F	G	H	I	J	K
B		•				•	•	•		•
C	•		•	•	•				•	
D		•				•	•			•
E		•				•	•	•		•
F		•				•	•	•		•
G	•		•	•	•			•	•	
H	•		•	•	•			•	•	
I	•			•	•	•	•		•	•
J		•				•	•	•		•
K	•		•	•	•			•	•	

Albayzín'10 Evaluation: Posters

Aholab Speech Synthesizers for Albayzin2010

I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, I. Odriozola, I. Luengo

Aholab Signal Processing Laboratory, University of the Basque Country, Spain
{inaki, derro, eva, inma, ion, ibon, igor, ikerl}@aholab.ehu.es

Abstract

This paper describes the two Text-to-Speech (TTS) systems presented by Aholab-EHU/UPV in the Albayzin2010 evaluation campaign. The first system is a statistical parametric TTS based on HTS, with the incentive of using our own vocoder. The other one is a hybrid system in which we try to take advantage of the consistency of the statistical averaging and the segmental naturalness of the unit selection approach. It uses the acoustic parameters generated by the statistical system as the target sequence during the unit selection process. Informal listening tests and some objective measures show that adding the Intonation Break information during the voice building process improves the performance of both systems.

Index Terms: speech synthesis, statistical parametric, unit selection, evaluation

1. Introduction

The Albayzin TTS evaluation compares the performance of different TTS systems built with a common Spanish speech database. This year is the second edition of Albayzin, as well as our second participation in it. After seven weeks for voice building, participants are asked to synthesize several hundred test texts that will be evaluated to determine the quality of the synthetic voices in terms of: naturalness, similarity to the original speaker and intelligibility.

AhoTTS [1] is the synthesis platform for commercial and research purposes that Aholab Laboratory has been developing since 1995. It has a modular architecture, and written in C/C++ it is fully functional in both UNIX and Windows operating systems. Up to this date, synthetic voices for Basque, Spanish (Albayzin2008 voice) and English languages have been created.

This paper is organized as follows. First, we describe the two systems presented. In Section 3 the voice building process is explained. The evaluation results are presented and discussed in Section 4. And finally, some conclusions are drawn in Section 5.

2. Systems Overview

In order to take part in Albayzin 2010 evaluation, we have developed two TTS systems: an HMM-based system and a hybrid one [2] [3] [4]. Both systems share the linguistic analysis module. Besides, the parametric output of the statistical system is used as an input of the hybrid system. Therefore, instead of explaining each system on its own, a sequential analysis of the hybrid TTS synthesis process is going to be described in this section.

The architecture of the hybrid is shown in Figure 1. In short, HTS [5] output is used as target prediction in the unit selection module. Pitch and duration predictions from HTS are combined with internal ones and spectrum parameters are used in order to calculate the distance between target and candidate units. Our hybrid approach tries to combine the robustness of the average modelling with the segmental quality of natural speech units.

2.1. Language Processing

This first module performs several language dependent tasks. Text normalization and grapheme to phoneme conversion are conducted by means of rules, whereas POS tagging uses a specific lexicon and some simple disambiguation rules.

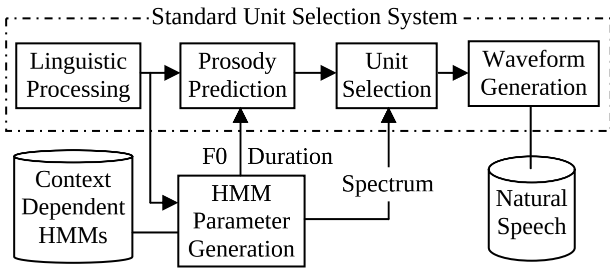


Figure 1: *Hybrid TTS Architecture*

2.2. Speaker-dependent HTS

Aholab had already built an HMM-based TTS system for Basque using HTS [6]. As HTS does not perform any kind of linguistic analysis, the output of the first module of AhoTTS had to be translated into proper labels containing phonetic and linguistic information. Taking into account that Basque language includes all the Spanish phonemes, only minor changes were necessary in order to adapt that system to Spanish (including the incorporation of Intonation Break feature (IB, see section 2.3.1)). The following features have been encoded into the context labels used by HTS:

- **Phoneme level:**
  - SAMPA label of the current phoneme.
  - Labels of 2 phonemes to the right and 2 phonemes to the left.
  - Position of the current phoneme in the current syllable (from the beginning and from the end).
  - Position of the current phoneme after the previous pause and before the next pause.
  - Position of the current phone after the previous IB and before the next IB.
- **Syllable level**:
  - Number of phonemes in current, previous and next syllables.
  - Accent in current, previous and next syllables.
  - Stress in current, previous and next syllables.
  - Position of the current syllable in the current word (from the beginning and from the end).
  - Position of the current syllable in the current accent group.
  - Position of the current syllable in the current sentence.
  - Position of the current syllable after the previous pause and before the next pause.

  - Position of the current syllable after the previous IB and before the next IB.
- **Word level:**
  - Simplified part-of-speech tag of the current, previous and next words (content/function).
  - Number of syllables of the current, previous and next words.
  - Position of the current word in the sentence (from the beginning and from the end).
  - Position of the current word after the previous pause and before the next pause.
  - Position of the current word after the previous IB and before the next IB.
- **Accent level:**
  - Type of current, previous and next accent groups, according to the accent position.
  - Number of syllables in current, previous and next accent groups.
  - Position of the current accent group in the sentence (from the beginning and from the end).
  - Position of the current accent group after the previous pause and before the next pause.
- **Pause context level:**
  - Type of previous and next pauses.
  - Number of pauses to the right and to the left.
- **Sentence level:**
  - Type of sentence.
  - Number of phonemes.
  - Number of syllables.
  - Number of words.
  - Number of accent groups.
  - Number of pauses.

In order to extract the framewise parametric representation of both the spectrum and the excitation, an HNM (Harmonics plus Noise Model) is used [7] that allows the reconstruction of speech too.

2.3. Prosody Prediction

This module performs several sequential tasks as IB insertion, duration prediction and intonation modelling. We have decided not to use our phrasing algorithm because its performance is still poor (too many false insertions that spoil the synthesis output). Therefore, we rely only on orthographic marks to assign phrase breaks. Nevertheless, thanks to the new IB prediction module, the absence of pause breaks in long word sequences is somehow alleviated. Besides, being the IB a more subtle phenomenon than the phrase break is, its miss-insertions are also less disturbing.

2.3.1. Intonation Break

IB is an important phenomenon not only related to the intonation contour (e.g. F0 reset), but also to the duration (e.g. syllable lengthening) and acoustic realization (e.g. relaxed pronunciation) of phonemes adjacent to this event. As the corpus provided by the Albayzin organization included IB labels, we have built a CART that predicts their location from input plain text [8]. Among the features used to accomplish that goal, the following ones can be highlighted: POS in a three word window around current word, and the number of syllables, words and accent groups to previous and next breaks (IB or pause). The IB information is used in both prosody prediction and unit selection acoustic module at several unit levels (phoneme, syllable and word).

2.3.2. Corpus Based Intonation

Our unit selection intonation modelling uses the voiced phoneme as the basic unit in a similar approach to [9]. Such a small unit provides greater flexibility, although the concatenations of non consecutive units inside syllables are significantly restricted. We implement a generic Viterbi search to find the sequence of candidate units from the database that minimizes a function cost composed by the target and concatenation subcosts [10] as shown in the following equations:

$$C(t,u) = \alpha \sum_{i=1}^{n} C^T(t_i, u_i) + (1-\alpha) \sum_{i=1}^{n-1} C^C(u_i, u_{i+1}) \tag{1}$$

$$C^T(t_i, u_i) = \sum_{j=1}^{P} w_j^T C_j^T(t_i, u_i) \tag{2}$$

$$C^C(u_i, u_{i+1}) = \sum_{j=1}^{Q} w_j^C C_j^C(u_i, u_{i+1}) \tag{3}$$

Where t_i are target units and u_i candidate ones. C^T and C^C are the target and concatenation cost respectively; w_j is the j-th weight of the P target subcosts and the Q join subcosts. The main features employed in the target function are these: Type of proposition, Type of Accent Group (AG), Segmental characteristics of neighbouring phonemes, Position (in the AGs, syllable, word and phonic group), Accent, Duration, IB boundary.

Target weights are adjusted using a similar approach to the one proposed in [10] for acoustic unit selection. We first measure the pitch distance between units in the database and the relative distance regarding the adjacent voiced units. Then, we try to predict that distance as the summation of the target subcosts defined above, solving the weights as a multiple linear regression problem.

When two intonation units are not consecutive in the corpus, the following join subcosts are calculated: Pitch difference at the join, Pitch difference among natural neighbours of the units to be concatenated.

Join weights are manually assigned and some penalizations are added in order to hinder the concatenation of non consecutive voiced units inside a syllable, and to a lesser extent, inside an AG. Finally, the intonation contour is combined with the one predicted by the MSD-HSMM (multi-space distribution hidden semi-Markov models) modelling output from HTS. We just perform a weighted linear combination of both pitch contours, after phone alignment and interpolation in unvoiced regions. That way, we try to take advantage of the consistency of the statistical averaging and the segmental naturalness of the unit selection approach.

2.3.3. Duration

CART zscore duration models were trained for voiced and unvoiced consonants, whereas Random Forests [11] were preferred for the vowels. In both cases, the same features were used: phoneme characteristics in a five phoneme window, stress, position (in syllable, word, IB and sentence), simplified POS, etc. Once again, the durations are combined with the ones predicted by HTS.

2.4. Acoustic Engine

Our acoustic engine performs the usual steps in a corpus-based concatenative system: pre-selection of candidate units, a dynamic programming step combining weighted join and target costs, and a concatenation step joining the selected units to form an output speech waveform. Halfphones are selected as the basic unit because of the flexibility they provide to form longer units.

In our hybrid approach, the spectral parameters generated by the statistical parametric synthesis are used as the target during the unit selection process, combined with prosody and linguistic features.

2.4.1. Unit Selection

Target cost function (2) is divided in various subcosts which are calculated at the halfphone level: Phoneme context, Pitch and its slope, Duration, Accent, Type .of proposition, Position. A new subcost is added for the Hybrid System:

- *Spectral Distance*: Frame based Euclidean distance between target (HTS output) and candidate units after DTW [12] alignment. The distance is manually weighted according to three reduced phonetic classes: vowels, voiced and unvoiced consonants.

The concatenation cost function (3) is composed of seven subcosts, all but the *inter-syllable pitch range* being only computed for non-consecutive units: Pitch, Inter-syllable pitch range, Duration, Power, Spectrum, Voiceness, and Penalizations depending on the transition type.

Target weights are adjusted solving a multiple linear regression problem, as stated previously for the pitch modelling. The Euclidean distance of MFCC parameters is used as the predictee and the subcosts as the predictors. Different weights are estimated for left and right halfphones and for each phoneme type. Concatenation weights and α from equation (1) are adjusted manually.

2.4.2. Waveform Generation

The selected candidate units are joined using glottal closure instant information to get smooth concatenations. It is well known that prosody modifications reduce the overall natural quality of the voice. Therefore, only minor prosody modifications are done by means of pitch synchronous overlap and add techniques. The energy is smoothed over non consecutive halfphone transitions and a gain contour is applied in order to normalize the amplitude in the middle of each phoneme.

3. Voice Building

Organizers provided a medium sized (two hours long) speech database [13] recorded at University of Vigo by a male voice talent. The database consists of 1217 phonetically balanced sentences, automatically extracted phone segmentation and IB labels.

The voice building process involves several sequential tasks that are performed almost automatically. After segmentation labels are ready, linguistic and acoustic features are extracted and then, unit selection databases and prosody models are built and weights are trained. The training process of the statistical parametric voice is automatically done, once proper questions to build the trees are set.

3.1. Segmentation

Although the organizers provided segmentation labels, we decided to segment the whole corpus again with HTK toolkit [14]. Before doing so, transcriptions of some foreign words were manually corrected (e.g. West Side Story, pronounced by the speaker as B.w.e.s.t-s.a.j.T-e.s.t.o.r.i). Then, tied-state triphone models were trained and new labels obtained by means of forced alignment. Finally, pause boundaries were automatically refined with a simple processing based on phone duration and energy threshold. No manual revision of the segmentation labels was done.

3.2. Feature Extraction

All the language related features were extracted from our linguistic processing module. The extraction of the acoustic features consists of several steps. First, power normalization is performed by measuring the mean power in the middle of the vowels for all the sentences, and then normalizing each inter-pause interval. Then, pitch contour is detected combining three different methods in order to avoid gross errors (our own PDA (Pitch Detection Algorithm) [15], get_f0 from Snack Toolkit and Praat). HTK is used to generate 13 MFCC parameters calculated with a fixed 5ms frame. As far as the HTS training is concerned, the following parameters are extracted: f0 + 40 MFCCs.

3.3. Impact of IB information

As we had never used IB information during the voice building, a shallow analysis of its impact has been done. Questions related to IB appear in the upper third of all trees trained with HTS: duration, logF0 and MGC (Mel Generalized Cepstrum). The same can be said for the Hybrid TTS: the correlation of the duration prediction improves and IB information is taken into account in the trained weights for prosody and acoustic module (especially at phoneme level and for vowels). Informal listening also revealed that including the IB information had a positive impact in the performance of both systems.

4. Evaluation Results

Each listener completed three evaluation tasks: (i) Mean Opinion Score (MOS) to measure the similarity with the original voice, (ii) naturalness MOS, (iii) and an intelligibility test in which evaluators were asked to transcribe the Semantically Unpredictable Sentences (SUS) they heard. Up to 132 listeners completed the whole test. These were their main characteristics: 44% were experts on speech technologies, 92.5% were native speakers and 82% used headphones. Ten synthetic systems took part in the evaluation (identified with letters B-G). Natural voice (letter A) was also evaluated in order to fix the ceiling score.

In the present section detailed results are shown for our two systems in each of the three evaluation tasks. Unless the contrary is expressed, results from all listeners are analyzed. System ranking or grouping is based on the pairwise Wilcoxon test provided by the organization [16], which is a useful tool to know whether differences among systems are statistically significant or not.

4.1. Similarity Test

It measures the similarity to the original voice in a likert type scale ranging from 1 (*Sounds like a totally different person*) to 5 (*Sounds like exactly the same person*). The results for all the listeners are shown in Figure 2, whereas results for different groups of evaluators are displayed in Figure 3.

4.1.1. Hybrid-TTS

Our Hybrid system (letter D) obtained the best results (4.07 MOS) together with system I. That ranking remains constant for the different listeners' characteristics, although the expert evaluators scored our system a litter higher than the non-experts (4.19 and 3.99 respectively). Being our system a concatenative one (i.e. concatenates segments of natural speech) this high similarity to the original speaker could be expected. However, listeners tend to score not only the segmental similarity but the supra-segmental one (prosody), and concatenation artifacts may play an important role in the subjective evaluation too. In any case, there is still a significant performance gap with respect to the system A (4.83 MOS).

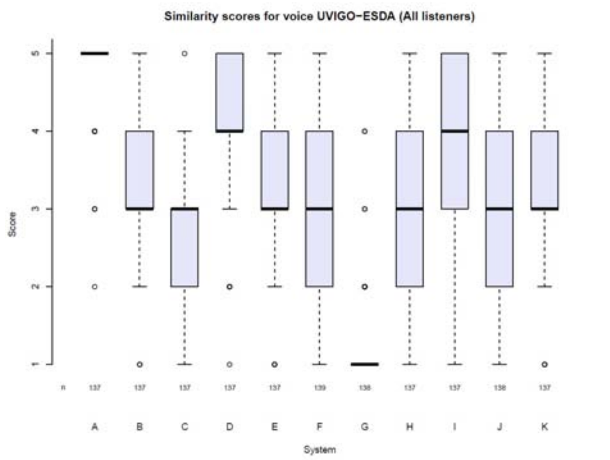


Figure 2: *Similarity to the original voice, all listeners.*

4.1.2. *HTS-based system*

Our system (letter F) gets a MOS of 2.91, and it shares not significant differences with a group of 5 systems (H, K, E, J and C). It seems that the vocoding nature of the system has slightly degraded the similarity to the original voice. Nevertheless, we think that this section of the test is the least important one for typical TTS applications. Besides, it must be stated that the scoring is almost the same for both expert and non-expert evaluators (and the same occurs in the second evaluation task).

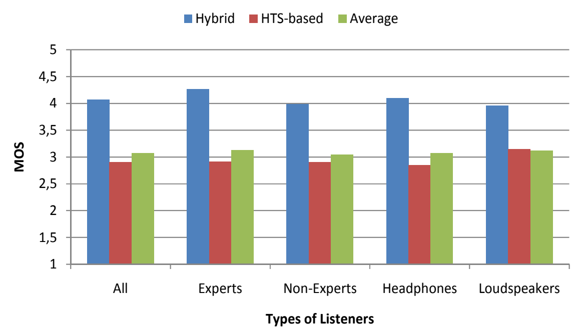


Figure 3: *Similarity to the original voice, different groups.*

4.2. Naturalness Test

It measures the naturalness of the systems in a likert type scale ranging from 1 (*Completely Unnatural*) to 5 (*Completely Natural*). The results are displayed in Figures 4 and 5.

4.2.1. *Hybrid TTS*

Our Hybrid TTS is significantly more natural than the rest of synthetics systems, with a 3.71 MOS. Once again, the scoring of expert evaluators (3.89) is higher than the one from non-experts (3.69). And the gap is even larger between listeners that used headphones (3.86 MOS) or loudspeakers (3.4 MOS). We think that the hybrid approach has succeeded in improving the consistency that unit selection systems usually lack. Just one bad join or incorrectly labeled unit can spoil a whole sentence. Introducing the spectral output of the HMM-based system in the unit selection algorithm has alleviated that problem. Besides, combining two prediction methods produced a more robust prosody.

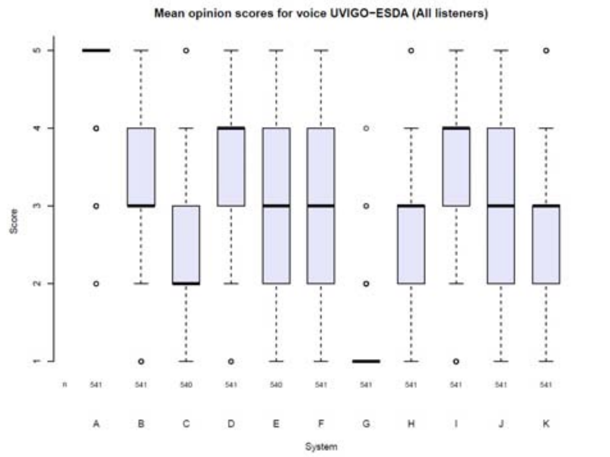


Figure 4: *Naturalness, all listeners.*

4.2.2. *HTS-based system*

Our statistical system obtained a MOS of 3.15, ranking as the fourth best TTS in this task, together with system E. The robustness of the statistical averaging in the modelling process has yielded quite good results. And the same can be said as far as the performance of our own vocoder is concerned.

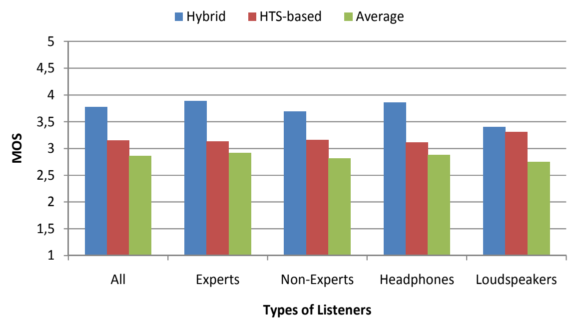


Figure 5: *Naturalness, different groups.*

4.3. Intelligibility Test

The organizers computed Word Error Rates (WER) for SUS as a measure of intelligibility. Unfortunately, no natural speech stimuli were available during the test due to the special structure of sentences needed. So it was not possible to make a comparison between synthetic and natural speech.

Non-native listeners might have introduced some noise in the evaluation (i.e. word errors due to their insufficient knowledge of the language). Therefore, they were not taken into account in the results presented here and displayed in Figure 6. Figure 7 shows WER for different listeners' groups.

4.3.1. *Hybrid TTS*

It achieved a WER of 17%, obtaining the best results together with systems E, B, J and F. The Hybrid approach seems to have alleviated the problems caused by labeling errors or poorly pronounced units, yielding a good performance.

4.3.2. *HTS-based system*

It managed a WER of 16%, obtaining the best results together with systems E, B, J and D. As happened in Albazyn2008 evaluation campaign, statistical modelling has yielded a pretty robust performance.

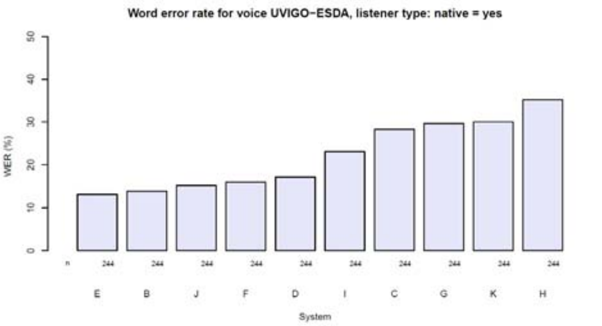


Figure 6: *WER, native listeners.*

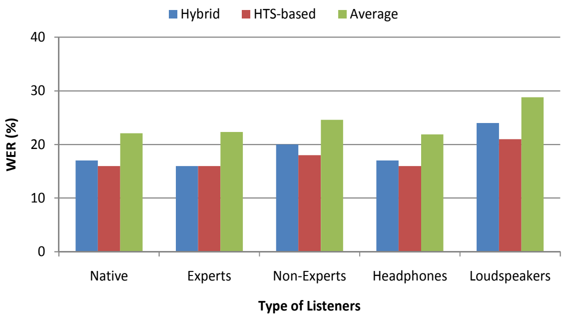


Figure 7: *WER, different groups.*

5. Conclusions

This has been our second participation in the Albayzin TTS evaluation campaign. Two synthetic voices have been built this year. One the one hand, an HTS-based TTS with a vocoder based on a parametric representation extracted from an HNM analysis. On the other hand, a Hybrid system that tries to combine the strong points of statistical and unit selection synthesis (i.e. robustness and segmental naturalness respectively). During the voice building process we introduced a feature we had never used so far: IB. We believe that its inclusion had a positive effect in the performance of both TTSs.

The Hybrid system got the best results (alone or together with other systems) in all the three sections of the evaluation, for experts and non-expert listeners. Those were very promising results, being this our first attempt to build a hybrid TTS. The HTS-based system scored pretty well too, above average in all sections but the first one.

A considerable gap between natural and synthetic voices still exists, but hybrid approaches seem to be an appropriate way to try to make the margin smaller.

6. Acknowledgements

The authors would like to thank the organizers of Albayzin TTS 2010 and the developers of all the tools employed during the voice building process.

This work has been partially supported by UPV/EHU (Ayuda de Especialización de Doctores), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Berbatek, IE09-262).

7. References

[1] I. Hernáez, E. Navas, J. Murugarren, and B. Etxebarria, "Description of the AhoTTS System for the Basque Language," *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, ISCA, 2001.

[2] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," *Fifth ISCA Workshop on Speech Synthesis*, r, 2004, pp. 179-184.

[3] S. Rouibia, O. Rosec, and T. Moudenc, "Unit Selection for Speech Synthesis Based on Acoustic Criteria," *Text, Speech and Dialogue*, 2005, pp. 281-287.

[4] T. Hirai and S. Tenpaku, "Using 5 ms segments in concatenative speech synthesis," *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 37-42.

[5] [Online], "HMM-based Speech Synthesis System (HTS)", http://hts.sp.nitech.ac.jp/

[6] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, I. Hernáez, "HMM-based Speech Synthesis in Basque Language using HTS," *Proceedings of Fala2010*, Vigo: 2010.

[7] D. Erro, I. Sainz, I. Saratxaga, E. Navas, I. Hernáez "MFCC+F0 Extraction and Waveform Reconstruction using HNM: Preliminary Results in an HMM-based Synthesizer," *Proc. De las VI Jornadas en Tecnología del Habla*, Vigo: 2010.

[8] F. Campillo, J. van Santen, and E. Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information," *Speech Communication*, vol. 51, 2009, pp. 452-465.

[9] A. Raux and A. Black, "A unit selection approach to f0 modelling and its application to emphasis," *ASRU, St Thomas, US Virgin Islands*, 2003.

[10] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *ICASSP '96: Proceedings of the Acoustics*, vol. 1, 1996, pp. 373-376.

[11] L. Breiman, "Random forests," *Machine learning*, vol. 25, 2001, pp. 5-32.

[12] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, 1978, pp. 43-49.

[13] E.R. Banga, C. G. Mateo, "Documentation of the UVIGO_ESDA Spanish database", *Grupo de Tecnoloxías Multimedia*, Universidade de Vigo, Vigo, España, 2010.

[14] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland:"The HTK Book (for HTK version 3.2)", Cambridge University Engineering Department, 2002.

[15] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, and I. Sainz, "Evaluation of Pitch Detection Algorithms Under Real Conditions," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, IEEE, 2007, pp. IV-1057-IV-1060.

[16] F. Méndez, L. Docío-Fernández, M. Arza, F. Campillo, "The Albayzín 2010 Text-to Speech Evaluation", *Proceedings of Fala2010*, 2010.

The GTM-UVigo Systems for Albayzín 2010 Text–to–Speech Evaluation

Francisco Campillo Díaz[1], *Francisco Méndez Pazó*[1],
Montserrat Arza Rodríguez[2], *Eduardo Rodríguez Banga*[1]

[1]Group on Multimedia Technologies, Universidade de Vigo, Spain
[2]Centro Ramón Piñeiro para a Investigación en Humanidades. Xunta de Galicia.
{campillo, fmendez, marza, erbanga}@gts.uvigo.es

Abstract

This paper describes the synthesizers Cotovía and Cotovía-hts developed by the Group on Multimedia Technologies of the University of Vigo, in cooperation with “Centro Ramón Piñeiro para a Investigación en Humanidades”. Cotovía is a state–of–the–art concatenative system based on a combined selection of acoustic and intonation units, while Cotovía–hts is a preliminary version of a synthesiser based on HMM technology.

Index Terms: speech synthesis, unit selection, statistical parametric, HMMs

1. Introduction

This paper describes the current state of the synthesizers Cotovía and Cotovía–hts at the moment of the Albayzín 2010 Text–to–Speech Evaluation. Cotovía is a state–of–the–art corpus based text to speech synthesizer [1]. In this kind of systems, synthetic speech is generated by concatenation of natural segments selected from a large database recorded from the same speaker. The underlying assumption is that synthetic speech will be indistinguishable from natural if segments are used in similar contexts to those from which they were originally extracted. In practice, this technology alternates chunks of completely close to natural speech, with sporadic and hard to predict concatenation artifacts that degrade severely the quality of the synthetic speech.

Regarding Cotovía–hts, is the first immersion of GTM on HMM based speech synthesis [2], and the result of three months of work, so it can only be considered as a preliminary version. Synthetic speech generated by this kind of systems is smooth and very stable, without the frequent artifacts of unit selection, but with a general artificial impression.

The article is outlined as follows: Section 2 describes the process of building the voice for the two systems, both regarding the information that is used and the system requirements; Section 3 summarises the main characteristics of the synthesizers: system structure, features that are taken into account and prosodic modelling; Section 4 shows the results obtained in the evaluation, and, finally, Section 5 is dedicated to the overall conclusions and future lines of research.

2. Building the voice

The Spanish corpus uvigo_esda was released by the FALA2010 organizers for comparison of the different systems. It consists of around 2 hours of speech (mono, 16 kHz sampling frequency and 16 bits/sample) of isolated sentences read in a neutral style by an amateur speaker. The corpus is phonetically balanced according to the frequency of appearance of phones in the Spanish language, and contains sentences of different lengths and belonging to four broad types: declarative, interrogative, exclamatory and suspended.

The organizers provided the wave files, the text files including information of intonation boundaries, the phonetic segmentation files (not manually revised), and the pitchmark files for voiced segments, as obtained directly from Praat [3].

In the process of building the voice for this evaluation, phonetic segmentation files were not manually corrected, besides some gross errors that were detected while testing the systems, specially regarding major phrase boundaries. Pitchmarks files were postprocessed with several tools developed by the research team, both to fill up unvoiced segments with equally spaced pitch marks, and to ensure that pitchmarks were always positioned at the same point of the local period, in order to avoid concatenation artifacts related to phase mismatch. Also, pitch–synchronous MFCC vectors (12 coefficients) were computed with Festival [4] to model spectral envelope continuity in joints, for the unit selection system. As for the intonation model, stylized intonation contours were extracted from the wave files using Praat, and postprocessed later to smooth out wrong values mainly related to unvoiced segments.

The automated process of building the voice for the unit selection synthesizer took around one hour in an Intel© Xeon™ server, 2.50 GHz processor and 8 GB RAM.

With regards to Cotovía–hts, we used Cotovía, Straight [5] and HTS [6] for the training process of the voice. This process took around 48 hours in an Intel© Xeon™ server, 2.40G GHz processor and 18 GB RAM. Please refer to Section 3.3 for a detailed description of the system.

3. Systems description

Cotovía and Cotovía–hts share a common linguistic module that extracts the information needed for the next stages, so this Section begins with the description of the main features of this module. After that, the distinctive characteristics of both synthesizers are presented.

3.1. Linguistic module

The linguistic module comprises several stages through which the input sentence is translated into a sequence of acoustic target units characterized by a set of features that are used both for prosodic modeling and waveform generation. In this sense, features are extracted related to phone identity, phonetic context, phone boundaries, accentuation, syllabic structure, type of sentence, position in the phonic group, part–of–speech (POS) labels and syntactic information.

With respect to POS, a hybrid analysis is performed [7]. First, a reduced set of highly reliable linguistic rules is used to eliminate from each word those categories that are not possible according to the context. Second, a statistical tagger makes up the final decision combining a contextual 5–gram model of sequences of categories, and a lexical model that considers the probability of each word having a certain category. Figure 1 is an example of the use of the 5–gram window to consider the ambiguous context around each word.

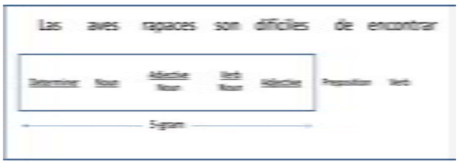


Figure 1: *Example of POS disambiguation*

3.2. Cotovía

3.2.1. Prosody estimation

Cotovía includes different modules to estimate duration, energy, intonation and phrasing.

Intonation Similarly to acoustic unit selection, Cotovía integrates a corpus based intonation module [8], with the accent group (defined as a sequence of unaccented words finishing in an accented one) as the basic unit for concatenation. This model is characterized by:

- Accent groups are described by their position within the phonic group and the sentence, the types of boundary surrounding them, the number of syllables, the position of the accent, the type of sentence, the POS of the accented word and the syntagma following it.
- The target cost penalizes the differences from the aforementioned features to the estimated ones. Perhaps the most interesting detail is that syntactic and morphosyntactic information is used to decide both the strength of the accent and the insertion of minor phrase boundaries [9].
- The concatenation cost only takes into account f0 continuity and boundary continuity, since it was found that joining two accent groups with different boundaries degrades severely the quality of the synthetic contour.

Phrasing Although a combined approach would probably yield better results, in Cotovía major and minor phrasing algorithms are implemented as different stages. First, major phrasing is accomplished by means of a decision tree, with factors such as the distance in syllables from the last pause and the distance in syllables to the next pause, and a POS window of three places to the left and right of the current word. And second, minor phrasing is integrated into the intonation module [9], taking major phrasing as an input. This way, minor phrasing is modeled as another subcost in the intonation target cost function, considering the POS and syntactic information as input. For every target accent group, candidate groups that can be followed or not by a minor phrase boundary are considered. Therefore, the best sequence of candidate accent groups resulting from the Viterbi search includes the best prosodic structure for the input sentence. Figure 2, where shaded and unshaded circles represent candidate accent groups with different boundaries, depicts this situation.

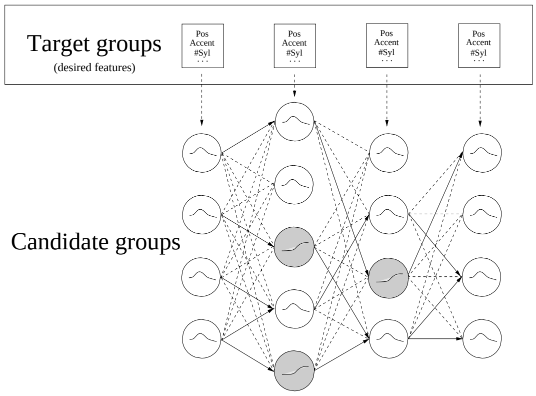


Figure 2: *Combined selection of minor phrasing and intonation contours*

Duration Phones are clustered into ten classes (open vowels, mid vowels, close vowels, voiced plosives, voiceless plosives, fricatives, laterals, nasals, vibrants and silence), and multivariant linear regression models are trained for all of them. The identity of the phone, the phones surrounding it in a window of size five, the position within the word and the phonic group, the type of sentence and the lexical accent are the features used in each model.

Energy Similar to duration, phones are divided into eleven classes (silence, open vowel, mid vowel, close vowel, voiced plosive, voiceless plosive, aproximant, fricative, nasal, vibrant and lateral), and multilayer perceptrons are trained for each of them, with features such as the identity of the phone, the energy of the previous phone, the lexical accent, the position within the sentence and the type of sentence.

3.2.2. Acoustic unit selection

As mentioned before, Cotovía is a corpus based synthesizer [1], with the demiphone as the basic unit for concatenation. Perhaps the most interesting difference lies in considering more than a single candidate intonation contour, as most of the other synthesizers do. In natural speech, a sentence can be realized in many different ways just by changing prosody, without affecting the meaning of the message that is conveyed. This way, in Cotovía several coherent candidate intonation contours are extracted from the intonation module, giving another degree of freedom to the acoustic unit search, and improving the quality of synthetic speech [8]. For example, Figure 3 shows two candidate intonation contours for the same sentence.

Summing up, the main characteristics of the acoustic unit selection stage are the following:

- Demiphones are parameterized according to the identity of the surrounding phonemes (in a window of size five), the lexical accent, the position within the phonic group, the type of sentence and the types of boundary. Regarding prosody, duration, energy and fundamental frequency at the beginning and end of the demiphones are considered.
- Target cost comprises two parts. First, similarity of phonemic context is computed using only symbolic features (unlike previous versions of the synthesizer, where MFCC were used with the same goal [10]). And second,

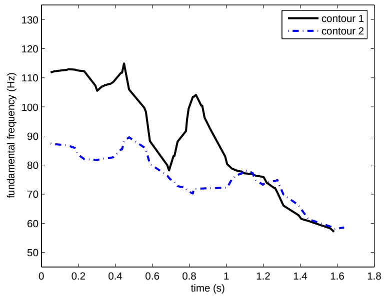


Figure 3: *Two intonation contours with different prosodic structure. Notice the minor phrase break in the solid line contour, around 0.9 seconds*

differences to the target prosody, as given by the modules mentioned before, are included.

- Concatenation cost: continuity of fundamental frequency, energy and spectral envelope are considered.

3.2.3. Waveform generation

Synthetic speech is generated by concatenation of the waveforms of the sequence of candidate acoustic units resulting from the Viterbi search. Demiphones close enough to the target duration and fundamental frequency values (40 ms and 5 Hz, respectively) are not prosodically modified, in order to preserve the micropsody and quality of the original recording.

3.3. Cotovía–HTS

In this section we describe Cotovía–HTS, our first statistical parametric speech synthesis system based on hidden Markov models (HMMs). This kind of systems [2] are now very popular, largely due to the release of the HMM–based Speech Synthesis System (HTS) [6][11]. Moreover, the results obtained by such systems in the Blizzard Challenge and also in the last edition of the Albayzín TTS Evaluation show that this is a very interesting and promising field in the speech synthesis research.

In brief, HTS works in two different phases: training and synthesis. In the training part, both spectrum and excitation (and its dynamic features) parameters are extracted from the speech database. This features are modeled by context–dependent HMMs (CD–HMMs), taking account of phonetic, linguistic and prosodic contexts extracted from the labelled speech database.

In the synthesis stage, the input text has to be processed to obtain a context–dependent label sequence, which HTS uses to obtain a sequence of CD–HMMs. Next, excitation and spectral parameters are obtained using the speech parameter algorithm [12], which in turn are used to generate the synthetic speech output.

Figure 4 shows an overview of our system. We use the text processing module from our unit–selection TTS synthesizer, Cotovía, to provide the context–dependent labels used for the CD–HMMs training. The linguistic features used are:

- Phoneme level:
  - Current phoneme identity.
  - Start and end time instants.
  - 2 previous and 2 next phonemes identity.

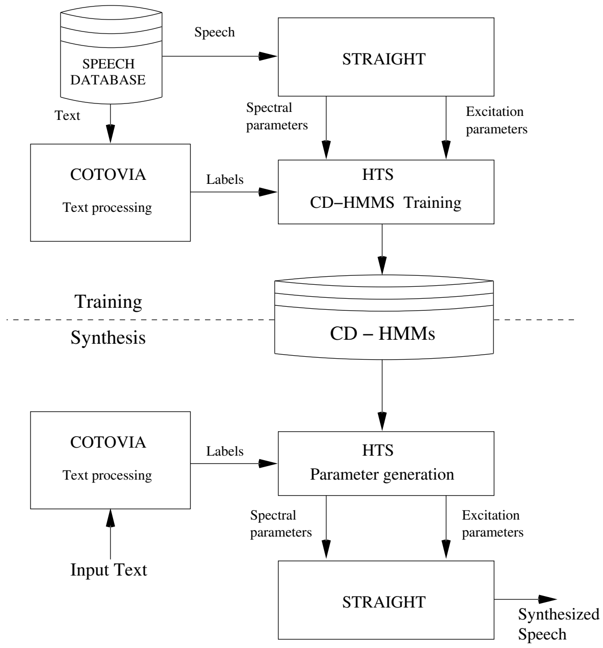


Figure 4: *Overview of Cotovía–HTS*

  - Position of the current phoneme in the current syllable (backward and forward).
- Syllable level:
  - Stress and number of phonemes in the previous, current and next syllable.
  - Position of the current syllable in the current word and phrase (backward and forward).
  - Number of stressed syllables after and before the current syllable in the current phrase.
  - Number of syllables, counting from the previous stressed syllable to the current syllable in this utterance.
  - Number of syllables, counting from the current syllable to the next stressed syllable in this utterance.
  - Vowel in the current syllable.
- Word level:
  - Part of speech (content or function) of the previous, current and next words.
  - Number of syllables of the previous, current and next words.
  - Position of the current word in the current phrase (backward and forward).
  - Number of content words after and before the current word in the current phrase.
  - Number of words, counting from the previous content word, to the current word in this utterance.
  - Number of words, counting from the current word to the next content word in this utterance.
- Phrase level:

- Number of syllables and words in the previous, current and next phrases.
- Position of the current phrase in the utterance (backward and forward).
- Type (declarative, interrogative, exclamatory and suspensive)

- Utterance level:
  - Number of syllables, words and phrases in the current utterance.

Straight [5] was used to obtain the spectral and excitation parameters. In our case 39th order Mel–cepstrum, logf0 and 5 band–aperiodicity coefficients together with their dynamic features (first and second derivatives) were extracted from the speech database. These parameters were used for training, amounting in total 88878 multi–stream —5 stream and 7 states— context–dependent HMMs.

For the synthesis part, again we used Cotovía to extract the context–dependent labels from the test sentences. Then, HTS converts this label sequence into a sequence of CD–HMMs, and the speech parameter generation algorithm provides the spectral and excitation parameters. The final speech waveform is synthesized from these parameters using Straight.

4. Results

The results of the Albayzín 2010 Text–to–Speech evaluation [13] can be considered very positive for both systems (Cotovía was system "I" on the evaluation, while Cotovía–hts was system "E").

Since Cotovía participated too on the Albayzín 2008 evaluation [14] (system "B"), we can compare the results in both cases to have an idea of the improvement. Table 1 shows the results of the evaluations regarding MOS (Mean Opinion Score) and similarity to the original voice, on a scale of 1 to 5, with 1 being the worst and 5 being the best. The results of Cotovía–hts and the best system in the 2010 evaluation are also included. Comparing the output of a system with two different voices might lead to wrong conclusions, but the relative performance of Cotovía regarding the other participants in both evaluations shows an improvement as well: in 2008 Cotovía was third in both MOS and similarity to the original voice, while in 2010 it was second in MOS and similarity, with no statistical differences to the best system in this last test.

With regards to WER (Word Error Rate), there was a surprising decrease in performance, from 4.95% in 2008 to 24% in 2010. The authors consider this to be a result of a more difficult task in the intelligibility test of 2010: while in 2008 the WER ranged between 3.49% and 8.19%, in 2010 it ranged between 15% and 36%.

	MOS		Similarity	
	Mean	Median	Mean	Median
Cotovía–2008	2.91	3	3.36	3
Cotovía–2010	3.50	4	4.02	4
Cotovía–hts (2010)	3.10	3	3.15	3
Best–2010	3.78	4	4.07	4

Table 1: Cotovía: comparison between 2008 and 2010

Cotovía-hts was first in intelligibility (15%), which is clearly remarkable on being compared with other synthesizers with many years of development.

5. Conclusions

This paper describes the current state of the synthesizers Cotovía and Cotovía–hts, as were presented at the Albayzín 2010 TTS evaluation, including both the steps to build a new voice and the process followed to generate synthetic speech. The results of the evaluation were very positive, confirming also the general trend of unit selection systems being more natural and similar to the original voice, and HTS being more intelligible. Cotovía showed a clear improvement when compared with the results of the Albayzín 2008 evaluation (from 2.91 to 3.50 in MOS, and from 3.36 to 4.02 in similarity to the original voice), while Cotovía–hts was first in the intelligibility test and also outperformed in MOS and similarity other synthesizers with many years of development.

6. Acknowledgements

The work reported here was partially supported by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04 "Búsqueda de Información en Contenidos Audiovisuales plurilingües", Xunta de Galicia "Isidro Parga Pondal" research programme and *Centro Ramón Piñeiro para a Investigación en Humanidades*.

7. References

[1] A. Hunt, A. Black, "Unit selection in a concatenative speech synthesis using a large speech database", Proceedings of ICASSP, Volume 1, pp. 373-376, Atlanta, 1996

[2] Heiga Zen, Keiichi Tokuda, Alan W. Black, "Statistical parametric speech synthesis", Speech Communication, Volume 51, Issue 11, November 2009, pp. 1039-1064

[3] http://www.praat.com

[4] http://www.cstr.ed.ac.uk/projects/festival

[5] Hideki Kawahara, Ikuyo Masuda-Kasuse, Alain de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, 27, pp.187-207, 1999

[6] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, "The HMM-based speech synthesis system version 2.0" , Proceedings of ISCA SSW6, Aug. 2007

[7] F. Méndez, F. Campillo, E.R. Banga, "Análisis morfológico estadístico en lengua gallega", Procesamiento del lenguaje natural, Volume 31, pp 159-166, 2003

[8] F. Campillo, E.R. Banga, "A method for combining intonation modelling and speech-unit selection in corpus-based speech synthesis systems", Speech Communication, Volume 48, Issue 8, pp. 941-956, 2006

[9] F. Campillo, Jan Van Santen, E.R. Banga, "Integrating phrasing and intonation modelling using syntactic and morphosyntactic information", Speech Communication, Volume 51, Issue 5, pp. 452-465, 2009

[10] F. Campillo, E.R. Banga, "On the design of the cost functions for a unit selection speech synthesis", Proceedings of Eurospeech, Volume 1, pp. 289-292, Geneva, 2003

[11] Online: http://hts.sp.nitech.ac.jp/

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM–based speech synthesis", Proceedings of ICASSP, pp.1315-1318, June 2000.

[13] Francisco Méndez Pazó, Laura Docío-Fernández, Montserrat Arza Rodríguez, Francisco Campillo Díaz, "The Albayzín 2010 Text–to–Speech Evaluation", Proceedings of Fala2010, 2010.

[14] Iñaki Sainz, Eva Navas, Inma Hernáez, Antonio Bonafonte, Francisco Campillo, "TTS evaluation campaign with a common Spanish database", Proceedings of LREC, 2010.

The GTH-CSTR Entries for the Speech Synthesis Albayzin 2010 Evaluation: HMM-based Speech Synthesis Systems considering morphosyntactic features and Speaker Adaptation Techniques

R. Barra-Chicote[1], *J. Yamagishi*[2], *J. M. Montero*[1], *O. Watts*[2], *S. King*[2], *J. Macias-Guarasa*[3]

[1]Speech Technology Group, Universidad Politecnica de Madrid
[2]Center for Speech Technology Research, University of Edinburgh
[3]Geintra Group, University of Alcala

barra@die.upm.es, jyamagis@inf.ed.ac.uk, macias@depeca.uah.es

Abstract

This paper describes the GTH-CSTR systems developed for the *Albayzin 2010 Speech Synthesis Evaluation*. We have developed three different HMM-based systems to build synthetic voices in Spanish, using two hours of speech of a male speaker. We have improved our baseline system (GTHCSTR-2008) by using morphosyntactic features, iterative segmentation algorithms, enhanced feature analysis and speaker adaptation techniques.

Index Terms: text to speech synthesis, statistical parametric speech synthesis, morphosyntactic features, speaker adaptation, speech synthesis evaluation

1. Introduction

The quality of HMM-based speech synthesisers has been improving in the recent years, also showing good intelligibility rates. However, the over-smoothing tendency, typical of these synthesisers, causes that most of the sentences are spoken in a very closely form. We have incorporated morphosyntactic features to the system, looking to improve the prosody generation of our text-to-speech system (TTS) and to enrich the way it reads complex sentences.

One of the features of HMM-based synthesis is their flexibility as compared to unit selection synthesis. Since we have an explicit speech model, its parameters can be modify more easily modified to obtain new voices. The application of model adaptation techniques to an average voice [1], allows the possibility of building a target speaker voice using only a few minutes of speech. We have incorporated those techniques to our baseline system and present an additional entry to the evaluation.

2. Albayzin 2010 Speech Synthesis Evaluation

The *Albayzin 2010 Speech Synthesis Evaluation* is an event, similar to the *Blizzard Challenge*, promoted in order to compare different techniques for building corpus-based speech synthesisers applied to Spanish. The challenge consists of building a voice from a released data set and synthesising a predefined set of test sentences, which are perceptually evaluated through listening tests by volunteers and speech experts.

Each voice is evaluated in terms of:

- Similarity with the target speaker
- Naturalness
- Intelligibility

3. Corpora

The organisation has released the UVIGO_ESDA corpus as the target speaker synthetic voice for this challenge. In addition we have also used our Spanish Expressive Voices (SEV) corpus as part of the training data in one of our three submitted systems.

3.1. UVIGO_ESDA Corpus

The UVIGO_ESDA Database contains speech recordings from an amateur male speaker that read prompted texts in “neutral” style. The database approximately contains two hours of speech and 1217 phonetically balanced sentences, automatically extracted from journalistic texts by means of a greedy algorithm. Data collection was performed at a recording studio. Audio files with the original sampling frequency (44100 KHz) were also provided for training tasks (test audio files should be in 16 KHz).

3.2. SEV Corpus

The *Spanish Expressive Voices* (SEV) corpus [2] comprises speech and video recordings of an actor and an actress speaking in a neutral style and simulating six basic emotions: *happiness*, *sadness*, *anger*, *surprise*, *fear* and *disgust*.

The SEV corpus covers speech data in several genres such as isolated word pronunciations, short and long sentences selected from the SES corpus [3], narrative texts chosen from a novel “Don Quijote de la Mancha”, a political speech, short and long interviews, question answering situations and short dialogues. The texts of all utterances are emotionally neutral.

More than 100 minutes of speech duration per emotion have been recorded, allowing for comprehensive studies in emotional speech synthesis, prosodic modelling and speech conversion. The amount of data per emotion and speaker is close to one hour of speech.

For this challenge, we have used the neutral voice of the SEV male speaker as the *Base Voice* in the training process of the GTHCSTR-2010 Adaptation-based system described in the following section. However, for further research we are also interested in using positive emotional voices as the *Base Voice*.

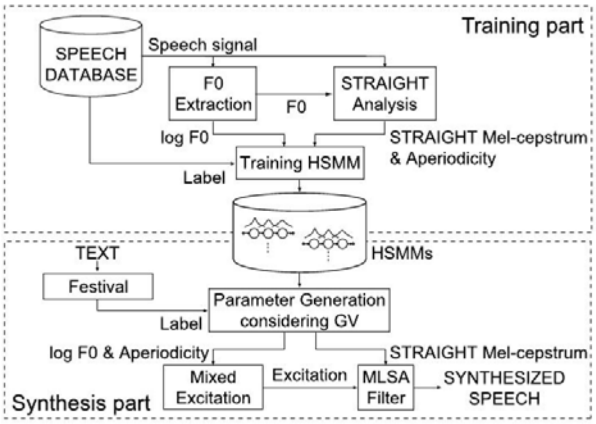


Figure 1: *Blocks diagram GTHCSTR-2008 system.*

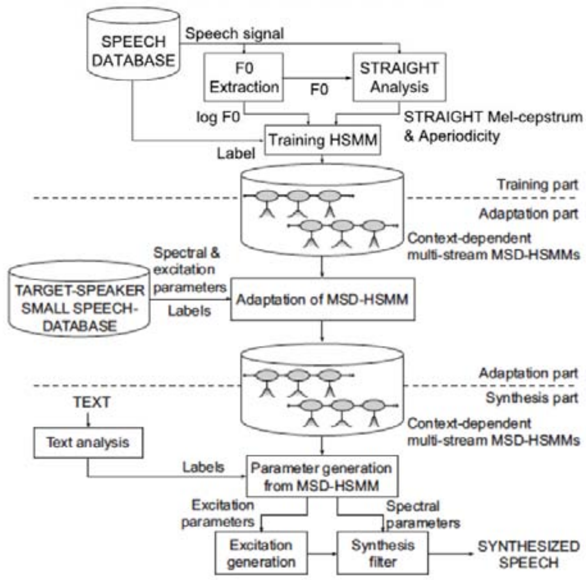


Figure 2: *Block diagram of GTHCSTR-SA-2010 adaptation-based system.*

4. Systems Description

4.1. GTHCSTR-2008: Baseline System

Our HMM-based voices have been built using a method similar to the Nitech-HTS 2005 system [4] which is publicly available from the HTS toolkit website [5].

The HMM-based speech synthesis system comprises three components: speech analysis, HMM training, and speech generation. In the speech analysis part, three kinds of parameters for the STRAIGHT [6] mel-cepstral vocoder with mixed excitation (the mel-cepstrum, $\log F0$ and a set of aperiodicity measures) are extracted as feature vectors for modelling by the HMMs. These are as described in [4], except that the F0 values we used were more robustly estimated using a vote amongst several F0 extraction algorithms [7]. In the HMM training part, context-dependent multi-stream left-to-right MSD-HSMMs [8] are trained using the maximum likelihood criterion. In the speech generation part, acoustic feature parameters are generated from the MSD-HSMMs using the GV parameter generation algorithm [9]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) [10]. This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter corresponding to the STRAIGHT mel-cepstral coefficients, generating the speech waveform. Figure 1 plots the blocks diagram of the GTHCSTR-2008 system.

The GTHCSTR-2008 system exhibited very good performance in the previous *Albayzin 2008 Speech Synthesis Evaluation* [11]. Emotional synthetic voices have been recently developed with this system [12].

4.2. GTHCSTR-SA-2010: Adaptation-based System

The GTHCSTR-2008 system has been improved with the inclusion of adaptation techniques. We have incorporated CSMAPLR an MAP adaptation algorithms in the voice training processes, fully described in [13, 14].

An average voice [1] is usually used in speaker independent TTS systems [14] as the *Base Voice* that is adapted to the target speaker. However, for the challenge we used the neutral male voice (built with GTHCSTR-2008 system) from the SEV corpus as the *Base Voice*. This voice was adapted to the target speaker of UVIGO_ESDA corpus only using the first 50 sentences (5% of training data) from the training set (approximately 5 minutes of speech). Figure 2 plots the block diagram of the GTHCSTR-SA-2010 system.

The objective of this GTHCSTR-2008 system is comparing the results between the speaker dependent voices of the target speaker (GTHCSTR-SD-2010 system described below) and an adapted voice trained using only a very small amount of training data.

4.3. GTHCSTR-SD-2010: Speaker Dependent System

4.3.1. Acoustic processing improvements

We have modified our speaker dependent system by using some acoustic improvements, as compared to the GTHCSTR-2008 implementation:

- A bigger spectral bandwidth using the original sampling frequency (44100 KHz) in the training process. In the synthesis stage, the speech signals were down-sampled to 16000 KHz.
- A higher number coefficients in the analysis of the spectral component.
- An iterative segmentation process based on building partial voices used for relabelling.

4.3.2. Morphosyntatic Features

In order to improve the basic HMM-based system, we have also included new features coming from a morphosyntactic analysis of the input sentences. As the natural language processing (NLP) of the speech synthesis sentences should be very robust (in order to deal with whatever grammatical structures the author of the target texts could use), shallow techniques seem to be a good choice. The first module in our NLP chain is a Spanish Part-Of-Speech tagger (Montero 2003), based on ESPRIT-860's EAGLES-like 10-byte labels (more than 250 possible tags), using a set of dictionaries such as RAE's 159898-word dictionary, richly-tagged ESPRIT-860's 9883-word dictionary, Onomastica's 58129-proper-noun dictionary, GTH's 1960-multiword

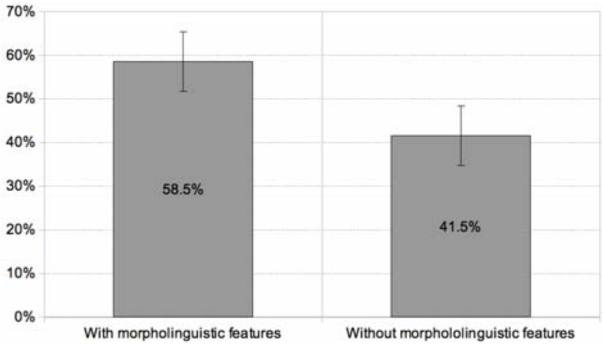


Figure 3: *Comparison between Speech Quality (SQ) scores obtained by GTHCSTR-SD-2010 speaker dependent system considering morphosyntatic features or not.*

expression dictionary, and GTH's verb conjugation analyser (including 102 irregular paradigms and 4325 infinitives).

After assigning all possible tags to the words, several sets of hand-written rules are used for cascade-filtering impossible tag sequences: GTH's 77 high-recall rules, CRATER's 148 rules and GTH's 162 low-recall high-precision rules. On the 38172-word test-set of the ESPRIT-860 Spanish corpus, the recall is as high as 0.9987 when averaging 1.6984 tags per word.

Finally, the TnT stochastic tagger (Brants) is used for disambiguation. This tagger uses an interpolated language model based on trigrams, bigrams and unigrams, resulting in a 98.99% accuracy for a 1-tag-per-word basis, or 99.45% if 1.0238 tags are assigned per word on average.

After tagging the sentence, 2 features are available to be used in the speech synthesis training and testing process:

- A gross 10-category feature (based on 860 tag).
- A 3-byte set of tags in 860 coding scheme, (including a first byte for the 10 main tags and 2 additional bytes for a more detailful sub-tagging).

The final NLP processing module is a shallow parser based on a CYK botton-up algorithm and a set of 2179 hand-written general-purpose CYK parsing rules. As these rules are very ambiguous, many possible parser trees are assigned to each sentence. In order to control the exponential growth of this analysis, a small set of splitting rules were developed (trying to reduce the length of the text to be analysed) and a final filtering process was used, selecting only one tree using a Minimum Description Length approach. In a subset of the test set, for a total 5703 shallow syntactic phrases, there were 0.35% cutting errors, 0.55% tagging-recall errors, 1.10% tagging-precision errors and 1.49% syntactic-analysis errors. These shallow syntactic phrases are the third feature to be used in the synthesis process.

reader to understand the figure Figure 3 shows the results of an internal perceptual test to validate the improvements when adding the morphosyntatic features. Based on these results (a 41% statistically significant relative improvement in Speech Quality) our final system considered the morphosyntatic features described.

5. Results and Discussion

Figure 4 shows the similarity scores for all systems submitted to the evaluation (for all listeners). GTHCSTR-2008 is

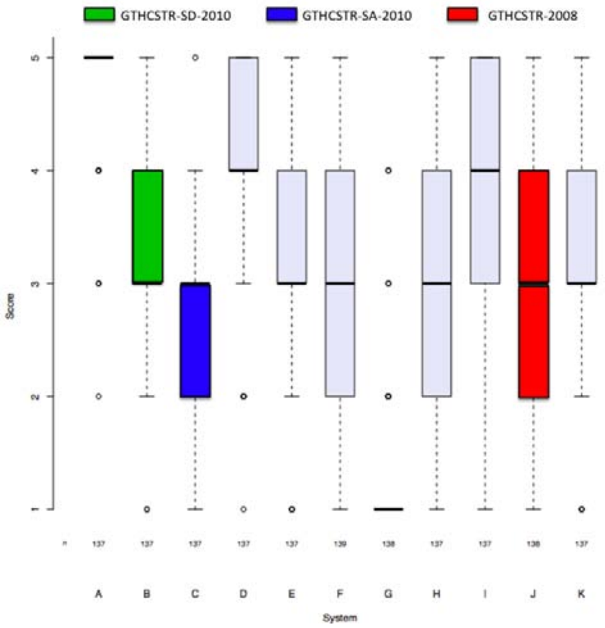


Figure 4: *Similarity scores for voice UVIGO-ESDA (All listeners).*

plotted using a red box, GTHCSTR-SD-2010 is plotted using a green box, and GTHCSTR-SA-2010 is plotted using a blue box. GTHCSTR-SD-2010 has good similarity scores and significantly improves our baseline GTHCSTR-2008 system.

Our speaker adapted system, GTHCSTR-SA-2010, obtains excellent similarity results (median value of 3, equivalent to the speaker dependent systems) considering that we have only used 50 sentences (5% of the whole training data) of the target speaker to build its voice. This result strongly supports the goodness and the high potential of the speaker adaptation algorithm.

Figure 5 shows the MOS scores (considering all listeners). Again, GTHCSTR-SD-2010 significantly improves GTHCSTR-2008. In this case, as expected, the MOS scores obtained by GTHCSTR-SA-2010 are lower than the speaker dependent system. These results are reasonable since the *Base Voice* was built only with 50 minutes of speech (in comparison with the 2 hours of UVIGO_ESDA) and we only used 5 minutes of adaptation data.

Figure 6 shows the MOS scores only considering the listeners that did not used headphones. In this case, our GTHCSTR-SD-2010 system did not show significant diferences with the two systems which obtained better MOS scores. We presume that the over-smoothing introduced by our synthesis technique is filtered by the channel when synthetic speech is heard using speakers instead of headphones.

Figure 7 shows the Word Error Rate in the intelligibility tests (WER considering all listeners). GTHCSTR-SD-2010 has lower WER than GTHCSTR-2008. Also, there are no significant differences between the best system (system E) and GTHCSTR-SD-2010 and GTHCSTR-2008.

6. Conclusions

This paper described the GTH-CSTR systems submitted to the *Albayzin 2010 Speech Synthesis Evaluation*. All of them are

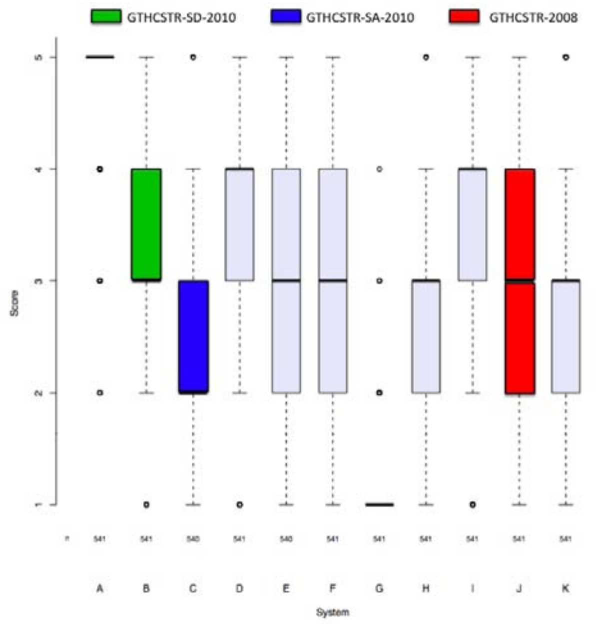


Figure 5: *Mean opinion scores for voice UVIGO-ESDA (All listeners).*

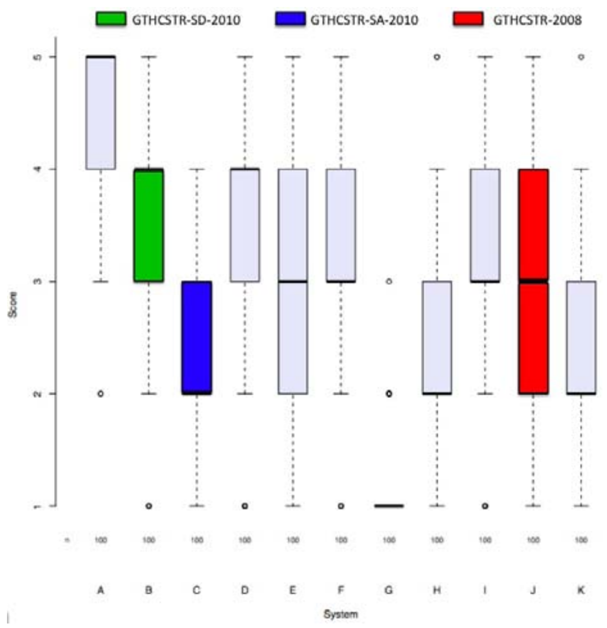


Figure 6: *Mean opinion scores for voice UVIGO-ESDA (listeners headphones=NO).*

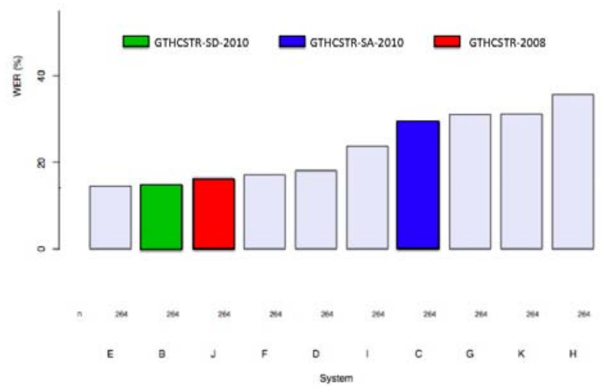


Figure 7: *Word error rate for voice UVIGO-ESDA (All listeners).*

based on HMM-based synthesis to build synthetic voices in Spanish.

Three systems have been presented:

- A baseline system submitted to the 2008 Albayzin Evaluation (GTHCSTR-2008)
- A system based on speaker adaptation algorithms (GTHCSTR-SA-2010)
- An improved speaker dependent system (GTHCSTR-SD-2010).

The synthetic voice built with GTHCSTR-SA-2010 system is reasonably perceived as the target speaker, in spite of having used only the 5% of the training data.

An internal evaluation validated the goodness of adding morphosyntatic features in order to improve the quality of our GTHCSTR-SD-2010 synthetic voices.

7. Acknowledgements

The authors which to express their gratitude to the members of GTH and CSTR for their collaboration in the preparation of this work.

8. References

[1] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[2] R. Barra-Chicote, J. Montero, J. Macias-Guarasa, S. Lufti, J. M. Lucas, F. Fernandez, L. D’haro, R. San-Segundo, J. Ferreiros, R. Cordoba, and J. Pardo, “Spanish Expressive Voices: Corpus for emotion research in Spanish,” in *Proceedings of 6th international conference on Language Resources and Evaluation*, 2008.

[3] J. M. Montero, J. M. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, and J. M. Pardo, “Emotional speech synthesis: From speech database to TTS,” in *Proc. ICSLP-98*, Dec. 1998, pp. 923–926.

[4] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[5] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, *The HMM-based speech synthesis system (HTS) Version 2.1*, 2008, http://hts.sp.nitech.ac.jp/.

[6] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, pp. 187–207, 1999.

[7] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, “A robust speaker-adaptive HMM-based text-to-speech synthesis,” *IEEE Trans. Speech, Audio & Language Process.*, 2009, (in press).

[8] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis system,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[9] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[10] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–468, 1990.

[11] R. Barra-Chicote, J. Yamagishi, J. Montero, S. King, S. Lutfi, and J. Macias-Guarasa, "Generacion de una voz sintetica en Castellano basada en HSMM para la Evaluacion Albayzin 2008: conversion texto a voz," in *V Jornadas en Tecnologia del Habla*, Nov. 2008, p. Please add page range.

[12] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, no. 5, pp. 394 – 404, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1C-4XY4GDS-1/2/7e701c2305a5ff0713d2c2e83af6e760

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[14] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007," in *Proceedings BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

Building an HMM-based Spanish TTS system for Albayzin 2010 challenge

Daniela Braga[1,2], *Pedro Silva*[1], *João Freitas*[1], *Difda Monterde*[1], *Miguel Sales Dias*[1,2]

[1] Microsoft Language Development Center, Portugal
[2]ADETTI – ISCTE, IUL, Lisbon, Portugal
{i-dbraga, i-pedros, i-joaof, v-dimont, Miguel.Dias}@microsoft.com

Abstract

In this paper we describe the process of building a new Text-to-Speech (TTS) system in Spanish from the materials provided for the Albayzin 2010 challenge using the deployed Microsoft TTS technology in Spanish. The architecture of the system is described as well as the process of compiling a new voice font and producing a new synthetic voice. The main issues found during this process are presented and discussed.

Index Terms: Text-to-Speech, HTS, Spanish, voice quality, evaluation

1. Introduction

When talking about the Spanish language, one has to keep in mind that it is in the top 5 of the most spoken languages in the world [1]. Spanish is the official language in 21 countries, and it is spoken in nearly 30 countries around the world [2]. This amounts to almost 400 million of Spanish speakers, of which 350 million speak Spanish as the native language [3]. These numbers can be considered a huge motivation to work on Speech Technologies such as Text-to-Speech (TTS) systems.

For European Spanish, we find some university research groups developing text-to-speech systems based on concatenative synthesis, such as *Cotovía*, a TTS for Spanish and Galician from the University of Vigo [4]; the *Ogmios* UPCTTS for Spanish and Catalan from the Talp (Tecnologies i Aplicacions del Llenguatge i la Parla) group of the Politechnical University of Catalunya [5]; and the ViVoLab (Voice input Voice output Lab) from the University of Zaragoza [6]. The Politechnical University of Madrid tested both concatenative-based and HSMM-based synthesis systems for TTS and got better results using the last methodology mentioned [7].

Due to the large size of the Spanish speaking population, some private initiatives were interested in the market of Spanish TTS. Some of them focused on Latin American Spanish, such as *Marta* and *Miguel* from Cepstral [8], *Rosa* and *Alberto* from AT&T Labs and *Violeta* from Neospeech [9]. Some TTS from other companies offer only European Spanish or both varieties, among them we can find *Jorge, Carmen* and *Leonor* from Loquendo [10], *Amaya, Carlos* and *Laura* from Verbio [11], *Isabel, Monica, Diego, Paulina* and *Javier* from Nuance [12], and *Antonio* and *Maria* from Acapela [13]. Microsoft released a Spanish synthetic voice in Exchange 2010 deployment, named Helena, available for download both on server and client side[1], together with other languages for mobile and desktop interfaces.

[1] Microsoft Helena can be downloaded here: http://www.microsoft.com/downloads/en/details.aspx?FamilyID=f704cd64-1dbf-47a7-ba49-27c5843a12d5&displaylang=en

2. System description

The front-end of the system is dictionary-based, being composed by a lexicon with 599520 words, phonetically annotated with phonetic transcriptions, stress marks and syllable boundaries, and with Part-of-Speech (POS) information. The front-end is also composed by the text analysis, which involves the sentence separator and word breaker modules and includes a couple of other files, such as phone set, features and the POS tags set. It also includes a rule-based Text Normalization module and stochastic-based LTS (Letter-to-Sound) converter to predict phonetic transcriptions for out-of-vocabulary words. The prosody model was trained with 2000 utterances prosodically tagged in terms of breaks, boundaries and intensity.

The front-end outputs phonetic transcriptions that are subsequently input of the TTS runtime engine or back-end, which then outputs synthetic voice. Figure 1 illustrates the system workflow.

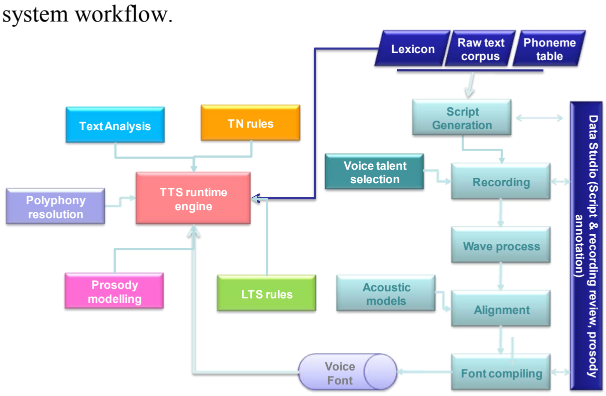


Figure 1: Microsoft TTS system pipeline.

The voice font building is also a very complex and demanding process that requires the following steps: script selection (using different text genders, phonetically balanced, with a broad prosody coverage (in terms of types of sentences – declaratives, interrogatives, exclamatory sentences), in a total of 11 500 prompts and nearly 13 hours of speech), recording process at 44 kHz, 16 bits of sampling rate), edition of the prompts, recording quality control, re-recording and edition of the prompts which failed in the quality control, wave process, automatic alignment and quality validation, font compiling and conversion of the original recorded waves to 8khz, 8 bits sample rate. Figure 2 depicts the voice font building process.

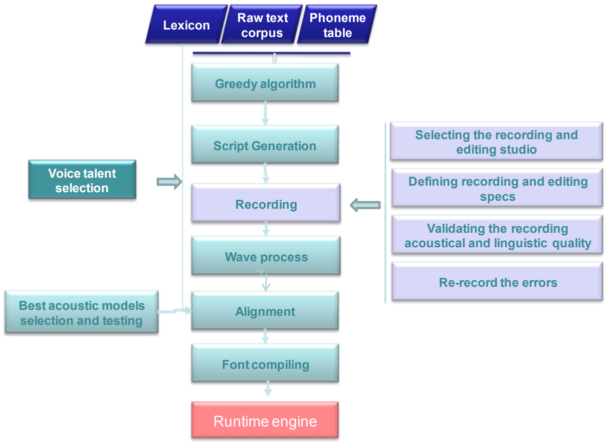


Figure 2: Voice font building process.

3. Building a new Spanish voice for Albayzin 2010

The Spanish Albayzin voice was developed using the same process and tools as the currently shipped Spanish (Spain) voice from Microsoft, which is presently available in several Microsoft commercial products and services. Additionally, the Microsoft Spanish voice is also available through a public Speech Platform SDK [14], which can be used for the development of speech-enabled applications.

As such, the current Microsoft Spanish text analysis front-end module was leveraged for preparing the input data, namely the phone set, lexicon and related components.

The provided Albayzin database consisted of 1217 phonetically balanced sentences (text sentences + wave files) recorded from an amateur male speaker in "neutral" style, comprising ~2 hours of speech.

To build the Albayzin voice the following steps were taken:

1. Prepare input wave files;
2. Prepare input script file;
3. Extract phone segmentation / alignment results;
4. Train HMM-based speech synthesis (HTS) voice font.

For step 1, all the 1217 wave files were first filtered and normalized before the training step.

In step 2, an input script file was prepared with the 1217 text sentences, including phonetic transcription + syllable + stress marks (all obtained from proprietary Spanish lexicon) as well as POS tags, per word. Before doing this, the input script file was first cleaned of some errors that were found originally, such as empty lines and more than one sentence per line.

Step 3 consists of an automatic process for extraction of phone segmentation labels, based on the script file and the wave files.

Finally in step 4, the HTS voice font was trained using the script, wave files and phone alignment results.

Once finished, the public Speech Platform [14] was used to synthesize the output test wave files (the original Spanish voice font was replaced with Albayzin trained voice font, keeping the runtime TTS engine and the frontend text analysis module).

4. System Evaluation

After training our system with the Albayzin database (1217 prompts and its correspondent wave files), a set of 430 sentences were provided in order to be synthesized and uploaded in an online platform designed for this purpose. Then Albayzin organization created an online evaluation tool where all synthesized utterances from the 10 systems participating in the challenge were randomly displayed, always having the original voice as baseline. Three tests were presented, each one testing naturalness (proximity to the original and MOS) and intelligibility (through a SUS test). Microsoft contributed to this evaluation with 13 participants, 11 of each from outside Microsoft. Participants' requirements to perform this evaluation were: being native or nearly native speaker and expert in language technology. The results of this evaluation are described in [17].

Microsoft's system (system G) results are depicted in Figures 3-5:

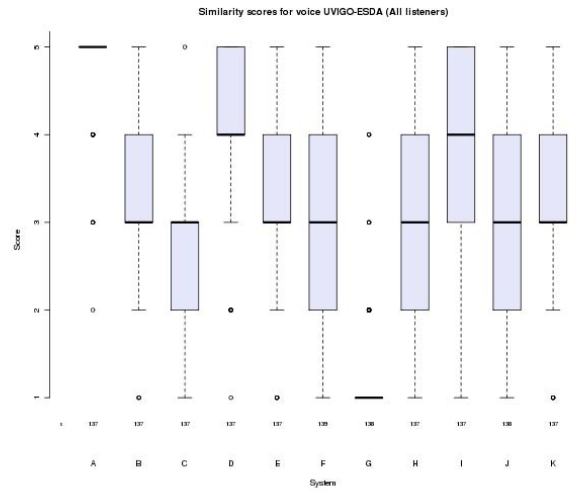


Figure 3: Test 1 results: similarity to the original voice overall scores.

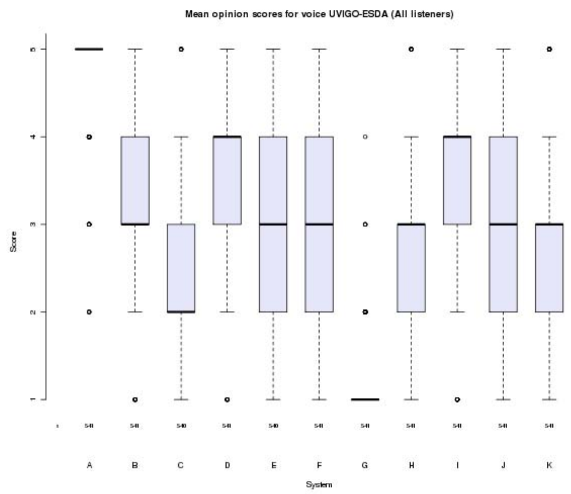


Figure 4: Test 2 results: overall Mean Opinion Scores.

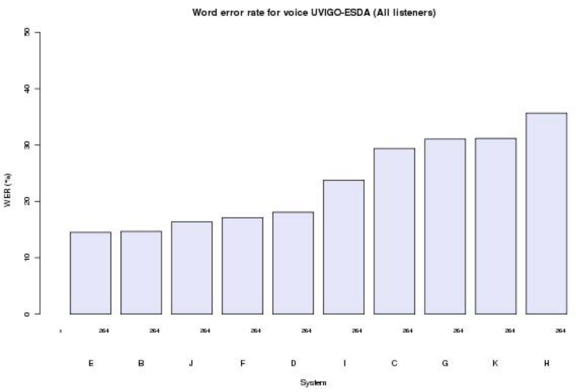


Figure 5: Test 3 results (SUS): overall WER.

The overall results show that our system was rated with 1 in a 5 points scale in Test 1 and Test 2. Regarding the SUS test, Microsoft's system had 30% of WER, which ranked it in the 8th position out of 10 systems presented in the challenge. These results were very surprising to us especially because they don't reflect our internal assessment of the released technology. For confidentiality reasons, we cannot publish those results against the competitors' systems, but we can state that our system passed Microsoft's quality bar regarding intelligibility and voice quality (MOS) after several test rounds with more than 40 native listeners each. Not to mention other performance and back-end tests, where our technology passed with very good scores. The reasons for these bad results may be the following: 1) our system is HMM-based which means it has a good intelligibility rate but not so good naturalness assessment, especially if compared against unit selection based systems; 2) in our system, voice quality improves dramatically when trained with at least 5000 sentences. In this challenge, only 1217 utterances were provided, which explains the bad results: 3) the listeners' profiles are different: we used more listeners, balanced in gender and explicitly with no experience with speech technology, whereas in Albayzin listeners were in less number and preferably experts in speech technology.

5. Conclusions and future work

When leveraging Microsoft Helena for Exchange 2010, we had very good MOS-scale results (which cannot be published for confidentiality reasons), especially regarding intelligibility of synthetic speech. The success rates of our system's intelligibility, when compared with other available systems, may be explained mainly by two reasons: the HTS technology enabled in the back-end, which largely increases intelligibility by making the segmental phone transitions smoother, and the application of several rule-based modules in the front-end, which allows a better accuracy rate in the grapheme-phoneme conversion. Another key aspect that has a significant impact in the synthetic speech voice quality is the choice of the voice talent or the professional speaker. In Speech at Microsoft, we have an accurate process to select the voice talent and to control voice quality with objective and acoustic measurements, which will be published soon. Preliminary work on this topic can be found here: [15], [16]. Albayzin results presented here are not consistent with our internal results and this difference will be investigated. The reasons for the bad results of our system in Albayzin may be related with the fact that most of the systems presented in the challenge are unit selection based, our systems requires more training data and listeners' profiles are different from the ones we used in our internal assessment. More work on naturalness and expressiveness of synthetic speech is ongoing.

6. References

[1] Francisco Moreno Fernández; Jaime Otero Roth. *Demografía de la lengua española.* Instituto Complutense de Estudios Internacionales y Fundación Telefónica. 2006. Spain.

[2] https://www.cia.gov/library/publications/the-world-factbook/fields/2098.html last visited on: 15/10/10

[3] http://es.wikipedia.org/wiki/Anexo:Hablantes_de_espa%C3%B1ol_como_lengua_materna_en_el_2003_(seg%C3%BAn_el_Britannica_Book) last visited on: 15/10/10

[4] Eduardo R. Banga; Francisco Méndez; Francisco Campillo; Gonzalo Iglesias; Laura Docío. *Descripción del sintetizador de voz Cotovía para la evaluación Albayzin 2008.* Grupo de Teoría de la Señal. Universidad de Vigo. 2008. Spain.

[5] Antonio Bonafonte; Ignasi Esquerra; Albert Febrer; José A. R. Fonollosa; Francesc Vallverdú. *The UPC Text-to-Speech System for Spanish and Catalan.* Universitat Politècnica de Catalunya. 1998. Spain.

[6] http://dihana.cps.unizar.es/investigacion/voz/ctv.html last visited on: 15/10/10

[7] R. Barra-Chicote; J. Yamagishi; J. M. Montero; S. King; S. Lufti; J. Macias-Guarasa. *Generación de una voz sintética en castellano basada en HSMM para la evaluación Albayzín 2008: Conversión texto a voz.* Universidad Politécnica de Madrid, University of Edinburgh, Universidad de Alcalá. 2008. Spain.

[8] http://www.cepstral.com/ last visited on: 15/10/10

[9] http://www2.research.att.com/~ttsweb/tts/demo.php last visited on: 15/10/10

[10] http://tts.loquendo.com/ttsdemo/default.asp last visited on: 15/10/10

[11] http://www.verbio.com/webverbio3/html/demos_ttsonline.php last visited on: 15/10/10

[12] http://www.scansoft.com/realspeak/languages/ last visited on: 15/10/10

[13] http://www.acapela-group.com/text-to-speech-interactive-demo.html last visited on: 15/10/10

[14] Microsoft Language Development Center website – Downloads section (http://www.microsoft.com/portugal/mldc/downloads.mspx)

[15] Daniela Braga, Luis Coelho, Fernando Gil Resende Jr., and Miguel Dias. 2007. "Subjective and Objective Assessment of TTS Voice Font Quality", XII International Conference Speech and Computer - SPECOM 2007, October 15-18, 2007, Moscow, Russia. ISBN 6-7452-0110-x

[16] Daniela Braga, Luis Coelho, Fernando Gil Resende Jr., and Miguel Dias. 2007. "Subjective and Objective Evaluation of Brazilian Portuguese TTS Voice Font Quality", Kacic, Zdravko and Markus, Aleksandra (eds) Advances in Speech Technology – International Workshop. June 28-29, 2007. Maribor, Slovenia. ISBN: 978-961-248-105-6. pp. 129-138.

[17] Francisco Méndez Pazó, Laura Docío-Fernández, Montserrat Arza Rodríguez, Francisco Campillo Díaz, "The Albayzín 2010 Text-to Speech Evaluation", Proceedings of Fala2010, 2010.

Adaptation of the URL-TTS system to the 2010 Albayzin Evaluation Campaign

Lluís Formiga, Alexandre Trilla, Francesc Alías, Ignasi Iriondo and Joan Claudi Socoró

GTM - Grup de Recerca en Tecnologies Mèdia.
La Salle - Universitat Ramon Llull. C/Quatre Camins 2, 08022 Barcelona, Spain
{llformiga,atrilla,falias,iriondo,jclaudi}@salle.url.edu

Abstract

This paper presents the text-to-speech (TTS) synthesis system of La Salle (Universitat Ramon Llull, URL) and its adaptation to the Albayzin Evaluation Campaign of FALA2010 conference. The URL-TTS system follows the classical scheme of unit selection TTS synthesis systems. However, it presents two distinguishable particularities: *i*) prosody prediction learned from labelled data by means of Case-Based-Reasoning (CBR) and perceptual weight tuning by means of active interactive Genetic Algorithms (aiGA). The aiGA-based weights are compared to multilinear regression (MLR) weights both considering classical averaged cost function and its root-mean squared variant. The internal validation tests and the results of the evaluation campaing are described, and finally discussed.

Index Terms: speech synthesis, unit selection, weight tuning, prosody prediction, interactive genetic algorithms, case-based reasoning

1. Introduction

The text-to-speech (TTS) synthesis system of the Grup de Recerca en Tecnologies Mèdia (GTM) of La Salle (Universitat Ramon Llull) (URL-TTS) is based on the original mid-90's second generation [1] Catalan concatenative TTS system, which considered diphones as basic units and *TD-PSOLA* for waveform generation [2, 3]. Subsequently, the system has been improved across years until the current unit selection TTS (US-TTS) synthesis system (see [4] for further details). The unit selection based URL-TTS synthesis engine presents two principal particularities (see figure 1): *i*) a case-based reasoning (CBR) prosody prediction module based on learning prosodic patterns from recorded corpora [5], and *ii*) a unit selection module, which integrates real human perceptual preferences through weights tuned by active interactive genetic algorithms (aiGA), which are adjusted at cluster level [6, 7, 8]. Moreover, great effort has been done to obtain automatic corpus development tools in order to speed up the set-up of the URL-TTS synthesis system for new voices [9]. This additional work involves features such as improving the selection of texts to be used during the recording process, including rules for avoiding ambiguity on phonetic transcription, refining unit segmentation [9] and reliable pitch marking [10]. In addition, there has been some further research focused on new acoustic parametrizations based on voice quality (VoQ) and harmonic plus noise models (HNM) [11], besides new approaches for expressive speech corpus parametrization [12]. All those improvements have been developed with the support of several research projects: SALERO, (IST-FP6-027122), SAVE (TEC2006-08043/TCM), evMIC (TSI-020301-2009-25).

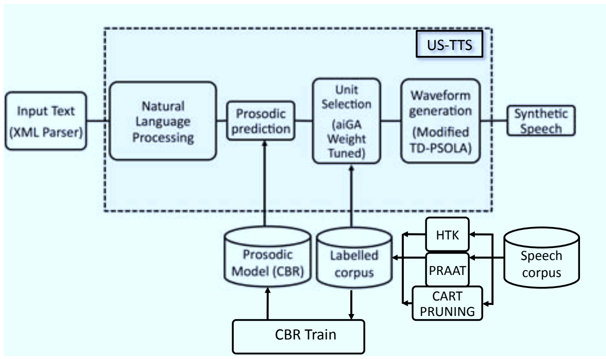


Figure 1: Block diagram of the URL-TTS synthesis system based on unit selection (US-TTS).

In this contest, we have incorporated some contributions with respect to the previous competition [4]: *i*), the speech corpus preparation includes a (quite simple) corpus pruning process based on detecting outlier voiced units with the aid of a of a clustering tool (the *wagon* tool of Festival [13]); *ii*) the prosody prediction module [5] has been incorporated in the TTS system, providing a richer prosodic reference for driving the unit selection process. This module has been used to guide the unit selection module only, and it has not been applied for conducting the posterior *TD-PSOLA*-based signal processing modification. In contrast, it has been considered the use of natural prosody of the retrieved acoustic units. This decision has been taken to recover the natural micro-prosody of units while minimizing the need of signal manipulation, following the classical idea of US-TTS systems to "*choose the best to modify the least*" [14]. Regarding the unit selection module, *iii*) the selection process has been updated through the use of 14 subcosts and 3 types of parameters (acoustic and linguistic parameters in target subcosts, and acoustic parameters in concatenation subcosts), and *iv*) the cost function weights are adjusted by using a three-stage process, which involves clustering and perceptual weight tuning [8]. Finally, *v*) TD-PSOLA is used for the waveform generation, minimizing both pitch and energy discontinuities around concatenation points.

This paper is organized as follows. Sections 2, 4, 5 and 6 describe the main modules of the current URL-TTS synthesis system based on unit selection. Section 3 is devoted to the Albayzin 2010 Evaluation Campaign corpus preparation process. The internal validation tests and the evaluation campaing results are presented in section 7, and conclusions and future work are outlined in sections 8 and 9.

2. Phonetic transcription

The *phoneme-set* used by the URL-TTS system is derived from SAMPA [15]. The phonetic transcription module consists of a rule-based system [16]. The rules are applied on a data structure that is a list of grapheme-phoneme pairs within a statement. It is possible to use insertion (I) or deletion (D) rules. The rules are applied only when the evaluation (E) of a phoneme characteristic yields a positive result.

$$E(gr ==' h') \quad \rightarrow \quad D(gr) \tag{1}$$
$$E(gr ==' x') \quad \rightarrow \quad I(/ks/) \tag{2}$$

Rule (1) indicates that the grapheme (gr) $'h'$ must be deleted, while rule (2) indicates that the grapheme $'x'$ must be transformed into the phonemes pair $/ks/$, thus implying a phonetic insertion. Regarding the exceptions, the system includes a dictionary that is consulted before applying any rule.

3. Creation of Unit Inventory

The voice of this competition (internally named as uvig_da_es) has been the 4th Spanish voice adapted to the URL-TTS synthesis system. Previously, we created the url_sam_es voice for weather forecasting restricted domain, the url_pat_es emotionated voice with 5 basic emotions (anger, joy, neutral, sad and sensual), and adapted the upc_ma_es voice for the 2008 Albayzin competition. It has to be added that url_pat_es voice is the main voice for the Spanish version of the URL-TTS synthesis system, which has been used in the projects mentioned in the introduction. In addition, it is worth noting that URL-TTS system is multilingual. The system also supports 2 Catalan voices (upc_pau_ca and upc_ona_ca from FesCat) and 4 English voices (url_sam_en, url_pgp_en, url_lau_en and url_rog_en) making a total of 8 public voices available. All of them may be tested on the GTM public website[1].

3.1. Segmentation and Labelling

3.1.1. Phonetic segmentation with pauses detection

In the segmentation process, the speech corpus is labelled indicating the temporal limits at the phoneme level. Our research group has been working to improve the segmentation process in recent years, in terms of the quality of the labelling process, the ease of use with the inclusion of user interfaces and language independence. Presently, the training and the posterior segmentation processes are based on Hidden Markov Models (HMM). To this end, a proprietary *Matlab*® code has been developed, also using the HTK tool (Hidden Markov Model Toolkit) [17].

The corpus that has been provided for the 2010 Albayzin Evaluation Campaign has been recorded by a male in neutral voice. It consists of 1217 utterances, 17797 word instances and a total vocabulary size of 5465 words. Regarding its analysis for the competition, the apparition and omission of silences have been controlled. Therefore, the pauses are correctly set according to the text. Alternatively, occlusive sounds are treated in special so that voice bursts and the previous silences are modeled as different units. At the end of the process we have obtained 826 different diphones with a total of 88571 diphones on the corpus.

[1] http://www.salle.url.edu/portal/departaments/home-depts-DTM-projectes-demos

3.1.2. Pitch marking

The PRAAT tool [18] has been used for signal pitch marking. It performs an acoustic periodicity detection on the basis of an accurate autocorrelation method [19]. In a first step, the voiced parts of the spoken utterance are pitch marked using this procedure. The pitch mark values are allowed to range between 75 and 600 Hz. In a second step, the unvoiced parts of the spoken utterance are given a sequence of pitch marks corresponding to the linear interpolation between the values of the previous pitch mark and the following one.

3.2. Corpus *pruning*

The process of recording and automatically labeling (segmentation and pitch marking) a speech corpus is prone to make errors. During the recording process, the speaker may introduce variants in the pronunciation or changes in the speed of delivery. Hence, elisions may be performed by speeding-up the speaking rate, or breaks may be introduced in the case of slowing it down, among others. A low-rate error labeling process is crucial for the general success of our US-TTS synthesis system, since the unit selection process itself is not capable of guaranteeing the retrieval of an error free unit sequence. In contrast to considering an exhaustive manual revision, a quite simple pruning process that attempts to detect errors in the recording and labeling phases has been implemented. In this work, the pruning has been performed at the phoneme level, by only considering voiced phonemes as they present more consistent parameters for the analysis. For each phoneme, the pruning process takes into account its prosodic parameters (pitch, energy and duration) and the first 3 spectral formants (obtained with [18]). Next, the 6-dimensional space (3 prosodic dimensions plus 3 spectral dimensions) is clustered using the *wagon* tool of Festival [13]. Once the phoneme groups are defined, the labelled phonemes out of their corresponding region are removed. As a result, 4908 recorded units are removed from the overall 88571 units (i.e. a 5.54% corpus size reduction).

4. Prosody Prediction

The URL-TTS synthesis system incorporates a corpus-based method for the quantitative modelling of prosody [5], following the case-based reasoning (CBR) algorithm proposed by [20]. This module predicts three main prosodic parameters: the fundamental frequency (F0) contour, the segmental duration and the energy, with the purpose of guiding the unit selection.

The automatic extraction of prosodic features from text starts from our linguistic analysis tool [21]. It carries out the phonetic transcription of text (based on SAMPA), annotating intonation groups (IG), stress groups (SG), words and syllables. The IG in Spanish is defined as a structure of coherent intonation that does not include any major prosodic break [22]. Prosodic breaks take place due to pauses or significant inflections of the F0 contour. The SG is defined as a stressed word preceded by one or more unstressed words, if they appear.

For the F0 contour modelling, the SG has been chosen following the proposal of [23]. The SG incorporates the influence of the syllable (it includes one stressed syllable plus some unstressed ones) and the pitch structure at IG level is achieved by the concatenation of SG contours. However, this model lacks variations due to micro-intonation. Up to now, we only differentiate between declarative, exclamatory, interrogative and suspended/unfinished IGs [24], which can be reliably identified from punctuation signs. Another attribute is the placement of the tonic syllable in the SG. Finally, other considered attributes are the number of syllables of the SG and the positions of the SG relative to the IG and the sentence.

A quantitative representation of the F0 contour has been used, by means of the coefficients of the polynomial that minimizes the error between the original set of points and the polynomial. Therefore, F0 parameters consist of the coefficients of the polynomial that are adjusted to minimize the distance between the polynomial and a collection of points that represent the value of the average F0 of every phoneme. This mean value of F0 is referenced to the centre of each phoneme of the IG.

For segmental duration and energy modelling, the phoneme has been chosen the basic acoustic unit (as [25, 26]). These parameters depend on basically the phoneme identity and the context where it is placed (attributes related to position and stress).

5. Unit Selection

5.1. Framework

The unit selection module follows the classical scheme described by Hunt and Black in [27]. The corpus units are retrieved by means of the Viterbi dynamic programming algorithm [28], which seeks the best sequence of units by minimizing a cost function. This cost function is defined as a weighted sum of several normalized subcosts (see equation (5)). In general terms, these subcosts are composed of target and concatenation measures [27]. For each possible candidate unit, target subcosts measure the difference between the ideal unit on that position (either by linguistic definition or prosodic prediction) and the candidate unit. Moreover, for each possible pair of candidate units, concatenation subcosts measure the acoustic discontinuity at the concatenation point.

Thus, the unit selection cost function of unit i jointly with unit j is defined by the following equations:

$$C_T(i) = \sum_{k=0}^{param\_t} w_T^k \cdot {SC_T}^k(i) \quad (3)$$

$$C_C(i,j) = \sum_{k=0}^{param\_c} w_C^k \cdot {SC_C}^k(i,j) \quad (4)$$

$$C(i,j) = C_T(i) + C_C(i,j) \quad (5)$$

where ${SC_T}^k(i)$ and ${SC_C}^k(i,j)$ represent target and concatenation subcosts, which are weighted by w_T^k and w_C^k, respectively, and they are computed as:

$${SC_T}^k(i) = D\left[P(u_i)^k, P(t_i)^k\right] \quad (6)$$

$${SC_C}^k(i,j) = D\left[P(u_i^R)^k, P(u_j^L)^k\right] \quad (7)$$

where u_i is the candidate unit, t_i is the target unit, u_i^R is the parametrization on the right concatenation point of the candidate unit and u_j^L is the parametrization of the left concatenation point of the candidate unit. $D[\cdot,\cdot]$ is the distance function (Manhattan, euclidean, cubic, etc.) and $P(\cdot)^k$ is the measured value of parameter k for the corresponding unit.

Moreover, for this particular competition, we wanted to analyse the effects of changing classical averaged cost function (AVG) (see equation (5)) [27] for the root mean squared (RMS) cost function variant proposed in [29]. RMS cost function considers quadratic weighted sum of different subcosts instead of computing the lineal weighted sum of subcosts (see equation (10)).

$$C_T(i) = \sum_{k=0}^{param\_t} \left(w_T^k \cdot {SC_T}^k(i)\right)^2 \quad (8)$$

$$C_C(i,j) = \sum_{k=0}^{param\_c} \left(w_C^k \cdot {SC_C}^k(i,j)\right)^2 \quad (9)$$

$$C(i,j) = \sqrt{C_T(i) + C_C(i,j)} \quad (10)$$

In terms of target subcosts, we consider four acoustic subcosts (pitch, energy and left/right half phone durations) and seven linguistic subcosts (position in utterance, position in word, position in syllable, previous and next phonemes, part-of-speech and syllable stress). That makes a total of 11 target subcosts. As concatenation subcosts, we consider discontinuity of pitch, energy and cepstral coefficients at the concatenation point. Cepstral distance is computed considering the first 12 Mel-Cepstral coefficients along their derivatives. Overall, the cost function is composed by 14 subcosts of 3 different types (acoustic and linguistic for target and acoustic for concatenation subcosts).

5.2. Weight Tuning

The weights w_T^k and w_C^k of the cost function are tuned by a 3-step process:

i) Automatic weight tuning is performed using Multilinear Regression (MLR) [27, 30]. In order to avoid negative weight values, we used non-negative least squares implementation [31]. For each recorded unit in the corpus, MLR performs regression across the 20 acoustically nearest units considering the cepstral distance and their related subcosts.

ii) Once unit weights are automatically tuned at unit level, in a second phase, these weights are clustered by expectation maximisation (EM) algorithm in order to obtain weight patterns for each cluster [32]. EM is chosen since it is the method that obtains better validation clustering indices [33]. Afterwards, phonetic and linguistic information of each unit is mapped to weight patterns clusters by means of a classification and regression tree (CART). At this point we have weight patterns at cluster level, where the cluster is defined by linguistic and phonetic specifications.

iii) In the final stage, the weights for each cluster are tuned perceptually. The number of clusters is set to 5 after reaching a consensus among different validity indices [34]. Once the groups of units are defined, four representative sentences of each cluster (mainly containing units of that cluster) are selected. The utterances are chosen through an entropy maximization algorithm [35]. These 20 sentences (4 sentences for each of the 5 clusters) are then used for conducting the perceptual weight tuning process based on active interactive Genetic Algorithm (aiGA), following the scheme described in [8]. It is worth noting that no prediction of prosody is considered for the weight tuning, assuming an ideal process by extracting the prosody values of the target sentence. Finally, the aiGA-based weights are obtained and a new CART tree is built for determining the final perceptual weights pattern per cluster.

6. Waveform Generation

The waveform generation process included in the URL-TTS synthesis system is based on *TD-PSOLA* [36]. In that original work, all units are pitch-synchronously resynthesized overlapping their frames in order to match the duration and pitch of the target unit sequence. Discontinuities of pitch are minimized by interpolating pitch marks around the concatenation points between units that are not consecutive in the corpus. In this work, informal listening tests have shown that the synthetic speech quality is better when the target F0 and duration are recovered from the corpus instead of considering the CBR-based prosodic prediction.The original pitch marks structure is kept in the speech segments generated from units that are consecutive in the corpus. At each concatenation point the signal frames are interpolated, following new pitch marks values in order to achieve a smoother pitch contour. Also, signal amplitude adjustment is conducted to avoid energy discontinuities.

7. Experiments

In this section, the experiments conducted to set-up the URL-TTS synthesis system and the 2010 Albayzin Evaluation Campaign results are described. The validation experiments are perceptual tests considering Mean Opinion Score (MOS) [37]. Some investigations [38, 39] state that pairwise direct comparison (pairwise preference tests) overcomes MOS in terms of obtaining preference for final users in the case of comparing similar systems. To that effect, we adapted the classical MOS methodology to a double stimuli input in order to obtain the advantages of both methods. That is, the same input utterance was presented to the user synthesized by two different TTS system configurations, but the user had to rank them independently instead of choosing which one was the best. For testing the stimuli, we used the TRUE platform [40], which is capable to perform MOS, pairwise comparison tests or both at the same time. After presenting the validation tests, the results collected from the evaluation campaing are described and discussed.

7.1. Validation of weights with copy-prosody

Once the weights have been perceptually tuned, they are submitted to a subjective validation process to confirm their appropriateness. We consider the weights obtained by MLR [30] as the baseline for validating the aiGA-based weights.

To that effect, 20 utterances different from the ones involved in the perceptual adjustment were chosen from the speech corpus to be part of a preference test. The utterances were synthesized by 4 different unit selection configurations (aiGA-rms, aiGA-avg, MLR-rms, MLR-avg). aiGA/MLR identifies the weights used and rms/avg identifies the cost function involved in the unit selection process. The original recorded prosody from the utterance (copy-prosody) was used, as done during the perceptual weight tuning stage. In addition, the units composing the utterance were removed from the corpus in order to avoid the selection of those units, and thus, obtain a more reliable evaluation of the compared unit selection processes. Moreover, the natural recorded version of the utterance was also presented to the evaluators along with each pair of stimuli in order to provide an ideal *target*.

Six evaluators participated in the validation tests, obtaining the results depicted in figure 2. As it can be observed, better synthesis is achieved by aiGA-based methods: their corresponding averaged MOS results are 3.54 for aiGA through AVG cost function and 3.41 through RMS cost function, although if

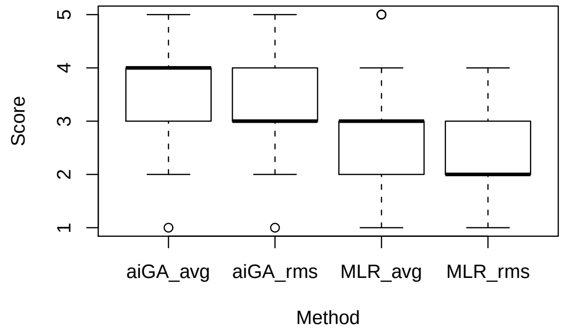


Figure 2: Internal MOS results comparing different weight tuning (aiGA / MLR) and two different integration cost functions (averaged vs. root-mean squared) when the target prosody is extracted from the recorded units.

the Bonferroni correction method is applied to test the significance of the results [41], we can conclude that their difference (0.13) is not statistically significant ($p = 0.743$). MLR-based weights behave significantly worse than the perceptual weights within both cost functions ($p < 0.001$). However, the AVG cost function computed with the MLR weights (MOS: 2.94) behaves slightly better than RMS cost function (MOS: 2.36) with their difference (0.58) being statistically significant ($p < 0.001$). As a last step, we also analyzed the pairwise comparison significance through a signed ranked test and we obtained the same results.

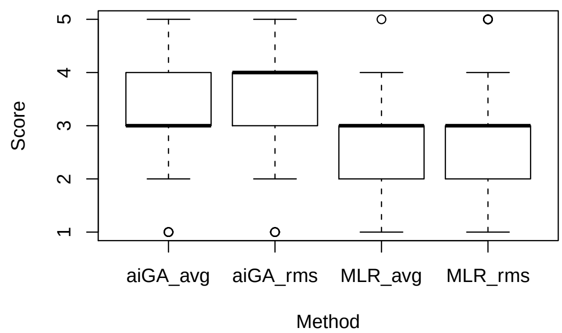


Figure 3: Internal MOS results comparing different weight tuning (aiGA / MLR) and two different integration cost functions (averaged vs. root-mean squared) when the target prosody is predicted by the CBR-based technique.

7.2. TTS final adjustment

In order to test the performance of the whole TTS synthesis system, we incorporated the CBR-based prosody prediction module with 20 utterances selected from the 2010 Albayzin Evaluation Campaign sets. As no natural prosody was available at that time, the natural recorded sentence was not presented to the evaluation users.

The same six evaluators participated in the final system validation tests, obtaining the new results depicted in figure 3. Again, better synthesis is achieved by aiGA-based methods: their corresponding averaged MOS values are 3.50 for aiGA through RMS cost function and 3.35 through AVG cost func-

Table 1: Groups detected by Bonferroni pairwise analysis

Group	Weight Tuning	Cost Function	Prosody	MOS Score
1	aiGA	AVG	COPY	3.54
	aiGA	RMS	CBR	3.50
	aiGA	RMS	COPY	3.41
	aiGA	AVG	CBR	3.35
2	MLR	AVG	COPY	2.94
3	MLR	RMS	CBR	2.61
	MLR	AVG	CBR	2.56
	MLR	RMS	COPY	2.36

tion, although their difference (0.15) is not statistically significant ($p = 0.517$). MLR-based weights again behave significantly worse than the aiGA-based weights within both cost functions ($p < 0.001$). However, in this case, the difference of averaged MOS values (0.05) between RMS (MOS: 2.56) and RMS cost functions (MOS: 2.61) is not statistically significant ($p < 1$).

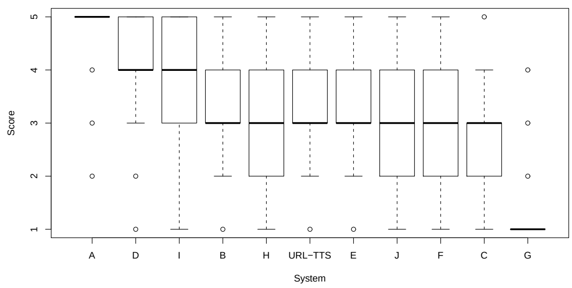
Next, the effects of including CBR-based prosody prediction to the unit selection module are discussed. The obtained results (copy-prosody and CBR-based prosody) were analyzed simultaneously. To that effect, we applied Bonferroni pairwise analysis in order to identify groups on the MOS evaluations. Groups are defined by configurations with no significant differences among them. The analysis found three groups in terms of the MOS results, as it can be seen on table 1. It can be observed that CBR-prosody prediction does not introduce major alteration to the copy-prosody results. Thus, the determinant factor for the URL-TTS synthesis system based on unit selection to obtain high quality speech is the weight tuning methodology. Under MLR-weight tuning methodology, synthesis with artificial (CBR-based) prosody is unable to reach the quality of natural prosody. Nevertheless, this difference is overcome by the aiGA-based weights.

As a result, the system presented to the 2010 Albayzin competition includes CBR prosody prediction, aiGA-based weight tuning and RMS cost function (as presented slight better results than AVG cost function, although not significant). This configuration achieved a MOS score of > 3.30 in the validation experiments.

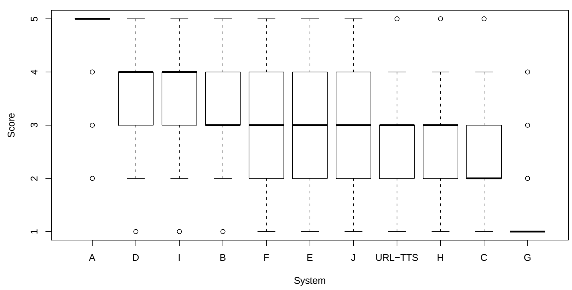
7.3. 2010 Albayzin Evaluation final results

Once the set-up of the system was completed, 400 synthesized sentences were presented to the 2010 Albayzin Evaluation Campaign. This evaluation campaign consisted of 3 separate analyses in order to assess different aspects of the evaluated TTS synthesis systems: *i*) similarity to the original recorded voice, *ii*) overall quality through mean opinion scores (MOS), and *iii*) intelligibility by computing word error rate (WER) on sentences composed of random words (i.e. with no clear meaning). The number of users involved in each test was substantially different depending on the test. Whether around 541 users were involved in the MOS test (see figure 4(b)), only around 137 users were involved in the voice similarity tests (see figure 4(a)) and 182 were involved in the WER tests.

In terms of similarity to the recorded voice, the URL-TTS system performs quite well since it is a US-TTS synthesis system, yielding a similar MOS value to the internal validation tests (average MOS$= 3.20$). However, on the overall quality MOS test, the URL-TTS decreases its score to 2.62. This significant decrease compared to the obtained MOS results may be motivated to several factors. Firstly, no natural voice was used on the validation tests, which makes the results not comparable. Secondly, few evaluators conducted the internal validation tests considering only relative improvements instead of considering the quality of other TTS systems. Finally, the URL-TTS synthesis system presents several intelligibility problems, reflected with a poor WER (0.31). This factor maybe caused by the presence of artifacts, that definitely affects the overall preceived synthetic speech quality. It is worth noting that, besides including a pruning process, the corpus creation has been fully automatic with no manual intervention at any stage of the process (neither using the given labelings or transcriptions).



(a) *Similarity to natural voice*



(b) *Overall quality*

Figure 4: FALA2010 results through different systems [42]

8. Conclusions

This paper describes the main advances included in the URL-TTS synthesis system with respect to the previous 2008 Albayzin competition. The two key elements are the CBR-based prosody model and the aiGA-based weight tuning. After several perceptual experiments, the URL-TTS synthesis system has obtained acceptable internal validation (MOS> 3.30) and similarity to the natural voice (MOS$= 3.20$) results. However, there has been a decrease on the overall quality according to the evaluation campaign results (MOS$= 2.64$), where the URL-TTS synthesis system has been challenged against to other TTS systems and some intelligibility problems (WER$= 0.31$). In favor of URL-TTS system, it is worth noting that these results were obtained after reasonable reduced time for the TTS set-up and tuning, thanks to the fully automatic voice building tools and tuning platforms.

In terms of the weight tuning of the cost function, it can be concluded that weight tuning is one of the key factors in order to obtain good synthetic speech quality for the US-TTS synthesis system at hand. In addition, the results present a significant improvement when considering perceptual tuned weights (aiGA-based) with respect to using automatically trained weights (MLR-based). However, the substitution of the cost function from averaged to root-mean squared does not yield notable quality changes. Moreover, the perceptual results obtained after including the CBR-based predicted prosody during the TTS execution remain almost unaltered. However, it is worth noting that other key factors for obtaining high quality synthetic speech through US-TTS synthesis (e.g. segmentation and pitch marking, pruning methodology, waveform generation, etc.) have not been explicitly analyzed in this paper, leaving their analysis and optimization for future works.

9. Future work

Future work will be focused on improving the intelligibility and naturalness of the URL-TTS synthesis system, improving the corpus building tools and revising the database pruning process accordingly. In addition, this work will be focused on improving synthesis flexibility so as to modify the speech identity and expressiveness. In this regard, we are currently working on adapting an HNM (Harmonic-plus-Noise Model) library to the current US-TTS synthesis system. The main objectives are: *i*) considering the CBR prosody predictions, besides improving the quality of concatenations smoothing (avoiding artifacts), by fully exploiting the potentialities of the HNM through interpolation techniques, and *ii*), gradually improving the flexibility of the system (i.e. using speech conversion methods) but keeping the final synthesis similarity to natural voice as high as possible.

10. References

[1] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[2] J. Camps, G. Bailly, and J. Martí, “Synthèse à partir du texte pour le catalan,” in *Proc. 19èmes Journées d'Études sur la Parole*, Bruselas, Francia, 1992, pp. 329–333.

[3] R. Guaus, F. Gudayol, and J. Martí, “Conversión textovoz mediante síntesis PSOLA,” in *Jornadas Nacionales de Acústica*, Barcelona, España, 1996, pp. 355–358.

[4] C. Monzo, L. Formiga, J. Adell, I. Iriondo, F. Alías, and J. Socoró, “Adaptación del CTH-URL para la competición Albayzin 2008,” *V Jornadas en Tecnología del Habla*, pp. 87–90, 2008.

[5] I. Iriondo, J. C. Socoró, and F. Alías, “Prosody Modelling of Spanish for Expressive Speech Synthesis,” in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Honolulu, EUA, Abril 2007, pp. 821–824.

[6] F. Alías, X. Llorà, L. Formiga, K. Sastry, and D. E. Goldberg, “Efficient interactive weight tuning for TTS synthesis: reducing user fatigue by improving user consistency,” in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, Toulouse, Francia, 2006, pp. 865–868.

[7] L. Formiga and F. Alías, “Extracting User Preferences by GTM for aiGA Weight Tuning in Unit Selection Text-to-Speech Synthesis,” in *Computational and Ambiental Intelligence - Proceedings on 9th International Work-Conference on Artificial Neural Networks (IWANN)*. San Sebastián, Spain: Springer (LCNS), June 2007, pp. 654–661.

[8] L. Formiga, F. Alías, and X. Llorà, “Evolutionary process indicators for active IGAs applied to weight tuning in unit selection tts synthesis,” in *IEEE Conference on Evolutionary Computation*. Barcelona, Spain: IEEE, July 2010, pp. 2322–2329.

[9] G. Kienast, G. Thallinger, R. Fach, S. M. Freixes, O. Mayor, T. Bürger, M. Yan, T. Stolt, R. Villa, M. Romeo, C. Goodman, M. Matthews, and L. Formiga, “Third annual on-line public report,” http://www.salero.eu/media/pdf/del/SALERO-D10.5.6-PublicAnnualReport2008.pdf, Semantic AudiovisuaL Entretainment Reusable Objects, Deliverable 10.5.6, 2008.

[10] F. Alias and N. Munne, “Reliable Pitch Marking of Affective Speech at Peaks or Valleys Using Restricted Dynamic Programming,” *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 481–489, 2010.

[11] C. Monzo, À. Calzada, I. Iriondo, and J. Socoró, “Expressive Speech Style Transformation: Voice Quality and Prosody Modification Using a Harmonic plus Noise Model,” *Proceedings of Fifth International Conference on Speech Prosody, Chicago, USA*, 2010.

[12] I. Iriondo, S. Planet, J. Socoró, E. Martínez, F. Alías, and C. Monzo, “Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification,” *Speech Communication*, vol. 51, no. 9, pp. 744–758, 2009.

[13] A. Black and P. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” in *Proceedings of EuroSpeech*, Rhodes, Greece, 1997, pp. 601–604.

[14] M. Balestri, A. Paechiotti, S. Quazza, P. L. Salza, and S. Sandri, “Choose the best to modify the least: a new generation concatenative synthesis system,” in *Proceedings of EuroSpeech*, vol. 5, Budapest, Hungary, 1999, pp. 2291–2294.

[15] J. C. Wells, *SAMPA computer readable phonetic alphabetHandbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter, 1997, ch. SAMPA computer readable phonetic alphabet, pp. Part IV, section B.

[16] I. Iriondo, “Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva,” Ph.D. dissertation, Universitat Ramón Llull, 2008.

[17] “HTK,” in *Recuperado el 19 de 09 de 2008, de http://htk.eng.cam.ac.uk*, 2008, pp. 149–150.

[18] P. Boersma and D. Weenink, “Praat: doing phonetics by computer (version 5.1.37),” http://www.fon.hum.uva.nl/praat/, 2010, as of July 1, 2010.

[19] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proceedings of the Institute of Phonetic Sciences 17*, Amsterdam, The Netherlands, 1993, pp. 97–110.

[20] A. Aamodt and E. Plaza, “Case-based reasoning: foundational issues, methodological variations, and system approaches,” *Artificial Intelligence Communications*, vol. 7, no. 1, pp. 39–59, 1994.

[21] S. Sánchez, “Sincat/2. lenguaje para la conversión grafema-fonema,” Enginyeria i Arquitectura La Salle, Tech. Rep., 1997.

[22] J. M. Garrido, “Modelling spanish intonation for text-to-speech applications,” Ph.D. dissertation, Departament de Filologia Espanyola. Facultat de Lletres. Universitat Autònoma de Barcelona, 1996.

[23] D. Escudero and V. Cardeñoso, “Applying data mining techniques to corpus based prosodic modeling,” *Speech Communication*, vol. 49, no. 3, pp. 213–229, March 2007.

[24] F. Campillo and E. Rodríguez, “A method for combining intonation modelling and speech unit selection in corpus-based speech synthesis systems,” *Speech Communication*, vol. 48, no. 8, pp. 941–956, 2006.

[25] E. Navas, I. Hernáez, and J. M. Sánchez, “Modelo de duración para conversión texto a voz en euskera,” *Procesamiento del Lenguaje Natural*, vol. 29, pp. 147–152, 2002.

[26] J. P. Teixeira and D. Freitas, “Evaluation of a segmental durations model for tts,” in *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, ser. Lecture Notes in Computer Science, N. Mamede, J. Baptista, I. Trancoso, and M. Nunes, Eds., vol. 2721. Heidelberg: Springer, 2003, pp. 40–48.

[27] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, Atlanta, EUA, 1996, pp. 373–376.

[28] A. Viterbi, "Error bounds for convolutional codes and an asymtotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–267, 1967.

[29] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis," in *Speech Communication, Elsevier*, vol. 48, no. 247, 2006, pp. 45–56.

[30] Y. Meron and K. Hirose, "Efficient weight training for selection based synthesis," in *Proceedings of EuroSpeech*, vol. 5, Budapest, Hungary, 1999, pp. 2319–2322.

[31] C. Lawson and R. Hanson, *Solving least squares problems*. Society for Industrial Mathematics, 1995.

[32] C. Fraley and A. Raftery, "MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering," 2006.

[33] S. Günter and H. Bunke, "Validation indices for graph clustering," *Pattern Recognition Letters*, vol. 24, no. 8, pp. 1107–1113, 2003.

[34] K. Kryszczuk and P. Hurley, "Estimation of the Number of Clusters Using Multiple Clustering Validity Indices," *Multiple Classifier Systems LNCS*, pp. 114–123, 2010.

[35] J. Gauvain, L. Lamel, and M. Eskénazi, "Design Considerations and Text Selection for BREF, a large French read-speech corpus," in *First International Conference on Spoken Language Processing*. Citeseer, 1990.

[36] E. Moulines and F. Charpentier, "Pitch-Sinchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Speech Communication*, vol. 9, 1990, pp. 453–467.

[37] ITU-T, "Methods for subjective determination of transmission quality," Recommendation ITU-T P.800, Geneve, Suïssa, 1996.

[38] Y. Alvarez and M. Huckvale, "The reliability of the ITU-T P. 85 standard for the evaluation of text-to-speech systems," in *Seventh International Conference on Spoken Language Processing*. Citeseer, 2002.

[39] D. Sityaev, K. Knill, and T. Burrows, "Comparison of the ITU-T P. 85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems," in *Ninth International Conference on Spoken Language Processing*, 2006.

[40] S. Planet, I. Iriondo, E. Martínez, and J. Montero, "TRUE: an online testing platform for multimedia evaluation," in *Programme of the Workshop on Corpora for Research on Emotion and Affect*. Citeseer, 2008, p. 61.

[41] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, p. 800, 1988.

[42] F. Méndez Pazó, L. Docío-Fernández, M. Arza Rodríguez, and F. Campillo Díaz, "The Albayzín 2010 Text-to Speech Evaluation," in *Proceedings of Fala2010*, 2010.

Ogmios: the UPC entry for the Albayzin 2010 TTS Evaluation

Antonio Bonafonte[1], *Ignasi Esquerra*[1], *Asunción Moreno*[1], *Pablo Daniel Agüero*[2]

TALP Research Center, Universitat Politècnica de Catalunya, Spain [1]
Communications Lab, University of Mar del Plata, Argentina [2]

Abstract

This paper describes Ogmios, the UPC TTS system that was used in the 2010 Albayzin Evaluation. Ogmios is a concatenation system that builds the synthetic sentence from demiphones selected from the training database. In this evaluation round, the database was provided by the organization and it has been phonetically transcribed and segmented automatically using using the development tools included in Ogmios. Based on the segmentation the synthetic voice is build. Each voice includes the segmental inventory (waveforms of acoustic units), prosodic models (breaks, segmental duration and f0 prediction models) and the weights of the selection cost function. In order to have a better prosody models, in this evaluation external data has been used to ensure better prosodic coverage.

1. Introduction

This paper describes Ogmios, the UPC Text-to-Speech system used for the evaluation. The system was originally designed for Spanish and Catalan but has been extended to English and Mandarin [1, 2].

During the last 2 years, Ogmios has extended their functionalities: it is able to produce conversational speech to some degree [3] and to cope with multilingual text [4]. It also includes HTS as a possible back-end for speech synthesis. However, the system used in Albayzin'2010 is basically the same that was used in Albayzin'2008. We include here the description from [5] for completness. The main difference is that we have used prosody models estimated from other male speaker with a larger (10h) database. Our prosody algorithms, in particular, the phrassing component, did not perform well with the Albayzin'2010 database.

This paper is organised as follows: Section 2 describes the system and Section 3 explains the process of building the voices. Finally, section 4 describes the results of the evaluation.

2. System Description

2.1. Text and Phonetic Analysis

The first task of the system is to detect the structure of the document and to transform the input text into words. For this task we have used rules for tokenizing and classifying *non-standard words* in Spanish. The rules for expanding each token into *words* are language dependent, but are based in a few simple functions (spellings, natural numbers, dates, etc.) by means of regular expressions.

The second process is the POS tagger. Ogmios includes a statistical tagger based on FreeLing. The FreeLing package consists of a library providing language analysis services. Main services used of FreeLing library are PoS tagging and probabilistic prediction of unknown word categories. Freeling provides services for all currently supported languages: Spanish, Catalan, Galician, Italian, and English [6].

2.1.1. Phonetic Transcription

The goal of the *phonetic* module is to provide the pronunciation of the words. This is used not only for producing the test sentences but also for transcribing the training database which is used for building the voices.

For Spanish the pronunciation of each word is based on a set of rules that take into account the transcription rules of Spanish and phonotactics.

Some particular words are transcribed using a lexicon, specially foreign words, abbreviations and signs.

2.2. Prosody

Prosody generation is done by a set of modules that sequentially perform all the tasks involved in prosody modelling: phrasing, duration, intensity and intonation.

2.2.1. Phrasing

Phrasing is one of the key topics in the linguistic part of text-to-speech technologies and consists of breaking long sentences into smaller prosodic phrases. Boundaries are acoustically characterised by a pause, a tonal change, and/or a lengthening of the last syllable. Phrase breaks have strong influence on naturalness, intelligibility and even meaning of sentences.

In Ogmios phrasing is obtained using two algorithms. The first algorithm consists in a Finite State Transducer that translates the sequence of part-of-speech tags of the sentence into a sequence of tags with two possible values: break or non-break [7]. This uses the same tool

which was used for the grapheme-to-phoneme task: x-grams [8]. The method uses very few features, but the results are comparable to CART using more explicit features.
The second algorithm predicts phrase break boundaries combining a language model of phrase breaks [9] and probabilities of phrase breaks given contextual features [10]. Phrase break boundaries are found by maximizing the following equation:

$$J(C_{1,n}) = argmax_{j_{1,n}} \prod_{i=1}^{n} \frac{P(j_i|C_i)}{P(j_i)} P(j_i|j_{i-k,i-1}) \tag{1}$$

The latest algorithm was chosen in this evaluation for Spanish due to its better subjective performance in training data.

2.2.2. *Duration*

Phone duration strongly depends on the rhythmic structure of the language. For example, English is stressed-timed while Spanish is syllable-timed. Ogmios predicts phone duration with a two steps algorithm: prediction of the suprasegmental duration (syllable or stress unit), and then phone duration is predicted by factoring the suprasegmental duration.
The suprasegmental duration is predicted using CART. Features include the structure of the unit, represented by articulatory information of each phoneme contained in it (phone identity, voicing, point, manner, vowel or consonant), stress, its position in the sentence and inside the intonation phrase, etc.
Once the duration of the suprasegmental unit is calculated, the duration of each phoneme is obtained using a set of factors to distribute suprasegmental duration over its constituent phonemes. These factors are predicted using CART with a set of features extracted from the text, such as articulatory information of the phoneme itself and the preceding and succeeding ones, position in the unit, in the word and in the sentence, stress, and whether the unit is pre-pausal.

2.2.3. *Intensity*

The intensity of the phonemes is predicted by means of a CART. Features are again articulatory information of the actual, preceding and succeeding phone, stress, and the position in the sentence relative to punctuation and phrase breaks.

2.2.4. *Intonation*

Ogmios has two available intonation models: a superpositional polynomial model trained using JEMA (*Join feature Extraction and Modelling Approach* [11]), and a *f0 contour selection* model. In some cases, using the superpositional approach results in over-smoothed intonation contours with a loss of expressiveness.
Thus, in this evaluation we generate the f0 contour using the selection approach [12]. For each accent group we select a real contour from the database taking into account the *target cost* (position in the sentence, syllabic structure, etc.) and the *concatenation cost* (continuity). The selected contour is represented using a 4th order Bezier polynomial. The contour is generated using this polynomial, once the time scale is adapted to the required durations. The final result is a more expressive intonation contour than the JEMA model. However, in some cases, the contour is not adequate for the target sentence due to natural language understanding limitations of TTS systems.

2.3. Speech Synthesis

Our unit selection system runs a Viterbi algorithm in order to find the sequence of units $u_1 \dots u_n$ from the inventory that minimises a cost function with respect to the target values $t_1 \dots t_n$. The function is composed by a target and a concatenation cost: both of them are computed as a weighted sum of individual sub-costs as shown below:

$$\begin{aligned} C(t_1 \dots t_n, u_1 \dots u_n) = w^t \sum_{i=1}^{n} \left(\sum_{m=1}^{M^t} w_m^t C_m^t(t_i, u_i) \right) \\ + w^c \sum_{i=1}^{n-1} \left(\sum_{m=1}^{M^c} w_m^c C_m^c(u_i, u_{i+1}) \right) \end{aligned}$$

where w^t and w^c are the weights of the global target and concatenation costs ($w^t + w^c = 1$); M^t is the number of the target sub-costs and M^c the number of concatenation sub-costs; $C_m^t(.)$ is the m th target sub-cost which is weighted by parameter w_m^t; and $C_m^c(.)$ is the m th concatenation sub-cost weighted by w_m^c.
Tables 1 and 2 show the features used for defining the sub-cost functions. There are two types of sub-costs functions. Binary, which can only have 0 or 1 values, and continuous. For continuous sub-costs functions, a distance function is defined and a sigmoid function is applied in order to restrict their range to $[0 - 1]$.
To adjust the target weights, we applied a similar approach to the one proposed in [13]. For each pair of units, we compute their distance using feature vector (MFCC, f0, energy) taken every 5 msec. Let $\overline{d}$ be the vector of all distances for each pair of units, C a matrix where $C(i, j)$ is sub-cost j for unit pair i and $\overline{w}$ the vector of all weights to be computed. If we assume $C\overline{w} = \overline{d}$ then it is possible to compute $\overline{w}$ as a linear regression. In other words, the target function cost becomes a linear estimation of the acoustic distance. The weights of the concatenation sub-costs functions were adjusted manually.

phonetic accent	B
duration difference	C
energy difference	C
pitch difference	C
pitch diff. at sentence end	C
pitch derivative difference	C
pitch deviate sign is different	B
accent group position	B
triphone	B
word	B

Table 1: Target costs: B stands for binary cost and C for continuous cost.

energy	C
pitch	C
pitch at sentence end	C
spectral distance at boundary	C
voice-unvoiced concatenation	B

Table 2: Concatenation costs: B stands for binary cost and C for continuous cost.

Concerning the waveform generation process, in our experience, listeners assign higher quality scores to the synthetic utterances where the prosodic modifications are minimal. Thus, most of the units selected for generating synthetic speech are simply concatenated using glottal closure instant information, without any prosodic manipulation. Therefore, the use of the information provided by the prosody generation block is restricted to the unit selection process.

3. Building the Albayzin voice

Once the normalization and phonetic transcription rules are ready (section 2.1), our system is able to build a new voice automatically from the audio files and their corresponding prompts. This automatic procedure consists of four main steps: automatic segmentation of the database, training of the prosodic models, selection weights adjust plus database indexing. The prosody training and the selection weights adjust procedures have been described in previous sections. Therefore, in the present section, we will describe the segmentation process and the database indexing.

Once the database was supplied we built the unit inventory. In our system, the units are context dependent demiphones. However, the selection algorithm forces the use of diphones imposing a high cost in phone transitions. The database is automatically segmented into phones by means of the HMM-based aligner named Ramses [14]. We used the front-end described in section 2.1 to automatically transcribe the whole database into phones.

Afterwards, we trained a different set of context dependent demiphone HMM models from each data set, corresponding to each of the three voices. The phone boundaries are determined using a forced alignment between the speech signal and the models defined by the phonetic transcription. A silence model, trained at punctuation marks, was optionally inserted at each word boundary during the alignment. In addition, the detected silences are also used for the pause prediction model (see Section 2.2).

Previous experiments have shown that when a correct phonetic transcription is given, HMM models can achieve similar speech synthesis quality than manual segmentation [15, 16]. Therefore, additional effort was devoted to phonetic transcription and database pruning to obtain correctly segmented voices, as show in the following paragraphs.

Automatic phonetic transcription of a speech synthesis database has to cope with pronunciation variants, pronunciation errors and recording noise. In order to overcome the former problem, the alignment took into account all possible transcriptions of a single word. At this point, the alignment may have errors either because there is a mismatch between front-end and speaker production or because there is an alignment error.

We assume that wrong units will never represent a big portion of the database and that it is affordable to reject such part of it. Therefore we tried to detect undesired units in order to remove them from the inventory by means of a pruning procedure. After computing the alignment likelihood for every unit, 10% of them, those with worst scores, were removed. Previous experiments have shown that it is possible to remove 90% of wrong units by means of this pruning procedure [17].

In this evaluation we do not include any pruning due to the small amount of data provided to generate the synthetic voice. Therefore, we rely on spectral measures at unit selection to avoid problematic units.

Once the speech signals were segmented and the list of sentences are ready, we can start building the voices for our TTS system. The process consists of three main steps: feature extraction, unit indexing and voice generation. The first step extracts F0, duration, energy and MFCC for each speech unit. The index file contains the relevant information needed for computing the target and concatenation costs. In the last step, the parameters of the prosody models and the weights of the unit selection algorithm are computed.

4. Experimental Results and Discussion

All the voice models were estimated automatically from the voice using Ogmios training modules.

The similarity of the synthetic voice with the original one is relatively high, with mean value > 3. Even using prosodic models derived from other speaker, the listeners can perceive that the synthetic speaker is the same than the original voice. This should be a common feature to

any concatenative system. This question is more relevant for statistical or parametric methods.
With respect to naturalness, the MOS is relatively low, 2.60. We think that the reason is that even with models trained from data, the algorithms (segmentation, unit selection weights, signal processing) are tuned to larger databases. For smaller databases different tuning, or even different synthesis methods should be applied.

5. Acknowledgements

This work has been funded by the Spanish Government (project BUCEADOR: Busqueda de Informacion en Contenidos Audiovisuales Plurilingues TEC2009-14094-C04-01)
The authors thank to the Albayzin 2010 Team for organising the evaluation and providing the speech data.

6. References

[1] Bonafonte, A., Agüero, P. D., Adell, J., Perez, J., and Moreno, A., “Ogmios: The UPC text-to-speech synthesis system for spoken translation”, Proceedings of TC-STAR Workshop, Barcelona, Spain, June, 2006.

[2] Bonafonte, A., Moreno, A., Adell, J., Agüero, P.D., Banos, E., Erro, D., Esquerra, I., Perez, J., and Polyakova, T., “The UPC TTS System Description for the 2008 Blizzard Challenge”, Blizzard 2008, Brisbane, Australia, September, 2008.

[3] Adell, Jordi and Bonafonte, Antonio and Escudero, David, “Synthesis of Filled Pauses Based on a Disfluent Speech Model”, Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),Dallas, USA, March, 2010

[4] Polyakova, Tatyana and Bonafonte, Antonio, “"Nativization of English words in Spanish using analogy” Proceeding of the 7th ISCA Speech Synthesis Workshop, Kyoto, Japan, September 2010

[5] Bonafonte, Antonio and Agüero, Pablo Daniel , “The UPC TTS Systems Description”, Actas de las V Jornadas en Tecnologías del Habla, Bilbao, Spain,November 2008

[6] Atserias, J., Casas, B., Comelles, E., Gonzalez, M., Padro, L., and Padro, M., “FreeLing 1.3: Syntactic and semantic services in an open-source NLP library”, Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genoa, Italy. May, 2006.

[7] Bonafonte, A., and Agüero, P. D., “Phrase break prediction using a finite state transducer”, Proceedings of the 11th International Workshop on Advances in Speech Technology, Maribor, Slovenia, July, 2004.

[8] Bonafonte, A., “Language modeling using x-grams”, Proceedings of International Conference on Spoken Language Processing, 1996.

[9] Black, A., and Taylor, P., “Assigning Phrase Breaks from Part-of-Speech Sequences”, Proceedings of Eurospeech, 1997.

[10] Agüero, P. D., and Bonafonte, A., “ Phrase break prediction: a comparative study”, XIX Congreso de la Sociedad Española para el procesamiento del Lenguaje Natural. Alcala de Henares, Spain, September, 2003.

[11] Agüero, P. D. and Bonafonte, A., “Intonation Modeling for TTS Using a Joint Extraction and Prediction Approach,”, Proceedings of the International Workshop on Speech Synthesis, Pittsburgh, USA, 67-72, 2004.

[12] Malfrère, F., Dutoit, T., and and Mertens, P., “Automatic prosody generation using suprasegmental unit selection”, Proceeding of the 3rd ISCA Speech Synthesis Workshop, Jenolan Caves, Australia, December, 1998.

[13] Hunt, A., and Black, A., “Unit selection in a concatenative speech synthesis system using a large speech database”, Proceedings of ICASSP, Atlanta, Georgia, 1996.

[14] Bonafonte, A., Mariño, J. B., Nogeuiras, A., and Rodriguez Fonollosa, J. A., “RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC”, VIII Jornadas de Telecom I+D (TELECOM I+D ’98), Madrid, Spain, October, 1998.

[15] Makashar, M. J., Wightman, C. W., Syrdal, A. K., and Conkie, A., “Preceptual evaluation of automatic segmentation in Text-to-Speech synthesis”, Proceedings of ICSLP, Beijin, China, October, 2000.

[16] Adell, J., Bonafonte, A., Gómez, J. A., and Castro, M. J., “Comparative study of automatic phone segmentation methods for TTS”, Proceedings of ICASSP, Philadelphia, PA, USA, March, 2005.

[17] Adell, J., Agüero, P. D., and Bonafonte, A., “Database pruning for unsupervised building of text-to-speech voices”, Proceedings of ICASSP, vol. 1, Toulouse, France, May, 2006.

UEF-NTNU System Description for Albayzin 2010 Language Recognition Evaluation

R. Saeidi[1], *M. Soufifar*[2], *T. Kinnunen*[1], *T. Svendsen*[2], *and P. Fränti*[1]

[1] School of Computing, University of Eastern Finland, Joensuu, Finland
[2] Department of Electronics and Telecommunications, NTNU, Trondheim, Norway
{rahim,tkinnu,franti}@cs.joensuu.fi, {soufifar,torbjorn}@iet.ntnu.no

Abstract

We are describing University of Eastern Finland and Norwegian University of Science and Technology joint submission for Albayzin 2010 language recognition evaluation. We are employed several several approaches including acoustic and phonotactic based algorithms in our final submission. A short description of the systems are given.

Index Terms: Language Recognition, GLDS, GMM, MMI, VSM.

1. Submission Overview

Our submission for Albayzin 2010 is a score-level fusion of 3 sub-systems as follows:

- MMI-GMM
- GLDS-SVM-NN
- HMM-VSM-GMM
- PR-VSM-GMM

2. MMI-GMM

This system is build based on [1] and specifications are:

- SDC features of 49 dimension extracted.
- Language dependent 256 Gaussian GMMs are used for modeling.

3. GLDS

This system is build based on [2] and specifications are:

- SDC features of 49 dimension extracted.
- Up to 3rd order polynomial expansion used.
- Neural network with one hidden layer applied for language classifier. Other specifications are: Activation function tansig for hidden nodes and purelin for output layer, MSE measure and trained with trainlm (I used MATLAB terminology here).

4. PR-VSM-GMM

This system is build based on [3] and specifications are:

- Brno university phone recognizers [4] used here. There are 4 phone recognizers in their website. English phone recognizer trained on 16kHz TIMIT data which we used it for Albayzin evaluation.
- Up to 2-gram counts used and 300 dimensions retained after SVD.
- Two GMMs trained for each language in language classifier stage; one GMM for target scores (with 2 Gaussians) and another for non-target scores (with 20 Gaussians).

5. HMM-VSM-GMM

This system is build based on [3] and specifications are:

- Train a UBM on all languages data with 128 Gaussian.
- Tokenize the same data with UBM.
- Using the labeled data train a HMM.
- Treat the HMM as an *event recognizer* and proceed with VSM back-end.

6. Tasks

The task is defined to be language detection for 30s-30s train-test in closed-set detection. Performance measure is Equal Error Rate (EER) and cost function C_{avg} defined by NIST [4].

6.1. Albayzin 2010

- Six languages: Spanish, Catalan, Basque, Galician, Portuguese and English.
- Speech data are extracted from multi-speaker TV broadcast recordings.
- Almost 10 hours of data per language is available for training. There are also some extra noisy data for training systems to deal with noisy situation.
- 836 test samples of 30s length for development set.
- 4992 test samples of different lengths (3, 10 and 30s) for evaluation. We should report our results without considering the length (or the state of being clean or noisy) of the utterance.

7. References

[1] Matejka Pavel, Burget Lukas, Schwarz Petr, Cernocky Jan: Brno University of Technology System for NIST 2005 Language Recognition Evaluation, In: Proceedings of Odyssey 2006: The Speaker and Language Recognition Workshop, San Juan, PR, 2006, p. 57-64.

[2] Campbell, W., Campbell, J., Reynolds, D. A., Singer, E., Torres-Carrasquillo, P., Support Vector Machines for Speaker and Language Recognition, Computer Speech and Language, Vol. 20, No. 23, pp. 210229, April 2006.

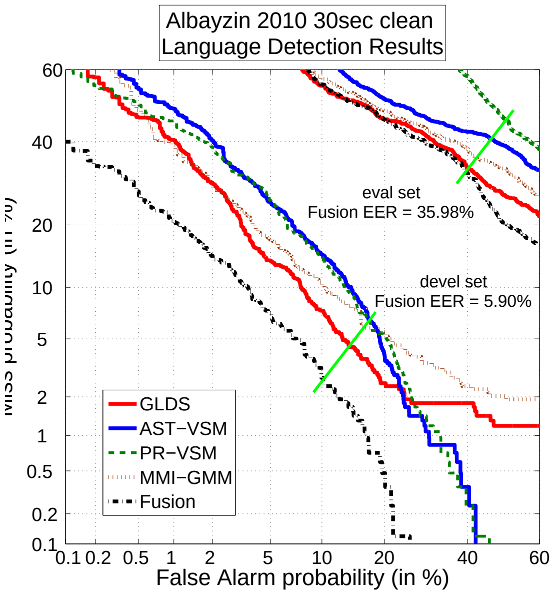


Figure 1: Submitted system result on 30sec clean data.

[3] Haizhou Li; Bin Ma; Chin-Hui Lee; , ”A Vector Space Modeling Approach to Spoken Language Identification,” Audio, Speech, and Language Processing, IEEE Transactions on , vol.15, no.1, pp.271-284, Jan. 2007.

[4] P. Schwarz, ”Phoneme Recognition based on Long Temporal Context, PhD Thesis”, Brno University of Technology, 2009.

[4] www.itl.nist.gov/iad/mig//tests/lre

ViVoLab UZ Language Recognition System for Albayzin 2010 LRE

David Martínez, Jesús Villalba, Antonio Miguel, Alfonso Ortega and Eduardo Lleida

Voice Input Voice Output Laboratory (ViVoLab)
Aragon Institute for Engineering Research (I3A)
University of Zaragoza (UZ), Spain
(david | villalba | amiguel | ortega | lleida)@unizar.es

Abstract

This paper describes the 2 systems submitted by ViVoLab UZ for the Albayzin 2010 Language Recognition Evaluation (LRE) [1]. Both submissions are a fusion of 5 phonotactic and 3 acoustic subsystems. The only difference between them is the normalization and fusion of the scores. We have investigated the state-of-the art methods for Language Recognition (LR) in the KALAKA-2 database [2]. Our group obtained the first position in the evaluation.

Index Terms: language recognition, phonotactic LRE, acoustic LRE, channel compensation, discriminative training

1. Introduction

Language Identification (LID) has experimented a huge development in the last years. To compare the quality of the different LR systems around the world, NIST has coordinated several evaluations (1996, 2003, 2005, 2007 and 2009) [3]. In year 2008, the Spanish Network on Speech Technology coordinated a similar one, for research groups of the Iberian Peninsula [4]. Albayzin LRE 2010 is the second edition. The main difference with NIST is that the languages to be recognized are Spanish, Catalan, English, Basque, Galician and Portuguese, and they are extracted from multi-speaker TV broadcast recordings.

The systems submitted by ViVoLab UZ are a fusion of 5 phonotactic subsystems and 3 acoustic subsystems. Both systems are identical except for the normalization and fusion methods used at the back-end. In the first submission, we make a t-norm of scores and perform a discriminative fusion. In the second, we make a zt-norm of scores, and follow a generative gaussian backend by a discriminative calibration. Our submission includes closed- and open-set condition for the clean speech task, for the 3s, 10s and 30s tests.

The thresholds for each submission are set separately for each condition and for each duration of file, detecting the length of each one by counting the number of frames. For the closed- and open-set conditions, we have used the same systems, but setting the threshold to different values in order to minimize C_{avg}.

The rest of the paper is organized as follows: Section 2 specifies the data used for training; Section 3 describes the acoustic, phonotactic and fusion methods; Section 4 indicates the computational cost of the systems; in Section 5, results obtained in the evaluation are analysed; and Section 6 gives the conclusions and comments some next steps.

2. Data and Performance Measurement

The data used for training our system come from the training part of KALAKA-2 database, with the exception of the training of the phone recognizers. We have used phoneme recognizers trained in Czech, with the Czech SpeechDat-E database [18], in Hungarian, with the Hungarian SpeechDat-E database [19], in Russian, with the Russian SpeechDat-E database [20], in English, with the TIMIT database [21], and in Spanish, with the Albayzin [22] and Speech Dat Car [23] databases.

Calibration of the results was performed with the development data of KALAKA-2 database.

3. System Description

The 2 submitted systems are a fusion of 8 subsystems: 3 acoustic and 5 phonotactics

3.1. Features

The features used for the acoustic systems are MFCC concatenated to their Shifted Delta Cepstra Coefficients (SDC) [5]. We obtain 6 MFCCs plus energy, perform cepstral mean normalization, and then we calculate the SDC with a 7-1-3-7 configuration. After that, we transform the features with a Short Time Gaussianization (STG) [6].

3.2. Acoustic Systems

The 3 acoustic subsystems are a GMM Maximum Likelihood (ML) subsystem, a GMM Maximum Mutual Information (MMI) subsystem and a GMM Factor Analysis (FA) subsystem which performs channel compensation.

3.2.1. GMM ML subsystem

The ML GMM subsystem is based on a calculation of one ML GMM model for each language using the EM algorithm. We perform 10 iterations to obtain a 2048 gaussians model. This method tries to maximize the likelihood of the data for each class.

3.2.2. GMM MMI subsystem

Starting from the GMM ML model, we perform a discriminative re-training based on MMI to obtain the final models. 10 iterations of MMI re-training are run. Unlike the ML training, this method tries to maximize the posterior probability of recognizing all training utterances given the labelled data. The objective function is [7]

$$F_{\mathrm{MMI}}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(\chi_r|s_\mathrm{r})^{K_\mathrm{r}} P(s_\mathrm{r})^{K_\mathrm{r}}}{\sum_{\forall s} p_\lambda(\chi_r|s)^{K_r} P(s)^{K_\mathrm{r}}} \quad (1)$$

where $p_\lambda(\chi_r|s_\mathrm{r})$ is the likelihood of r-th training segment, χ_r, given the correct language identity of the segment, s_r, and model parameters λ. R is the number of training segments, and the denominator represents the overall probability density, $p_\lambda(\chi_r)$. We consider the prior probabilities of all classes equal and drop the prior terms $P(s_\mathrm{r})$ and $P(s)$. Usually, segment likelihood $p_\lambda(\chi_r|s)$ is computed as simple multiplication of frame likelihoods incorrectly assuming statistical independence of feature vectors. Factor $0 < K_r < 1$, which is increasing the confusion between hypothesis represented by numerator and denominator, can be considered as a compensation for underestimating segment likelihoods caused by this incorrect assumption.

3.2.3. GMM FA subsystem

This system is based in a FA for the mean of the models based on [8], in which we have defined two factors, one for the language and one for the channel. Thus, we can obtain a channel compensated model for each language. This is a two-level hierarchy model, since we assume a different GMM that generates every speech segment, and we also assume that for every speech segment, this GMM has been generated by a sub-model. Then, for the speech segment *s*, we have

$$M_s = t_{l(s)} + Ux_s \quad (2)$$

where $t_{l(s)}$ are the *language location vectors*, x_s is a vector of C segment-dependent *channel factors*, and U is a 56-by-C *factor loading matrix*, which translates the channel factors from their low dimensional space to the high dimensional space where the model M_s lies.

The $t_{l(s)}$ matrix is obtained by MAP adaptation from a UBM model with mean m_0 and covariance matrix Σ, in the following way

$$t_{l(s)k} = \frac{\sum_s f_{sk}}{\tau + \sum_s n_{sk}} \quad (3)$$

being n_{sk} and f_{sk} the zero and first order statistics respectively, for the *k*th gaussian component.

The U matrix and the channel factors are calculated according to a ML criterion, using the EM algorithm iteratively, in a similar way to [8].

The scoring of each utterance is made via a linear scoring, as proposed in [9].

3.3. Phonotactic Systems

5 PRLM sub-systems [12] in different languages have been fused. 4 of them use the Brno University of Technology (BUT) phoneme recognizer, based on ANN/HMM and Temporal Patterns (TRAPS) with Split Temporal Context (STC) [10], and are trained on Czech, Hungary, Russian and English. The other one uses the phoneme recognizer of the UZ, which is based on GMM/HMM with conventional MFCC and is trained on Spanish. In this one, the phonemes are taken with right and left context, so we will call the recognition unit subphoneme instead. However, we will keep only the central unit for the posterior step, that is, the phoneme without context. The output phonemes are used to train a langauge model (LM) for each one of the target languages with the SRILM tool [11].

All LMs are built with an interpolated Witten-Bell discounting method. We use 4-grams for building them in all cases, and for testing we also use 4-grams for all cases except for the Spanish LM, in which we test with 3-grams. The reason is that we empirically experimented a better performance with this configuration. In addition, for the four phoneme recognizers based on GMM/ANN, we make use of lattices [13] to get more information out of the acoustic signal. Specifically, we create a 100-best list for the train and a 5-best list for the test.

3.4. Fusion for the Primary Submission

In our primary submission, the results coming from each system are T-Normalized [15], fused, and for the closed-set condition, another T-Normalization is applied after the fusion. The fusion is also a calibration [14] and the fused log-likelihood vector is

$$\mathbf{l}'(x_t) = \sum_{k=1}^{K} \alpha_k \mathbf{l}_k(x_t) + \beta \quad (4)$$

where the coefficients α_k and β are calculated via a discriminative Linear Logistic Regression (LLR) training, using the FoCal Multi-class toolkit [14], and $\mathbf{l}_k(x_t)$ is the output of system k when input in time t is x_t.

3.5. Fusion for the Alternative Submission

In this submission, we investigate the ZT-Normalization [15] technique, combined with a Gaussian Back-End (GBE) followed by a discriminative LLR fusion [14], as the one in the primary submission. In the closed-set condition, results after the GBE and after the LLR fusion are again T-Normalized.

In a GBE, the likelihood scores are obtained from multivariate Gaussians, with target language specific means and one common full covariance matrix. As explained in [16], the GBE can be seen as an affine transformation. The linear part of the transform is the same as a Linear Discriminant Analysis (LDA) transform, which tries to maximize the ratio of between-class to within-class variance. The translation part of the affine transform is equivalent to the calibration task of setting language dependent thresholds. The decision made by the GBE corresponds to the following normal distribution function:

$$\delta_l(\mathbf{x}) = (\mathbf{x} - \mu_l)^t \Sigma^{-1} (\mathbf{x} - \mu_l) \quad (5)$$

where μ_l is the mean for language l, Σ is the common covariance matrix, $\mathbf{x}$ is the input score and $\delta_l(\mathbf{x})$ is the transformed output.

The posterior discriminative LLR was added to give further calibration to the system, and we could check a further improvement in the results.

4. Computational Cost

Real time factor was approximately 0.9xRT for both submissions, since the normalization and calibration technique does not alter the overall processing time.

5. Analysis of Results

In this section we will show the results obtained in the evaluation for the described systems in terms of C_{avg}. We can see how accurate is the calibration on the DET curves comparing C_{avg} (marked with 'x') with C^*_{avg} (marked with 'o'). To analyse results, we will focus on the the primary system, since the results of the alternative system follow the same trend but with higher error rates.

5.1. Primary System - Closed Clean (CC)

In Fig. 1 we have the results of our primary system for the clean speech, closed set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0184, 0.0418 and 0.0943, respectively. The 30s test of this condition is the one used to rank systems in the evaluation. Our system was the best.

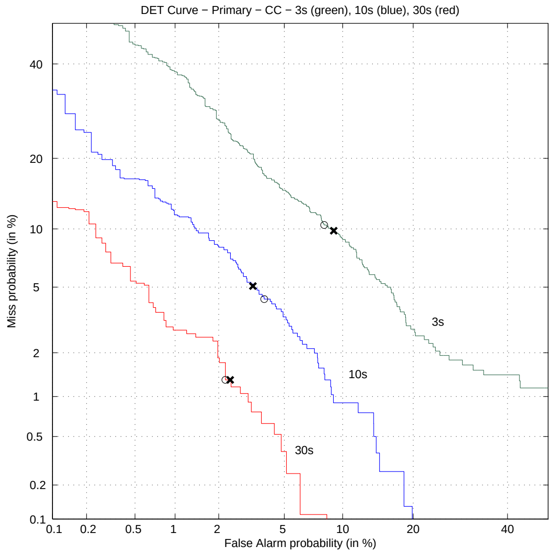


Figure 1: *DET Curves for Primary System CC condition*

$P_{fa}(L_t, L_n)$	Target Language L_t					
L_n	SPA	CAT	ENG	BAS	GAL	POR
SPA	–	0.016	0.000	0.000	0.232	0.000
CAT	0.020	–	0.000	0.000	0.007	0.007
ENG	0.000	0.000	–	0.000	0.000	0.000
BAS	0.000	0.000	0.000	–	0.000	0.000
GAL	0.413	0.016	0.000	0.000	–	0.000
POR	0.000	0.000	0.000	0.000	0.000	–
$P_{miss}(L_t)$	0.008	0.013	0.000	0.008	0.050	0.000
Avg. $P_{fa}(L_t)$	0.087	0.006	0.000	0.000	0.048	0.001
Avg. P_{miss} = 0.0131						
Avg. P_{fa} = 0.0237						

Table 1: *Error Rates for CC 30s condition in the primary system. We show the target language L_t in the columns and the segment language L_n in the rows. Labels of languages are SPA=Spanish, CAT=Catalan, ENG=English, BAS=Basque, GAL=Galician and POR=Portuguese*

If we analyse Table 1, we can see a very good performance recognizing all languages for the 30s CC condition, having a global P_{miss} of 0.0131. The highest figure is for Galician with a P_{miss} of 0.050. However, if we look at the false-alarm probabilities, we can check in general a good performance, but a low one when discriminating between Spanish and Galician. The false-alarm probability of saying that the language transmitted is Galician when it is really Spanish is 0.232 and of saying that it is Spanish when it is really Galician is 0.413. Several reasons could be considered for this behaviour, but we think that the most dramatic is, after listening some of the recordings, the fact that many Galician speakers are Spanish-native speakers. Therefore, their Galician accents are very influenced by the Spanish language.

The problem caused by people who speak several languages is a general one for language recognizers, and it would be beneficial to have into account this information when training LID systems. One solution could be to train systems for native and non-native speakers as different languages. Another approach to this problem could be to apply discriminative training techniques which place more gaussians (in GMM systems) at the borders between these languages for a better characterization of these areas of the vector space.

For the rest of languages, we can see small confusion rates, especially for English and Basque, which are 0. This is due to the highly different acoustic and phonotactic nature of these languages with regard to the others.

We detail the results for each individual subsystem for the 30 s condition in Table 2. We can see that the subsystem that performs the best is the FA. On the other hand, the PRLM_ES and PRLM_EN do not give good results by themselves. After evaluation, we could check that the back-end was not optimum, and with only a GBE we obtained great improvements in all subsystems.

Subsystem	C_{avg}
JFA	0.0357
ML	0.0855
MMI	0.0598
PRLM_CZ	0.0569
PRLM_HU	0.0501
PRLM_RU	0.0547
PRLM_EN	0.2618
PRLM_ES	0.1474

Table 2: *C_{avg} for the individual subsystems of the primary submission for the CC 30s condition*

5.2. Primary System - Open Clean (OC)

In Fig. 2 we have the results of our primary system for the clean speech, open set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0307, 0.0644 and 0.1202, respectively.

In Table 3 we can check that the performance of the system for the 30s OC condition has slightly dropped, compared to the CC condition. This is due to the introduction of out-of-set (OOS) languages. The average P_{miss} drops to 0.0193, but more dramatic is the decrease in the average P_{fa}, which drops to 0.0422, i.e. a relative decrease of 79% compared to the 30s CC condition.

We still observe the problem when distinguishing Spanish of Galician. And now, we can see that the OOS languages are mainly confused with English, with a $P_{fa}(EN, OOS)$ of 0.188.

The results for the individual subsystems for the 30 s condition are in Table 4. Again, the system that performs the best is the FA. In this case, the difference with the rest is much higher. However, we have to take into account that our system was tuned for the fusion of all subsystems and not for each individual one. If we tuned it for each one, we would obtain better results. As in the CC case, we could check after evaluation that the back-end was not optimum.

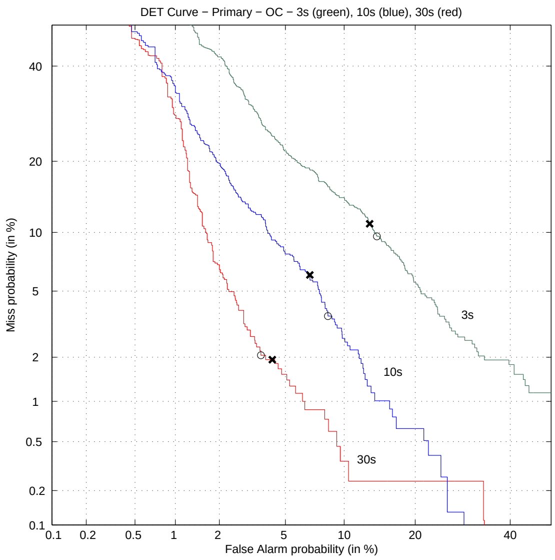


Figure 2: *DET Curves for Primary System OC condition*

Subsystem	C_{avg}
JFA	0.0510
ML	0.3658
MMI	0.3262
PRLM_CZ	0.4499
PRLM_HU	0.3998
PRLM_RU	0.4293
PRLM_EN	0.4991
PRLM_ES	0.4920

Table 4: C_{avg} *for the individual subsystems of the primary submission for the OC 30s condition*

$P_{fa}(L_t, L_n)$	Target Language L_t					
L_n	SPA	CAT	ENG	BAS	GAL	POR
SPA	–	0.008	0.000	0.000	0.208	0.000
CAT	0.0134	–	0.000	0.000	0.000	0.000
ENG	0.000	0.000	–	0.000	0.000	0.000
BAS	0.000	0.000	0.000	–	0.000	0.000
GAL	0.446	0.008	0.000	0.000	–	0.000
POR	0.000	0.000	0.000	0.000	0.000	–
OOS	0.026	0.064	0.188	0.052	0.003	0.094
$P_{miss}(L_t)$	0.024	0.013	0.007	0.008	0.050	0.013
Avg.$P_{fa}(L_t)$	0.092	0.003	0.000	0.000	0.042	0.000
Avg.$P_{fa}(L_t + L_o)$	0.066	0.028	0.075	0.021	0.026	0.038
Avg. P_{miss} = 0.0193						
Avg. P_{fa} = 0.0422						

Table 3: *Error Rates for OC 30s condition in the primary system. We show the target language L_t in the columns and the segment language L_n in the rows. Labels of languages are SPA=Spanish, CAT=Catalan, ENG=English, BAS=Basque, GAL=Galician, POR=Portuguese and OOS=Out-Of-Set*

5.3. Alternative System - CC

In Fig. 3 we have the results of our alternative system, for the clean speech, closed set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0238, 0.0498 and 0.1087, respectively. The detail analysis of results is similar to the primary system CC condition, but the error rates are higher.

5.4. Alternative System - OC

In Fig. 4 we have the results of our alternative system, for the clean speech, open set condition, for 30, 10 and 3 s of duration of utterance. C_{avg} is 0.0373, 0.0635 and 0.1309, respectively. The detail analysis of results is similar to the primary system OC condition, but the error rates are higher.

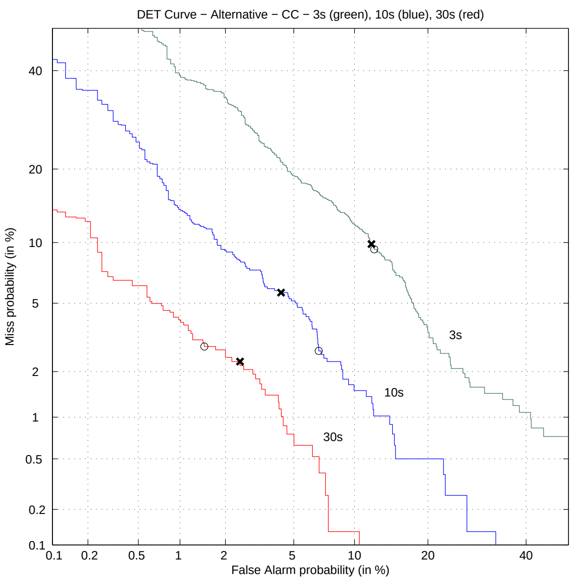


Figure 3: *DET Curves for Alternative System CC condition*

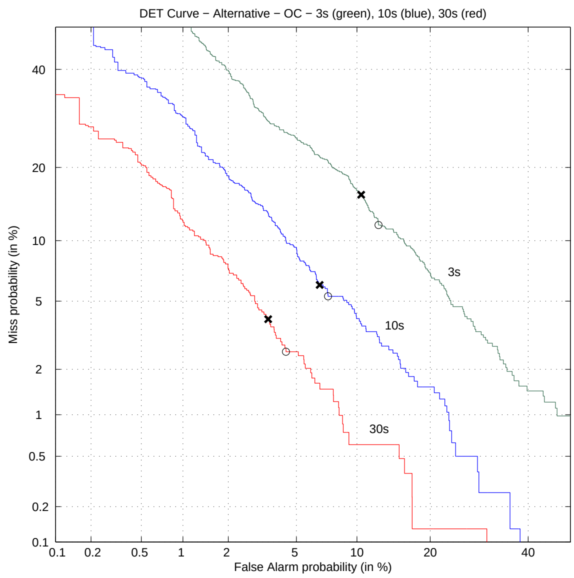


Figure 4: *DET Curves for Alternative System OC condition*

6. Conclusions and Future Work

In this edition of Albayzin LRE, ViVoLab UZ participates for the first time in a Language Recognition Evaluation. We have built several state-of-the-art systems that have been tested in the KALAKA-2 database. All these systems are fusioned into one, in which the characteristics of each are combined to get a higher performance. For the ranking of systems only the primary submission of the CC 30s condition was considered and our system was the best of all participant sites.

First, we built 5 phonotactic systems trained in different languages using 2 phoneme recognizers, one of BUT and one of UZ. Then, we combined them with 3 acoustic systems, ML, MMI and FA, coming from the Speaker Identification (SPKID) ideas, where our group has big experience.

Finally, we normalized and fused the results in two different ways, what generates our two submissions. One is based on a T-Norm and a discriminative LLR fusion, while the other is based on a ZT-Norm and a GBE followed by a discriminative LLR fusion.

Analysing the results, and focusing on the 30s CC condition, we can see very low P_{miss} values and, in general, low P_{fa} values. However, we detect a great confusion between Spanish and Galician, mainly caused by the fact that many Galician speaker are non-native speakers and their accent is influenced by the Spanish language. To solve this, we should think of having into account if the speaker is native or non-native when training the systems, and of turning our efforts to build a discriminative algorithm able to differentiate properly the borders between these languages.

As next steps we consider including several other approaches to our final system. First, the introduction of Vocal Tract Length Normalization (VTLN) should make the features more independent of the speaker. Secondly, we will experiment with a Probabilistic Linear Discriminant Analysis (PLDA) system on LID, since the performance on SPKID has shown to be excellent [17]. And finally, we would like to continue investigating new normalizing techniques, as one only based on the length of the files under test. In addition, we checked in a post-evaluation of the systems, that the selected configurations for the back-end were not optimal, and that only a GBE could have given a better performance.

7. Acknowledgements

We would like to thank the organizers of Albayzin 2010 LRE for his effort in preparing all the data of KALAKA-2 database, and also to all the organization of Fala 2010 for supporting the Albayzin 2010 LRE.

This work was funded by the Spanish Ministry of Science and Innovation under project TIN2008-06856-C05-04.

8. References

[1] The Albayzin 2010 Language Recognition Evaluation Plan (Albayzin 2010 LRE).

[2] KALAKA-2. Speech database created for the Albayzin 2010 Language Recognition Evaluation, organized by the Spanish Network on Speech Technology. Produced by the Software Technologies Working Group (GTTS, http://gtts.ehu.es), University of the Basque Country.

[3] http://www.itl.nist.gov/iad/mig//tests/lre/

[4] Luis Javier Rodríguez-Fuentes, Mikel Penagarikano, Germán Bordel, and Amparo Vasrona, "The Albayzin 2008 Language Recognition Evaluation", in Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 28 June - 1 July 2010.

[5] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstal Features", in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2002, pp.89-92.

[6] J. Pelecanos and S. Sridharan, "Feature Waroing for Robust Speaker Verification", Proc. Speaker Odyssey 2001 conference, June 2001.

[7] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition", Ph.D. thesis, Cambridge University, July 2004.

[8] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms", Technical Report CRIM-06/08-13, CRIM, 2005, http://www.crim.ca/perso/patrick.kenny/FAtheory.pdf.

[9] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny, "Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis", in Proc. ICASSP, pp.4057-4060, Taipei, Apr. 2009.

[10] P. Schwarz, "Phoneme Recognition Based on Long Temporal Context", Ph.D. Thesis, Brno University of Technology, 2009. http://speech.fit.vutbr.cz/cs/software/phoneme-recognizer-based-long-temporal-context.

[11] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit", in Proc. ICSLP, pp. 901-904, 2002. http://www.speech.sri.com/projects/srilm.

[12] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telepone Speech", IEEE Trans. Acoust., SPeech Signal Processing, vol. 4, no. 1, pp. 31-44, 1996.

[13] J.L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition using Phoneme Lattices", in Proc. International Conferences on Spoken Language Processing (ICSLP), Sept. 2004, pp. 1283-1286.

[14] N. Brümmer, "FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual-. http://sites.google.com/site/nikobrummer/focalmulticlass.

[15] R. Auckenhalter, M. Carey, and H. Lloyd-Thomas, "Dcore Normalization for Text-Independent Speaker Verification Systems, Digital Signal Processing, 10(1), 42-54.

[16] D.A. van Leeuwen and N. Brümmer, "Channel-Dependent GMM and Multi-Class Logistic Regression Models for Language Recognition", 2006 IEEE Odyssey: The Speaker and Language Recognition Workshop.

[17] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors", in Proceedings of Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, 28 June - 1 July 2010.

[18] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/czech.html

[19] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/hungarian.html

[20] http://www.fee.vutbr.cz/SPEECHDAT-E/sample/russian.html

[21] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[22] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.-B. M. no, and C. Nadeu, Albayzin speech database: Design of the phonetic corpus, in Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech-Interspeech), Berlin, Germany, September 1993, pp. 175178.

[23] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, Speech Dat Car. A large speech database for automotive environments, in Proceedings of the II Language Resources European Conference, Athens, Greece, June 2000.

The L^{2}F Language Verification Systems for Albayzin-2010 Evaluation

Alberto Abad[1], *Oscar Koller*[1], *Isabel Trancoso*[1,2]

[1]L^{2}F - Spoken Language Systems Lab, INESC-ID Lisboa
[2]Instituto Superior Técnico, Lisboa, Portugal
alberto.abad@l2f.inesc-id.pt

Abstract

This paper presents a description of INESC-ID's Spoken Language Systems Laboratory (L^{2}F) Language Verification systems submitted to the ALBAYZIN-2010 evaluation. The primary submission consists of the fusion of six individual sub-systems: one Gaussian supervector approach with support vector machines that relies on the acoustic characteristics extracted by a front-end of shifted deltas, and five individual Phone Recognition and Language Modeling detectors based on five different phone tokenizers. Additionally, two contrastive systems have been developed. Language detection results have been submitted for all the evaluation conditions for every system. The main particularity of the systems developed for this evaluation is that individual language models for clean and noisy conditions have been trained for each target language. Results for the different systems and evaluation conditions are reported.

1. Introduction

The "Red Temática en Tecnologías del Habla" (RTTH) has organized in the recent years a series of evaluations - so called ALBAYZIN evaluations - in some relevant speech processing topics devoted to encourage language research activities on the four official languages of Spain: Castilian, Catalan, Basque and Galician.

Similar to the well-known NIST Language Recognition Evaluation series, a Language Verification (LV) task was proposed in ALBAYZIN-08 with the objective of determining if each one of the four official languages of Spain was spoken (or not) in a given test file. In the new ALBAYZIN-2010 campaign the set of target languages is increased to cover also Portuguese and English.

This paper presents the LV systems developed by INESC-ID's Spoken Language Systems Laboratory (L^{2}F) for the ALBAYZIN-2010 campaign. A primary and two contrastive systems have been submitted, which differ in the number of employed sub-systems and in the followed back-end strategy for calibration. The *primary* system consists of the fusion of six different language detection sub-systems: an acoustic system based on Gaussian mixture model SuperVectors (GSV) [1] and five phonotactic Phone Recognition and Language Modeling (PRLM) [2] systems. Additionally, the *alt1* system explores a method for introducing segment duration score normalization to the calibration stage, while the *alt2* system is aimed at developing a simplified LV system. The next Section 2 presents a brief description of the task, the data provided for the evaluation and the evaluation metrics. Section 3 describes some commonalities of the systems developed (see Section 3.1) and details of each one of the six individual sub-systems: the GSV-LV system and the PRLM-LV detectors are described in Sections 3.2 and 3.3, respectively. Measurements of the computational deployment in the processing of the evaluation data set are also provided. The three submitted systems are described in Section 4. In Section 5 results obtained by the three systems in the different evaluation conditions with the development data set are presented. Finally, Section 7 presents our main conclusions.

2. ALBAYZIN-2010 LV: Task, Data and Metric Description

Detailed information on the ALBAYZIN-2010 LV campaign can be found in the evaluation plan document [3].

2.1. Task and Evaluation Conditions

The task consists of deciding whether a speech segment belongs to each one of the six target languages (Castilian, Catalan, Basque, Galician, Portuguese and English) or not. For each test signal six decision results (true or false) are produced together with a score, one for each of the target languages.

Four test evaluation conditions are proposed depending on the type of verification test (closed-set vs. open-set) and the type of speech (clean vs. noisy). In contrast to the closed mode, in the open mode speech segments from unknown languages different from the target ones may appear in the test data and are taken into account for the systems' assessment. The four evaluation conditions are referred to as closed-clean (CC), closed-noisy (CN), open-clean (OC) and open-noisy (ON).

2.2. Train, Development and Test Data

All the data provided for the ALBAYZIN-2010 evaluation are TV programs captured at 16 kHz. The training data set consists of more than 12 hours per target language, in several files of variable length separated in clean speech (more than 10 hours) and noisy speech (around 2 hours). The evaluation data set consists of 4992 files with speech of the six target languages and in other unknown languages of 3 different nominal durations: 3, 10 and 30 seconds. Additionally, a development data set consisting of 4950 files of similar characteristics to the evaluation set was provided with language identification, duration and type of speech labels.

2.3. Performance Metric

An average performance score based on the false positive and false alarm rates obtained by the evaluating systems is used. The performance score, hereinafter referred to as C_{avg}, is computed independently for each test length duration (3, 10 and 30 seconds). Further details about the metrics can be found in [3].

3. Language Verification Sub-system Description

3.1. Common Characteristics

3.1.1. Audio Data Pre-processing

The training data provided for each target language and for each speech type was pre-processed in order to segment long data files into a set of homogeneous reduced length speech segments. First, speech-non-speech (SNS) segmentation was applied [4]. The SNS module is a finite state machine that uses a binary Multi Layer Perceptron (MLP) trained with several hours of BN data to identify audio portions that do not contain speech, speech with too much noise or pure music. After this segmentation process, continuous speech segments (1 second of non-speech tolerance) of length above 8 seconds and below 40 seconds were selected. In the particular case of Castilian, these thresholds for segment filtering were fixed to 7 and 49 seconds. Notice, that this pre-processing segmentation was only applied to the training data and not to the development and evaluation data sets. Table 1 shows the amount of selected segments and the total duration in minutes for each target language and type of speech.

	clean		noisy	
	#segm	dur. [min]	#segm	dur. [min]
castilian	576	227.9	223	81.7
catalan	674	237.8	235	76.6
english	600	231.3	266	92.7
basque	722	260.7	268	80.8
galician	746	258.5	254	74.6
portuguese	583	222.3	233	83.3

Table 1: *Training data segmentation for each target language and speech type.*

3.1.2. Target Language Modeling

One of the main particularities shared among all the developed sub-systems is that a separate target language model was trained for clean and noisy speech. The two target models of each language are used to obtain two language-dependent scores for each speech test segment. Consequently, for every test segment a vector of 12 scores $\mathbf{x}_i$ is produced by every individual sub-system i.

3.1.3. Linear Gaussian Back-End

A linear Gaussian Back-End (GBE) follows every single sub-system to transform the 12 elements score-vector $\mathbf{x}_i$ to a 7 elements log-likelihood vector $\mathbf{s}_i$ (6 target languages plus 1 out-of-set language log-likelihoods):

$$\mathbf{s}_i = \mathbf{A}_i \mathbf{x}_i + \mathbf{o}_i \tag{1}$$

where $\mathbf{A}_i$ is the transformation matrix for system i and $\mathbf{o}_i$ is the offset vector.

A common characteristic of all the systems developed is that open-set and closed-set conditions have not been distinguished in back-end calibration (nor have they in the later fusion of the individual sub-systems). In other words, the same 7 log-likelihoods are produced independently of the type of verification test (closed-set or open-set) and they are used to obtain detection log-likelihood ratios and decisions using the adequate prior distributions over language classes in each verification test type.

3.2. GSV-LV sub-system

A method generally known as GSV [1] is known to be a successful approach for both speaker and language verification tasks. GSV-based approaches map each speech utterance to a high-dimensional vector space. Support Vector Machines (SVMs) are used for classification of test vectors within this space. The mapping to the high-dimensional space is achieved by stacking all parameters (usually the means) of an adapted GMM in a single supervector by means of a Bayesian adaptation of a universal background model (GMM-UBM) to the characteristics of a given speech segment. In language recognition, a binary SVM classifier is trained for each target language with supervectors of the target language as positive examples and supervectors of other non-target languages as negative examples. During test, the supervector of the testing speech utterance is used by the binary classifier to generate a score for each target language.

3.2.1. Feature Extraction

The extracted features are shifted delta cepstra (SDC) [5] of Perceptual Linear Prediction features with log-RelAtive SpecTrAl speech processing (PLP-RASTA). First, 7 PLP-RASTA static features are obtained and mean and variance normalization is applied in a per segment basis. Then, SDC features (with a 7-1-3-7 configuration) are computed, resulting in a feature vector of 56 components. Finally, low-energy frames detected with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment are removed.

3.2.2. Supervector Extraction and SVM Language Modeling

A GMM-UBM of 256 mixtures was trained with approximately 9 hours of speech randomly selected among the clean segments (around 1.5 hours per target language) of the training data set of Table 1.

One single iteration of Maximum a Posteriori (MAP) adaptation with relevance factor 16 is performed for each speech segment to obtain the high-dimensional vector of size 56x256.

The linear SVM kernel of [1] based on the Kullback-Leibler (KL) divergence is used to train the target language models with the LibLinear implementation of the libSVM tool [6]. For each target language and type of speech, all the training segments of that language are used as positive examples and all the segments from the other languages are used as negative background set.

3.2.3. Processing Time

Processing time measurements of the developed language recognition systems were carried out in a machine with two Quad Xeon 2.4GHz (E5530) processors with 48 GBytes of DDR3 RAM at 1333 MHz. Notice, however, that data is stored in a distributed file system with relatively slow transfer rates. Thus, disk access can become a bottleneck in some fast operations. The processing time was measured in a sub-set of 100 test files amounting to 1522.8 seconds. The feature extraction of the 100 files consumed 405 seconds, Bayesian adaptation and supervector extraction lasted 118 seconds, and scoring was performed in 6 seconds. These figures correspond to 0.35xRT approximately.

3.3. PRLM-LV Sub-systems

The Phone Recognition followed by Language Modeling (PRLM) systems used for ALBAYZIN-2010 exploit the phonotactic information extracted by five individual tokenizers: European Portuguese, Brazilian Portuguese, European Spanish (Castilian), American English and a mixed African/European Portuguese tokenizer using special mono-phonetic units [7]. The key aspect of this type of system is the need for robust phonetic classifiers that generally need to be trained with word-level or phonetic level transcriptions. In this case, the tokenizers are MultiLayer Perceptrons (MLP) trained to estimate the posterior probabilities of the different phonemes for a given input speech frame (and its context). For each target language and for each tokenizer a different phonotactic *n-gram* language model is trained. During test, the phonetic sequence of a given speech signal is extracted with the phonetic classifiers and the likelihood of each target language model is evaluated.

3.3.1. Phonetic Tokenizers

The tokenization of the speech data is done with the neural networks that are part of our hybrid Automatic Speech Recognition (ASR) system named AUDIMUS [8]. The tokenizers combine three MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), PLP with log-RelAtive SpecTrAl speech processing features (PLP-RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). A phone-loop grammar with phoneme minimum duration of two frames is used for phonetic decoding.

The networks were trained with different amounts of broadcast news (BN) annotated data. For the European Portuguese classifier, 57 hours of manually annotated data and more than 300 hours of automatically transcribed BN data were used. The Brazilian Portuguese classifier was trained with around 13 hours of BN data. The Spanish system used 14 hours of manually annotated data and 78 hours of automatically transcribed data. The English system was developed with the HUB-4 96 and HUB-4 97 data sets, that contain around 142 hours of TV and Radio Broadcast data. Finally, the mixed African/European system was trained with two times 6 hours of manually annotated broadcast news data containing both Portuguese varieties equally balanced. Through a particular training process (refer to [7]), this system is tuned to differentiate between the close Portuguese varieties.

The size of the input and hidden layers of the neural networks varies among the different parameterizations and languages, but in general all the MLPs are composed by two hidden-layers with a relatively small number of hidden units in order to accelerate the tokenization process. In the case of the output layer, its size corresponds to the number of phonetic units of each language, plus silence (no additional sub-phonetic or context-dependent units have been considered [9]).

3.3.2. Phonotactics Modeling

For every phonetic tokenizer, the phonotactics of each target language for every type of speech condition (clean and noisy) is modeled with a 3-gram back-off model, that is smoothened using Witten-Bell discounting. For that purpose the SRILM toolkit has been used [10].

3.3.3. Processing Time

Using the previously described machine, the total time deployed in processing the 100 files sub-set when running the 5 PRLM systems in parallel is 245 seconds, which corresponds to 0.16xRT. When the PRLM systems are run one after the other, the total amount of processing time increases up to 936 seconds. The phonetic tokenization operations account for 60% to 80% of the processing time (depending on the network) and the rest of the time is consumed in the scores generation.

4. The L^{2}F Submissions

The L^{2}F submitted systems consist of the fusion of some of the sub-systems described in previous Section 3. Linear logistic regression (LLR) has been used to fuse the log-likelihood outputs generated by the linear GBEs of the individual sub-systems to produce fused likelihoods **l**:

$$\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b} \tag{2}$$

where α_i is the weight for sub-system i and **b** is the language-dependent shift.

The GBEs and the LLR fusion have been trained and tested with the development data set using a jack-knifing strategy: data is partitioned in 5 random sets and each one of the sets is once held out for testing and the other 4 sets are used to estimate the calibration and fusion parameters. The initial randomization of the data is iterated 5 times and a jack-knife scheme is repeated resulting in total in 25 sets of estimated back-end and fusion parameters that are averaged to obtain the final back-end. Calibration was carried out using the FoCal Multiclass Toolkit [11].

4.1. Primary System (primary)

The *primary* system consists of the fusion of the GSV sub-system and the five PRLM sub-systems. Segment length duration dependent and type of speech dependent back-ends were trained. That is, for each combination of type of speech (clean and noisy) and segment duration (30, 10 and 3 seconds) different GBE and LLR parameters are estimated using only the development data of the corresponding type and duration. In test –the evaluation data set was split in 30, 10 and 3 seconds– the back-ends trained with only clean speech are used for the CC and OC evaluation conditions. On the other hand, CN and ON language detection is performed using the back-ends trained with noisy speech. It is worth remembering that, as explained in section 3.1.3, the same back-end is used for closed-set and open-set conditions applying different language priors for log-likelihood ratio and decision generation.

4.2. First Contrastive System (alt1)

The objective of the *alt1* system is to investigate an alternative back-end method that incorporates segment duration normalization. Like in the primary system, the *alt1* system consists of the fusion of the six sub-systems described in Section 3 and uses segment length duration dependent and type of speech dependent back-ends. However, in contrast to the *primary* system, a segment length normalization strategy similar to the one described in [12] was considered. In [12] a duration-independent back-end that uses duration-information in the fusion as side-information is proposed.

$$\mathbf{l} = \sum_i \alpha_i \mathbf{s}_i + \mathbf{b} + \mathbf{Cd} \tag{3}$$

where **d** is the vector of durations (that may be different for each sub-system). Additionally, the scores of each individual system are augmented with multiplied versions of the duration

d^p ($\mathbf{x}_i d^p$), where d is the segment duration and p can take different values.

$$\mathbf{s}_{i,p} = \mathbf{A}_{i,p}(\mathbf{x}_i d^p) + \mathbf{o}_{i,p} \tag{4}$$

In the *alt1* system, the segment duration information is ignored as a side-information in the fusion process ($\mathbf{C} = \mathbf{0}$), but it is used to produce duration normalized scores with p values of $0, -1$ and $-1/2$, that correspond, to the original scores, the scores normalized by the duration and by the square root of the duration of each individual sub-system. The duration d is measured as the number of speech frames in the GSV sub-system or the number of decoded phones in the case of PRLM sub-systems. Therefore, instead of 6 GBEs, 18 GBEs need to be estimated and the fused likelihood vector $\mathbf{l}$ is the result of the fusion of the "18 sub-systems".

$$\mathbf{l} = \sum_{i.p} \alpha_{i,p}\mathbf{s}_{i.p} + \mathbf{b} \tag{5}$$

Notice, that in contrast to [12], duration-dependent back-ends are estimated and, thus, we are applying a sort of within-class duration normalization. Consequently, like in the *primary* system, different GBE and LLR parameters for each type of speech and duration condition are estimated.

4.3. Second Contrastive System (alt2)

The aim of the *alt2* system is to assess the performance of a heavily simplified LV system. First, the number of sub-systems considered is reduced and only the GSV sub-system and the PRLM sub-system based on the European Spanish tokenizer – which was consistently the best performing one– are fused. Second, a single back-end is trained for all the conditions using all the development data ignoring the segment duration class and type of speech. A back-end that incorporates segment duration normalization scores like the one described above is used. Notice, that the *alt2* system generates the same scores and decisions for the OC and ON conditions and for the CC an CN conditions. Like in the previous submitted systems, the difference between the closed-set and open-set conditions relies on the target language priors applied.

5. Results on the Development Set

Table 2 presents the results obtained in the development set, for the primary and both contrastive systems. The three top-rows correspond to clean speech development data results and the last three rows to noisy speech. For the sake of clarity, 100xC_{avg} performance scores are reported.

With respect to the different submitted systems, similar detection performances of the *primary* and *alt1* systems can be observed. In fact, the *alt1* system consistently outperforms the *primary* system (except in the 10 seconds CC condition), showing the benefits of the within-class duration normalization method. However, we observed some calibration instabilities during the training of the calibration and fusion parameters (probably due to lack of data) that prevented us from presenting the *alt1* system as our primary submission. On the other hand, the *alt2* system shows a considerable lower performance, as we expected. However, in spite of its simplicity and the use of a general back-end for all conditions, it is still able to provide a quite significant language detection performance, particularly in closed-set and long segment duration conditions.

Regarding the evaluation conditions and segment duration, as it is well-known in LV tasks, the open mode is significantly more challenging than the closed one, and the use of longer segments contributes to smaller detection errors. The performance of all the submitted systems is also considerably affected by the speech quality. Significant performance degradations are obtained in noisy speech conditions, particularly, larger relative cost increase is observed in segments with longer durations.

System	30 sec		10 sec		3 sec	
	cl	op	cl	op	cl	op
primary	0.28	0.45	1.28	2.21	5.35	7.03
alt1	0.23	0.26	1.38	1.99	4.94	6.86
alt2	0.97	1.94	2.18	3.50	7.55	9.75
primary	1.32	1.88	2.02	3.61	6.73	7.90
alt1	0.92	1.16	1.82	2.33	5.08	7.07
alt2	1.90	3.09	4.71	6.87	12.64	15.50

Table 2: *100xC_{avg} performance on the ALBAYZIN-10 LV development set on closed-set and open-set mode and for clean (top three rows) and noisy speech (last three rows).*

6. Results on the Evaluation Set

Table 3 presents the results obtained in the evaluation set. Like in previous development results, 100xC_{avg} performance scores are reported. It must be noticed that an error in the PRLM American English sub-system that affected both the *primary* and *alt1* submissions was detected after submission: detection scores of the evaluation data were erroneously generated using the n-gram language models of the African Portuguese PRLM. In order to draw correct conclusions about the performance of the submitted systems, only the corrected results are provided here.

The most remarkable observation is the quite different language detection performance achieved by the *primary* and *alt1* systems with respect to the *alt2* and the results obtained in the development data set. While *primary* and *alt1* are still quite similar between them, a huge performance loss with respect to the development set is obtained, which is still more noticeable when they are compared to *alt2*. In longer segment duration conditions and particularly in noisy type of speech (30 seconds clean and 30 and 10 seconds noisy), *alt2* clearly outperforms the other submissions.

A post-evaluation analysis is still being conducted, however some preliminary explanations for these unexpected results can be provided. The *alt2* system mainly differs from the other submissions in two aspects: the number of sub-systems and the way the back-end parameters are estimated. With respect to the number of sub-systems, although an increased number of sub-systems does not necessarily imply improved detection, it is quite unlikely that it is the cause for large performance loss, specially when the individual sub-systems have been verified to provide significant language detection ability individually. The most likely reason for the different performance observed in development and evaluation data sets is the poor estimation of the back-end parameters due to the insufficient amount of data available for each evaluation condition, and the large number of back-end parameters. On the one hand, there are 1164 clean and 486 noisy development segments for every segment length duration. On the other hand, the back-end of the *primary* system is composed of around 550 parameters and the *alt1* has about three times this number of parameters. Given the fact that back-end parameters are individually estimated for each segment duration and type of speech, we believe that an overestimation problem to development data occurred. The fact that the most important performance degradations are observed in noisy conditions seems to verify this hypothesis. According to our current post-evaluation calibration experiments, language recognition improvements can be obtained by simply applying the back-end parameter estimation strategy of the *alt2* system to calibrate and fuse the six sub-systems.

System	30 sec		10 sec		3 sec	
	cl	op	cl	op	cl	op
primary	2.23	2.96	3.59	4.68	8.53	10.73
alt1	2.19	3.09	3.63	4.45	8.44	10.29
alt2	1.81	3.41	4.59	6.11	10.55	12.89
primary	4.16	7.00	8.10	9.81	12.73	15.51
alt1	4.03	8.39	7.54	9.48	12.17	16.09
alt2	2.53	4.75	6.36	9.36	13.42	16.54

Table 3: *100xC_{avg} performance on the ALBAYZIN-10 LV evaluation set on closed-set and open-set mode and for clean (top three rows) and noisy speech (last three rows).*

7. Conclusions

In the ALBAYZIN 2010 language recognition evaluation campaign, the L^2F has presented a primary system based on the fusion of 6 individual language recognition systems (one acoustic and five phonotactics) and two additional contrastive systems. The estimated processing time of the primary system is approximately 0.51xRT. The performance achieved by the submitted systems in the different evaluation conditions in the development data set was outstanding. However, a considerable performance loss was verified in the evaluation data and in contrast to our expectations, the simplest submitted system resulted in the most robust language detector for most of the conditions. The main reason for the differences observed in development and evaluation is most likely due to weak back-end parameters estimation of the systems that applied duration and type of speech dependent back-end calibration and fusion. With increased development data, or with a different back-end scheme, we believe that the primary system would be able to provide significant language recognition improvements, closer to the development data results. Anyway, the reported systems were still able to provide promising language detection performances in the different evaluation conditions.

8. References

[1] Campbell, W. M., Sturim, D. E. and Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification", IEEE Signal Processing Letters, vol. 13(5), pp. 308-311, 2006.

[2] Zissman, M., "Comparison of four approaches to automatic language identication of telephone speech", IEEE Transactions on Speech and Audio Processing, vol. 4(1), pp. 31-44, 1996.

[3] "The Albayzin 2010 Language Recognition Evaluation Plan (Albayzin 2010 LRE)", URL: http://jth2010.gts.tsc.uvigo.es/images/stories/pdfs/albayzin_lre10_evalplan_v2.pdf.

[4] Meinedo, H. and Neto, J., "Audio Segmentation, Classification and Clustering in a Broadcast News Task", in Proc. ICASSP 2003, Hong Kong, Apr 2003.

[5] Torres-Carrasquillo, P. A. et alt., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features", in Proceedings of ICSLP 2002, pp. 89-92, Denver, Colorado, Sep 2002.

[6] Lin, C-J, "LIBLINEAR - A Library for Large Linear Classification", URL: http://www.csie.ntu.edu.tw/ cjlin/liblinear/.

[7] Koller, O., Abad, A. and Trancoso, I. "Exploiting variety-dependent Phones in Portuguese Variety Identification", IEEE Odyssey 2010: The Speaker and Language Recognition Workshop, 2010.

[8] Meinedo, H., Alberto, A., Pellegrini, T., Neto, J. and Trancoso, I., "The L^2F Broadcast News Speech Recognition System", in Proc. FALA-2010, Vigo, Spain, Nov 2010.

[9] Abad, A. and Neto, J., "Incorporating Acoustical Modelling of Phone Transitions in an Hybrid ANN/HMM Speech Recognizer", in Proc. INTERSPEECH-08, Brisbane, Australia, Sep 2008.

[10] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", in Proc. ICSLP 2002, Denver, Colorado, Sep 2002.

[11] Brummer, N., "FoCal Multiclass Toolkit", URL: http://niko.brummer.googlepages.com/focalmulticlass.

[12] van Leeuwen, D. and Gonzalez-Dominguez, J., "The TNO system for LRE-2009", The 2009 NIST Language Recognition Evaluation (LRE09) Workshop, Baltimore, US, Jun 2009.

The UVigo-GTM Language Verification Systems for the Albayzin 2010 Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

Department of Signal Theory and Communications, Universidade de Vigo
plopez@gts.tsc.uvigo.es, ldocio@gts.tsc.uvigo.es, carmen@gts.tsc.uvigo.es

Abstract

This paper describes the two systems submitted by the UVigo-GTM group for the Albayzin 2010 language verification evaluation. They were initially thought to perform speaker recognition and verification, so they use language-independent information to apply the algorithms to language verification treating the problem as a pattern recognition task. The principal system consists in a dimensionality reduction approach that transforms the data into a lower dimensionality subspace by performing a two-stage process to reduce the dimensionality and extract a discriminative subspace. The alternative system uses the Nonnegative Matrix Factorization to obtain a representation of the data in terms of a set of basis functions, obtaining the utterances represented as a feature vector of lower dimensionality.

Index Terms: fishervoices, language identification.

1. Introduction

Language recognition is a task that may take into account different kinds of information: linguistic and phonetic information, which is language-dependent, and acoustic information, which is language-independent. The use of language-dependent information implies the training of language models and, sometimes, an in-depth analysis of the target language. On the other hand, the language-independent information does not require any prior knowledge of the target languages, they are all modeled in the same way without taking into account their own characteristics.

The systems developed for the Albayzin 2010 Language Verification Evaluation do not require language-dependent information for some reasons. On the one hand, the less information the system requires to decide among the target languages, the faster the decision. On the other hand, the language recognition task can be thought as a pattern recognition problem, where a sample of data has to be classified into one of the possible classes. This brings the possibility of using speaker recognition algorithms to perform language recognition because, from a pattern recognition perspective, both problems are the same: there are several classes (speakers, languages), and a series of utterances that have to be classified into these classes.

The main system that is described in this paper is based on a face recognition approach [2], where a transformation of the data is performed in order to reduce the dimensionality and to find a discriminative subspace. This Fisherface reduction technique was also applied in speech processing to perform speaker clustering [3]. In previous work [1], this technique was modified and employed to perform speaker identification, and this representation of the speech utterances was named after Fishervoice. As the Fishervoice approach consists in a transformation of the data in order to classify it into different classes (speakers), its application in language recognition is straightforward, and is described in detail in section 4.

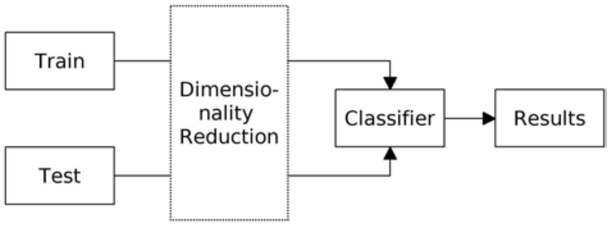


Figure 1: *Fishervoice Language Identification System.*

The alternative system presented in section 5 describes an approach called NMFvoices. It is based in the Nonnegative Matrix Factorization (NMF) [1], a technique to decompose a matrix into two new ones, where one represents a set of basis vectors and the other one the corresponding weights to obtain the original matrix with these basis vectors. NMF is used in this work to represent the speech utterances as the weights obtained with this factorization, obtaining a representation of the data using feature vectors of lower dimensionality.

2. The Language Verification System

The main structure of the language verification system is shown in Fig. 1. The system receives two inputs: a dataset to train the system and to build the models for each different language, and another one to test the performance of the system. A dimensionality reduction step is performed, where the approaches described here are applied, and after that a classifier decides if an utterance is spoken in the target language or not. This decision is taken by doing language identification, the utterance is assigned to the most likely language.

3. Datasets

Three different datasets are necessary to perform speaker verification with the proposed algorithm:

- A train dataset to train a GMM-UBM ($A_{GMM-UBM}$).
- A train dataset (A_{train}) to build a model of the different classes (languages).
- A test dataset (A_{test}) to test the performance of the system: each of its utterances has to be assigned to a class.

This three datasets are composed of several utterances of speech in the different target languages spoken by different speakers. Each utterance is represented as a matrix of dimension $m \times n$ that consists in the means obtained by performing a Maximum a Posteriori (MAP) adaptation of the GMM-UBM to the utterance. Thus, m is the number of gaussians of the GMM, and n is the dimension of the feature space. In this case, the acoustic features that represent the speech utterances are 12 Mel-frequency Cepstral Coefficients (MFCC), extracted using a 25 ms Hamming window at a rate of 10 ms per frame, and augmented with the normalized log-energy and their delta and acceleration coefficients. Thus, the dimension of the feature space (n) used in these experiments is 39. This features are normalized in order to fit a zero mean and a unit variance distribution.

The whole datasets A_{train} and A_{test} are represented as tridimensional matrices of dimension $m \times n \times L_{train}$ and $m \times n \times L_{test}$ respectively, where L_{train} is the number of utterances in A_{train}, and L_{test} is the number of utterances in A_{test}.

4. Fishervoices

4.1. The Fishervoice technique

The Fishervoices technique transforms a dataset A into a set C which represents the same information but with less features, i.e. it reduces the dimensionality of the subspace. To carry out this transformation, two transformation matrices X and Y have to be computed.

Matrix X is obtained by performing 2D-PCA in the A_{train} dataset. Three scatter matrices (between-class D_b, within-class D_w and total D_t) are defined:

$$D_b = \sum_{i=1}^{c} P_i (M_i - M)^T (M_i - M) \tag{1}$$

$$D_w = \sum_{i=1}^{c} \sum_{j, j \in i} (A_{train_j} - M_i)^T (A_{train_j} - M_i) \tag{2}$$

$$D_t = D_b + D_w \tag{3}$$

where c is the number of different languages (classes) in A_{train}, P_i is the a priori probability of the ith class, M_i is the mean matrix of the ith class ($i = 1, 2, \cdots, c$), M is the total mean matrix of A_{train}, and A_{train_j} is the $m \times n$ matrix of the jth segment in A_{train}. Thus, M represents the mean matrix of the whole set, and M_i is the mean matrix of language i.

The eigenvectors and eigenvalues of the total scatter matrix D_t are obtained, finding a matrix X that maximizes $J(X) = X^T D_t X$. The dimensionality reduction is achieved by dropping some of the eigenvectors in X. This is done by keeping only a percentage e_1 of the energy of the subspace E_X:

$$E_X = \sum_{i=1}^{n} \lambda_i \tag{4}$$

where λ_i is the ith greatest eigenvalue of X. Finally, a matrix $X \in \Re^{n \times u}$ is obtained, where u is the number of eigenvectors needed to absorb a percentage e_1 of the energy of the subspace.

The matrix X is employed to transform the set A_{train} into a lower dimensionality subspace by doing $B_{train} = A_{train} X$. Then, a LDA discriminative subspace is computed, obtaining the transformation matrix Y. New between-class and within-class scatter matrices (R_b and R_w, respectively) are computed:

$$R_b = \sum_{i=1}^{c} P_i (L_i - L)(L_i - L)^T \tag{5}$$

$$R_w = \sum_{i=1}^{c} \sum_{j, j \in i} (B_{train_j} - L_i)(B_{train_j} - L_i)^T \tag{6}$$

where L is the total mean matrix of the set B_{train}, and L_i is the mean matrix of the ith class in that set.

The Fisher criterion is applied, thus, a matrix Y that maximizes $J(Y) = \frac{Y^T R_b Y}{Y^T R_w Y}$ is obtained. As before, only a percentage e_2 of the energy of the subspace E_Y is kept, obtaining a matrix $Y \in \Re^{m \times v}$. Then, the dataset B_{train} is transformed into the final subspace $C_{train} = Y^T B_{train}$, where $C_{train} \in \Re^{v \times u}$.

Once X and Y are obtained, they are used to project A_{test} to the new subspace by doing

$$C_{test} = Y^T B_{test} = Y^T A_{test} X \tag{7}$$

4.2. Classifier

After performing the transformation of the datasets, the initial utterances are obtained, but represented in a space of lower dimensionality: the initial number of features to describe an utterance was $m \cdot n$, while after the transformation of the feature space it is reduced to $v \cdot u$, where $v \leq m$ and $u \leq n$.

The classification of each utterance in C_{test} is done by comparing these utterances with the models of the different languages C_{train}. An utterance S is compared to the different models by measuring the spatial distance between the utterance and the model:

$$T = \min_i d(C'_{test_S}, C'_{train_i}) \tag{8}$$

$d(.,.)$ is the euclidean distance between an utterance and a model (which is another utterance), C'_{test_S} is a vector that represents the utterance S in terms of a supervector obtained by concatenating the rows of C_{test_S}, and the same for C'_{train_i}.

The classifier in Eq. (8) decides which of the utterances in the model is spatially closer to the test utterance, assigning the language of the model to this test utterance. This is not an actual language verification system, it performs verification by performing language identification.

4.3. Processing speed

Three different processes had to be executed to run the experiments with the test dataset provided for the Albayzin 2010 Language Verification Evaluation. The CPU time necessary to run each process is shown below:

- Parameterization (with HTK-3.4): 722.83 s
- Normalization of the features (C code): 32.97 s
- Computing and generation of the results file (MatLab): 1464.96 s
- **TOTAL: 37 min 0.76 s**

$$\left.\begin{array}{l} 4992\ recognitions \\ \\ 2220.76\ s \end{array}\right\} 2.25\ recognitions/s$$

This processes were executed in a server with a processor Intel Xeon E5620 2.4 GHz and 18 GB of memory.

5. NMFvoices

5.1. The NMFvoices Technique

Nonnegative matrix factorization (NMF) is a dimensionality reduction technique employed over nonnegative data. Given a data matrix $V \in \Re^{\geq 0, F \times N}$, NMF finds a factorization

$$V \approx WH \tag{9}$$

where $W \in \Re^{\geq 0, F\times R}$ and $H \in \Re^{\geq 0, R\times N}$. R is the value that performs the dimensionality reduction, and it is usually chosen in a way that $FR + RN \ll FN$.

NMF is an iterative algorithm whose target is to reduce the euclidean distance between V and WH, its divergence, etc. In this paper, a multiplicative algorithm is employed, which is fast and easy to implement. Its update rules are:

$$H \leftarrow H \frac{W^T V}{W^T W H} \quad (10)$$

$$W \leftarrow W \frac{V H^T}{W H H^T} \quad (11)$$

The target of this iterative algorithm will be to minimize the euclidean distance $\|V - WH\|$.

The F row vectors of W can be interpreted as basis vectors, and the N column vectors of H would be the corresponding weights needed to obtain each of the vectors in V by combining the basis vectors.

5.1.1. Nonnegativity

A requirement of NMF is that matrices V, W and H have no negative or zero values, so a little adjustment of the data is performed in order to transform a matrix into a non-negative one. Given a matrix $M \in \Re^{I\times J}$:

$$M_+ = M \cdot \min_{i,j} m_{ij},\ i \leq I,\ j \leq J \quad (12)$$

where m_{ij} is the element (i, j) in matrix M, and M_+ is the obtained non-negative matrix. This adjustment will be applied if V or the initialization of W and H have negative or zero values.

5.1.2. Initialization of matrices W and H

The most usual way to initialize W y H is by doing it randomly. Nevertheless, as the matrices will be different for any trial of an experiment, it would be interesting to use a deterministic manner to initialize them. In this paper an initialization algorithm is proposed: given a matrix $V \in \Re^{F\times N}$, the R most different rows from this matrix will be selected, For each row f_i, the following distance measure is computed:

$$D(f_i) = \sum_{j\neq i, j=1}^{F} d(f_i, f_j) \quad (13)$$

where $d(f_i, f_j)$ is the euclidean distance between the rows f_i and f_j. Then, the initial matrix H will be composed by the R rows which obtained the higher values for $D(f_i)$.

Once H is obtained, W is computed by doing:

$$W = VH^T \quad (14)$$

5.2. Using W as a basis

A matrix V' can be represented by using the basis vectors obtained by performing NMF in another matrix V. V is decomposed into the two matrices W and H. Then, for matrix V', W' is initialized as $W' = W$ and $H' = W^T V'$. While executing NMF, the update rule (11) is not applied, so H' will be the only updated matrix, while W' will remain the same. Thus, as the weight matrices H' and H were obtained using the same basis vectors, they can be compared to each other.

5.2.1. NMF and its application in language recognition

In 5.1, it was explained that NMF makes it possible to decompose a matrix V into two matrices W, representing a set of basis vectors, and H, representing the corresponding weights. The algorithm to obtain the representation of the speech utterances in NMFvoices is as follows:

- Decompose the matrix A_{train} into two matrices W_{train} and H_{train}.
- Decompose the matrix A_{test} into two matrices W_{test} and H_{test}, but restricting the algorithm so that $W_{test} = W_{train}$ as explained in 5.2.
- The columns of H_{train} are the models for the possible languages, while the columns of H_{test} represent the test utterances that have to be assigned to the target languages.

5.3. Classifier

The procedure to perform language recognition is simple: each column of H_{test} will be compared to all the columns in H_{train}. The language of the column in H_{train} that is closer to the column in H_{test} is the one that will be assigned to that speech utterance. The classifier employed to compare the utterance is the same as in Eq. (8):

$$T = \min_i d(H_{test_S}, H_{train_i}) \quad (15)$$

where H_{test_S} is the column S of the matrix of the test weights, and H_{train_i} is the ith column of the matrix of the train weights.

5.4. Processing speed

As in the Fishervoice system, three processes had to be executed to run the experiments with the test data. The CPU time necessary to run each process was:

- Parameterization (with HTK-3.4): 722.83 s
- Normalization of the features (C code): 32.97 s
- Computing and generation of the results file (MatLab): 3352.52 s
- **TOTAL: 1 h 8 min 28.32 s**

$$\left.\begin{array}{l} 4992\ recognitions \\ \\ 4108.32\ s \end{array}\right\} 1.215\ recognitions/s$$

This processes were executed in a server with a processor Intel Xeon E5620 2.4 GHz and 18 GB of memory.

6. Train and development data

Initially, the Fishervoice algorithm was thought to perform speaker identification, as in [1]. The BANCA database [7] [8] was used for this experiments. When it was first developed to perform language recognition, the database that was employed is the COST278 Pan-European Broadcast Database [4]. This database includes broadcast news programs in 9 different European languages (Belgian Dutch, Portuguese, Galician, Czech, Slovenian, Slovak, Greek, Hungarian and Croatian) featuring clean speech and noisy speech. Nevertheless, in some of the experiments a GMM-UBM trained with data extracted from the Transcrigal Database [5] was employed.

In the case of the NMFvoices approach, it was initially developed to perform speaker verification, and was successfully tested using the BANCA database. After this, it was decided to transform it into a recognition system by changing the classifier, and it was tested with the development set of the KALAKA-2 database.

The only data employed to train the systems for the Albayzin 2010 Language Verification Evaluation was the database provided for that purpose, the KALAKA-2 database. This data is divided into three groups, namely *GMM*, *Devel* and *Test*. Attending to the datasets described in 3, the given data was distributed as follows:

- *GMM*: This is the only labeled data available, and includes clean and noisy utterances. Only the clean speech was employed, and it was divided into two groups of the same size: one employed to train the $GMM - UBM$ (corresponding to the dataset $A_{GMM-UBM}$) and the other one used to train the transformation matrices and, therefore, to build the models for the different languages (corresponding to the dataset A_{train}).
- *Devel*: All the data in this group was employed to tune the free parameters of the system, and to decide the best features for the algorithm. Thus, this corresponds to the dataset A_{test} in a first stage. The algorithm was run using different values for m, e_1 and e_2 in the case of Fishervoices, and different values of M and R in the case of NMFvoices. Different types of features were also tested(39 MFCCs, 39 MFCCs normalized to fit a zero mean and a unit variance distribution, and 39 MFCCs applying the M-norm after the MAP adaptation of the GMM-UBM). The parameters and features that obtained the best average cost were chosen:
  - Fishervoices: the best values were $M = 64$; $e_1 = 1.0$ and $e_2 = 0.5$ for clean speech, and $e_1 = 0.6$ and $e_2 = 0.8$ for noisy speech.
  - NMFvoices: the best values were $M = 32$; $R = 50$ for clean speech and $R = 100$ for noisy speech.
- *Test*: This data was employed to run the test that was submitted for the evaluation. Thus, in a stage after the tuning, this data corresponds to the dataset A_{test}.

7. Acknowledgements

The work reported here was partially supported by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04 “Búsqueda de Información en Contenidos Audiovisuales plurilingües” and by the Ministry of Science and Innovation under the FPI grant BES-2010-033358.

8. References

[1] P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, “A Fishervoice-based Speaker Identification System”, Proceedings of FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop.

[2] X.Y. Jing, H.S. Wong and D. Zhang, “Face Recognition Based on 2D Fisherface Approach”, Pattern Recognition, vol. 39, n. 4, pp. 707–710, 2006.

[3] S.M. Chu, H. Tang, T.S. Huang, ”Fishervoice and Semi-supervised Speaker Clustering“, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4089–4092, Los Alamitos, CA, USA, 2009.

[1] Daniel D. Lee and H. Sebastian Seung, “Algorithms for Non-negative Matrix Factorization”, Advances in Neural Information Processing Systems (NIPS) vol. 13, 556–562, 2000. Online: `http://citeseer.ist.psu.edu/lee01algorithms.html`

[4] A. Vandecatseye, J.-P. Martens, J. Neto, H. Meinedo, C. Garcia-Mateo, J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, C. Alexandris, “The COST278 pan-European Broadcast News Database”, Fourth International Conference on Language Resources and Evaluation (LREC ’04), Lisbon, Portugal.

[5] C. Garcia-Mateo, J. Dieguez-Tirado, A. Cardenal-Lopez, L. Docio-Fernandez, “Transcrigal: a bilingual system for automatic indexing of broadcast news”, Proc. International Conference on Language Resources and Evaluation, Lisbon, Portugal, May 2004.

[7] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Mariethoz, J. Matas, K. Messer, F. Poree, B. Ruiz, “The BANCA Database and Evaluation Protocol”, 2003.

[8] The BANCA Database Website, Online: `http://www.ee.surrey.ac.uk/CVSSP/banca/`

AhoLab Speaker Diarisation System for Albayzin 2010

Iker Luengo, Eva Navas, Ibon Saratxaga, Inmaculada Hernáez, Daniel Erro

Dpt. of Electronics and Telecommunications
University of the Basque Country
{iker.luengo, eva.navas, ibon.saratxaga, inma.hernaez, daniel.erro}@ehu.es

Abstract

This paper presents the speaker diarisation system presented by Aholab Signal Processing Laboratory to the Albayzin Speaker Diarization Challenge 2010. The system was built to run on-line, without any recording of the audio data to produce its output. As a result, the whole process must be done in a single iteration, which prevents the use of many optimisation processes that are usually implemented in diarisation systems. In order to minimise the reduction of the accuracy in the output and to maximise computational efficiency, the applied algorithms were carefully selected and some new modifications were implemented.

Index Terms: speaker diarisation, BIC, on-line audio processing

1. Introduction

The aim of speaker diarisation is to detect speaker changes in an audio recording, and to identify which of the resulting speech segments come from the same speaker. That is, the task is to detect 'who speaks when'.

The process is usually divided into different subtasks, each of them dealing with a specific subproblem. Typically the subtasks include speech detection, speaker segmentation, clustering and resegmentation of the audio stream. Usually these subtasks are executed one after another, as a series of processing steps, i.e., each subtask is applied to the whole audio recording before starting the next one. This architecture implies having the whole audio recording available for several processings, and also that no result can be obtained until the recording is finished.

Against this off-line architecture, we propose an on-line algorithm for speaker diarisation, which is capable of performing the whole process in a single iteration. Such algorithm can work with direct audio input or audio streaming, without needing to record it. Furthermore, it is more efficient in terms of execution time and memory requirements. Nevertheless, some accuracy loss is expected, as the system can only rely on past audio samples to make decisions, and not future ones. Also the impossibility of using multi-pass or iterative algorithms may provide suboptimal results.

2. Speaker diarisation: a short overview

2.1. Speech detection

This step is necessary in order to discard audio segments without speech, thus making the following steps easier and less error prone. Depending on which environment the system is going to operate, non-speech segments can contain a large variety of acoustic events: silence, noise, music, applause, shouting, etc. The most common approach is to make a Viterbi segmentation with GMM models trained on labelled data, although sometimes more elaborate models such as multistate HMM are used.

It is possible to use only two models (one for speech and another one for non-speech), although it is convenient to train more specific models when different acoustic events are expected. Often models for noise, music, clean speech, speech+noise and speech+music are used [1], but it is also possible to train separate models for male and female speech or to differentiate wideband and narrowband speech [2]. This way the detection accuracy is improved, assuming there is enough training data.

2.2. Speaker change detection

This is a critical step in the diarisation process. Once the segments without speech are discarded, the locations of the speaker changes must be located. Almost every diarisation system performs this step calculating a distance metric between two adjacent audio windows. If the distance is larger than a given threshold, a speaker change is assumed between both windows. The differences among specific algorithms lie in the choice of the distance metric and the windowing scheme.

One of the most widely used metric for change detection is the Bayesian information criterion (BIC) [3]. BIC is a likelihood criterion penalised by the model complexity, and is usually used for model selection. Therefore, it can be used to estimate if the windowed audio is better modelled with two different distributions (one for each window) or a single one (combining both windows), thus effectively detecting changes in the audio stream. Nevertheless, this algorithm is computationally very expensive. That is why some implementations prefer to use other metrics, which may be less accurate but faster. Examples of these metrics are Hotlling's T^2 [4], the Gaussian divergence [2] or the generalised likelihood ratio (GLR) [5]. It is also possible to use these faster metrics as a first approach, and then refine the detected changes with BIC [6].

Regarding the windowing, the simplest way to implement the algorithm is to use two adjacent fixed-size sliding windows. The distance metric between both windows is calculated, and the peaks in the distance function define the locations of the change points. A more elaborated windowing scheme that is usually applied together with the BIC metric uses a growing window. Every audio frame inside the window is a possible change point, and the BIC value for each of them is calculated. If the highest BIC value inside the window is greater than zero, a change point is detected at the position of that maximum, and the window is reset to its original size. If not, the window grows a fixed length and the process starts again. Generally, the growing window scheme provides better accuracy results, but is also computationally more expensive. Some implementations reduce this computation cost avoiding to search for changes in improbable places and defining a limit to the window length [7, 8].

2.3. Clustering

Once the boundaries among speakers are known, it is necessary to detect which speech segments belong to the same speaker. This step is usually implemented as a clustering process, the bottom-up clustering being the most common method. The distance between all cluster pairs is calculated, and the pair with least distance is selected and combined. Then, the distance matrix is updated and another pair is selected until the stop criterion is met. Many distance metrics can be used: BIC, GLR, Euclidean distance between GMM, etc.

2.4. Cluster recombination

Although this step is not absolutely necessary, it may improve the final result [1]. The idea here is to under-cluster the segments in the first clustering process, thus having more clusters than speakers, but at the same time ensuring that the clusters have a reasonable amount of speech. A GMM model is created for every cluster by MAP adaptation from a UBM. Finally, the GLR is used in order to identify which clusters to recombine. A new model is created for the newly recombined cluster, and the process is repeated until the stop criterion is met. It has been shown that feature normalisation is necessary to get any improvement with this technique [2].

2.5. Resegmentation

Now that each cluster has a reasonable amount of information, it is possible to train new models for each speaker, and use them in combination with the non-speech models in order to make a new Viterbi segmentation. This way it is possible to refine the segmentation boundaries. This process can be repeated iteratively for increased accuracy.

3. Proposed algorithm description

Figure 1 shows a schematic diagram of the proposed diarisation algorithm. The algorithm is based on an efficient implementation of a BIC change detector and an on-line speaker clustering. The change detector works with MFCC features without derivatives, whereas the speech detector appends first and second derivatives to this parametrisation. Furthermore, the BIC algorithm uses voiced frames only, discarding any unvoiced segment. In the following sections, each step in the algorithm will be explained with more detail.

3.1. Speech detection

A separate GMM model was trained for music, noise, clean speech, speech+music and speech+noise, using the development recordings and the audio segmentation labels provided by the contest organisation. These models are used in a Viterbi segmentation in order to detect audio segments with and without speech.

As the process ought to be on-line, an on-line Viterbi algorithm as described in [9] was implemented. This modified algorithm keeps track of the active paths and efficiently detects whether they converge in some point or not. In the case of all the paths converging in a point, the segmentation decision up to that point can be extracted, without loosing any accuracy and without needing to wait until the whole file has been processed. Furthermore, the part of the trellis that contained the consolidated path can be erased from memory, as it is not needed anymore.

Development experiments showed that the addition of first and second derivatives of MFCC provides slightly better segmentation results.

3.2. Speaker segmentation

For the initial speaker change detection, a growing window architecture and BIC metric are used. The growing window provides better results than a fixed-size sliding window, but the computational cost is also larger. In order to reduce the time of computation as much as possible, the solution described in [8] is used:

- No speaker change is searched in the first and last 2 seconds of the window.
- The window grows 2 seconds every time that no change is detected.
- Once the window reaches 20 seconds, instead of growing, it becomes a sliding window.
- For each window, a speaker change is searched every 250 ms. If a change is located, the search is refined to 50 ms.
- Once a change is found, the window size is reset to 5 seconds.

This solution provides the accuracy of a growing-window algorithm accuracy, while keeping the window size and the amount of calculation to a minimum. Furthermore, the calculation of the BIC values is also optimised by using a buffer of cumulative sums as described in work made by Cettolo and Vescovi [8].

Development results showed that discarding unvoiced frames and using only voiced ones decreased the diarisation error in a 12%. Therefore, only voiced frames were used for the speaker change detection. Similarly, it was confirmed that the use of feature derivatives was not convenient for this task.

3.3. Clustering

An on-line clustering that uses BIC metric was implemented, following the description given in [7]. Every time the speaker segmentation algorithm detects a new boundary, the newly extracted speech segment is immediately given to the clustering process. This process computes the BIC of this new segment against all known clusters, and selects the one with lowest BIC value. If this value is under a given threshold, the segment is assigned to that cluster and the cluster statistics are updated. If not, a new cluster is formed with the new segment.

This on-line clustering is theoretically suboptimal, since the clustering is performed without the information of forthcoming segments. However, in practice the on-line clustering may give better results than the bottom-up off-line clustering. The reason is that the speaker segmentation algorithm over-segments the speech, providing false speaker boundaries. Therefore, two consecutive segments are more likely to belong to the same speaker than segments far apart. The on-line clustering makes the clustering decision more locally, thus reinforcing the combination of adjacent segments. This algorithm not only does the clustering on-line, but it also provides better results.

3.4. Voiced unvoiced detection

As described before, the speaker change detection step uses only voiced frames, discarding the unvoiced ones. In order to

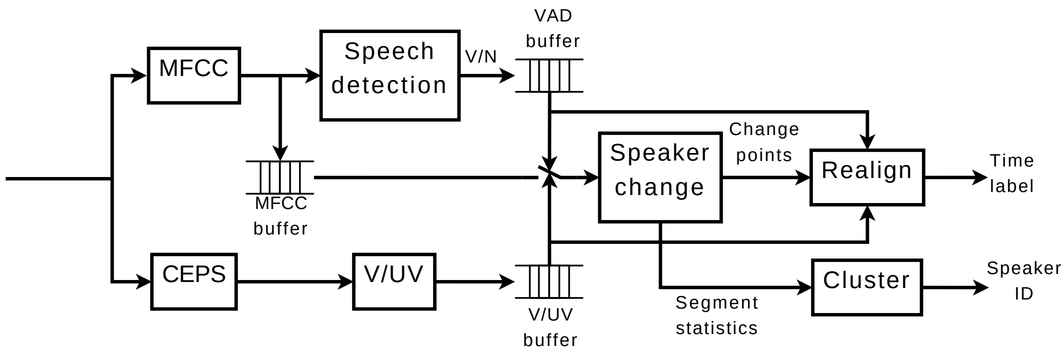


Figure 1: Schematic diagram of the diarisation algorithm.

make the voiced/unvoiced (VUV) estimation, the PTHCDP algorithm described in [10] was used. This algorithm uses cepstrum transformation and dynamic programming in order to estimate the F_0 curve and the VUV information. The algorithm was modified in order to use the on-line Viterbi algorithm, so that partial paths can be extracted with the VUV information every time the active paths converge.

3.5. System integration

The whole algorithm is meant to run on-line in a single iteration. But the clustering subprocess must wait until a speaker change is detected, and the speaker change subprocess is idle until the speech detector and the VUV detector make a decision. Both detectors make decisions asynchronously, depending on their own Viterbi algorithm and the convergence of their paths. Meanwhile, new feature frames are being generated from the audio input. For the synchronisation of all these subprocesses, a series of buffers and control points must be used.

Every time a new frame is collected, it is parametrised and delivered to the speech detection and VUV detection algorithms. Furthermore, the MFCC parametrisation is saved in a buffer for later use. In order to deal with the asynchronous behaviour of the speech detection and VUV detection algorithms, their outputs are directly stored in the corresponding buffer as soon as path convergence is detected. Now, let N be the number of speech decisions available in the VAD buffer, and M the number of VUV decisions available in the corresponding buffer. Then, we have a total of $min\{N, M\}$ new frames with all decisions made, so that they can be further processed. These frames are extracted from the MFCC buffer and the ones without speech and the voiced ones are discarded. The rest (if any) are provided to the speaker change algorithm.

As these frames are not needed anymore, they are deleted from the MFCC buffer, in order to save memory. At the same time, the time realignment process gets some information from the speech detection and VUV decisions for later use, and these decisions are also deleted from the buffers.

Whenever the speaker change algorithm finds a new boundary, it outputs the necessary statistics for the clustering process, together with the time instant in which this change happens. As this algorithm takes only voiced frames as input, this time mark does not consider non-speech or unvoiced frames. Therefore, a time realignment is necessary in order to convert the detected change times into absolute times. Finally, the clustering process outputs a unique speaker label, and the time realignment system outputs the corresponding time labels.

4. Analysis of the results

The presented algorithm was submitted as the primary system to the Albayzin Speaker Diarisation Challenge 2010. Nevertheless, two other systems were also presented as reference. The first one was an off-line version of the main algorithm, in which each subprocess (speech detection, VUV decision, speaker change detection and the clustering of the segments) was executed one after the other. The second one was an off-line version in which unvoiced frames were also considered for the speaker change detection algorithm, thus not needing a VUV decision step. This section compares the three systems in terms of diarisation accuracy and execution speed. See Table 1 for the error rates and Table 2 for the execution speed of each one of these systems.

The main difference between the on-line and off-line versions is post-processing. The algorithms used in both cases are the same, but the off-line architecture allows us to post-process the outcome of each step before going into the next one, which is not possible in the on-line system. For example, speech detection labels were post-processed in order to discard silences shorter than 500 ms, before feeding them to the speaker change detection algorithm. This provides a better estimation of speech activity, which is reflected in a much lower missed speaker error rate (see Table 1). However, the results from the speaker change detection or the clustering systems are not post-processed. As a result, the speaker error rate does not change much between the on-line and off-line architectures, and the small difference is due to the speech detection labels being more accurate.

Table 1 also shows the effect of discarding unvoiced frames for the speaker change detection step. When unvoiced frames were used in the off-line system, the speaker error rate increased a 16%. As the speech detection algorithm always uses both voiced and unvoiced frames, there is no difference in the missed speaker and false alarm error rates.

It is also interesting to compare the systems in terms of execution speed. Table 2 shows the average CPU time required in order to process one hour of speech. For the off-line implementations, the time required for each step is also shown. These measures were made on a quad-core Intel Xeon 2.27 GHz computer with 6 GB memory. Nevertheless, these values are not meant to be an absolute measure of the complexity of each system, but can be used in order to see which one is faster.

The on-line architecture makes a single iteration over the speech data, whereas the off-line systems must perform several iterations and post-processings. As a result, there is a significant difference in the processing speed. The off-line architec-

Error time (in %)	Off-line w/ unvoiced	Off-line	On-line
Missed speaker	2.8	2.8	4.9
F-alarm speaker	1.2	1.2	1.5
Speaker error	26.9	23.1	23.9
Diarisation error	31.00	27.17	30.38

Table 1: Error rates for each system on the Albayzin Speaker Diarisation Challenge 2010.

CPU time (in seconds)	Off-line w/ unvoiced	Off-line	On-line
Speech Detection	36.0	36.0	–
VUV detection	–	81.0	–
Diarisation	47.1	44.2	–
Total	83.1	161.2	126.1

Table 2: Mean CPU time for one hour speech for each one of the systems.

ture needs 161 CPU seconds to process one hour of speech, whereas the on-line systems needs 126 seconds. This means that the on-line implementation is a 22% faster, mainly due to the reduction in the number of iterations.

It is outstanding that half of the execution time of the off-line system without unvoiced frames is spent in the VUV detection step. When unvoiced frames are not discarded, this step is unnecessary, and the total execution time decreases to 83 seconds. Although we have not the corresponding measurements, is is expected that an on-line system without VUV detection would be even faster.

5. Conclusions

Most of the current speaker diarisation systems rely on an off-line architecture, in which several processing steps are performed over the same audio recording, one after the other. In this paper an on-line diarisation system has been described. This algorithm requires a single iteration in order to process the audio, and can be used with direct audio input or audio streaming. This features makes it suitable for applications where the recording of the signal is not a possibility.

As all the processing must be done in a single iteration, it is not possible to post-process the outcome of each step before going into the next one. Therefore, a certain increase of the error rate is unavoidable. The results show that the on-line architecture has indeed a 12% increase in the total diarisation error rate when compared to the off-line system. This increase is mostly due to a higher missed speaker error rate, which in turn occurs because it is not possible to post-process the speech detection labels.

Nevertheless, making the system run on a single iteration has its advantages in terms of speed. It has been shown that the on-line architecture is a 22% faster than the same algorithms running in an off-line fashion.

The possibility of including or discarding unvoiced frames in the speaker change detection algorithm has also been studied. On the one hand, discarding these frames provided a significant reduction in the speaker error rate. On the other hand, in order to discard these frames, a VUV detection step is mandatory, which increases the computational cost of the system. For example, the VUV detection algorithm that was used in the experiments [10] takes half of the total execution time. If the diarisation algorithm must run on devices with severe processing restrictions, faster VUV detection algorithms should be used, or even the whole VUV detection step can be avoided if the speaker change detection is performed with both voiced and unvoiced frames.

6. Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and The Basque Government (Berbatek, IE09-262).

7. References

[1] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations," in *NIST Rich transcription Workshop*, Palisades, NY, USA, Nov. 2004.

[2] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The cambridge university march 2005 speaker diarisation system," in *Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 2437–2440.

[3] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA speech recognition workshop*, 1998, pp. 127–132.

[4] B. Zhou and J. Hansen, "Unsupervised audio stream segmentation and clustering via the bayesian information criterion," in *ICSLP*, Beijing, China, Sep. 2000, pp. 714–717.

[5] S. Meignier, D. Moraru, C. Fredouille, J. F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, vol. 20, pp. 303–330, 2006.

[6] P. Delacourt and C. J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Communication*, vol. 32, pp. 111–126, 2000.

[7] A. Tritschler and R. A. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Eurospeech*, Budapest, Hungary, Sep. 1999, pp. 679–682.

[8] M. Cettolo and M. Vescovi, "Efficient audio segmentation algorithms based on the bic," in *International Conference on Acoustics, Speech, and Signal Processing (ICCASP'03)*, vol. 6, april 2003, pp. 537–5340.

[9] R. Šrámek, "The on-line Viterbi algorithm," Master's thesis, Comenius University, Bratislava, 2007.

[10] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sánchez, and I. Sainz, "Evaluation of pitch detection algorithms under real conditions," in *ICASSP*, Honolulu, USA, Apr. 2007, pp. 1057–1060.

GTTS System for the Albayzin 2010 Speaker Diarization Evaluation

Mireia Diez, Mikel Penagarikano, Amparo Varona,
Luis Javier Rodriguez-Fuentes, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain

mireia_diez@ehu.es

Abstract

This paper briefly describes the diarization system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the Albayzin 2010 Speaker Diarization Evaluation. The system consists of three decoupled elements: (1) speech/non-speech segmentation; (2) acoustic change detection; and (3) clustering of speech segments. Speech/non-speech segmentation is performed by means of one of the systems presented to the Albayzin 2010 Audio Segmentation Evaluation. With the aim to detect speaker changes, speech segments are further segmented by means of a naive metric-based approach which locates the most likely spectral change points. The third element is based on a dot-scoring speaker verification system: speech segments are represented by MAP-adapted GMM zero and first order statistics, dot scoring is applied to compute a similarity measure between segments (or clusters) and finally an agglomerative clustering algorithm is applied until no pair of clusters exceeds a similarity threshold.

Index Terms: Speaker Diarization, Dot Scoring, Sufficient Statistics

1. Introduction

This paper briefly describes the dot-scoring speaker diarization system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the Albayzin 2010 Speaker Diarization Evaluation. The system is based on three subsystems: an audio classifier developed for the Albayzin 2010 Audio Segmentation Evaluation, an acoustic change detector which was part of the system submitted to the Albayzin 2006 Speaker Tracking Evaluation [1], and a speaker verification system developed for the NIST 2010 Speaker Recognition Evaluation [2].

2. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features. The MFCC set, comprising 13 coefficients, including the zero (energy) coefficient, was computed in frames of 32 ms at intervals of 10 ms for the two first modules (audio segmentation and acoustic change detection). In the clustering approach, the MFCC set was computed in frames of 20 ms at intervals of 10 ms and augmented with dynamic coefficients (13 first-order and 13 second-order deltas), resulting in a 39-dimensional feature vector. Also, an energy based voice activity detector (VAD) was applied to remove those fragments (short silences) with an energy level of 30 dB (or more) under the maximum. All the speech processing computations were done by means of the Sautrela toolkit [3].

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

3. Speech/non-speech segmentation

For this task, a simple audio segmentation system was developed, which considered five acoustic classes: (1) music, (2) clean speech, (3) speech with music in the background, (4) speech with noise in the background and (5) other (noise, long silence fragments, etc.). An ergodic Continuous Hidden Markov Model with 5 states and 512 mixtures per state was defined, using the Sautrela toolkit, under the Layered Markov Models framework [4].

The emission distributions were independently estimated for each state, applying the Baum-Welch algorithm on the corresponding sets of segments extracted from the reference segmentations of 12 development sessions. The number of mixtures per state and the transition probabilities (auto-transitions fixed to 0.999999, transitions between states and final state transitions fixed to $2 \cdot 10^{-7}$) were optimized on audio segmentation experiments over the remaining 4 development sessions. Though system performance was quite poor for the 4-class setup defined in the evaluation, when considering a 2-class speech/non-speech classification setup, the false alarm error rate was 1.16% and the miss error rate was around 1.78% for the speech class (including the three sub-classes mentioned above). Note that, since around 3% of the speech frames are mistaken, our speaker diarization error will be, at best, of that order. More details can be found in the description of the GTTS submission to the Albayzin 2010 Audio Segmentation Evaluation.

4. Acoustic change detection

Speech segments produced by the speech/non-speech detector may contain various speakers, so before clustering, a further segmentation is needed to detect speaker changes. We presented a very simple approach to detect acoustic changes (i.e. any change of speaker, background or channel conditions) in our submission to the Albayzin 2006 Speaker Tracking Evaluation (see [1] for details).

Though it was found that not only speaker changes were detected, but also many other changes, even those related to the presence of spontaneous speech events (filled pauses, coughs, etc.), the key point was that *almost all the speaker changes were detected*. Note that consecutive short segments corresponding to the same speaker can be grouped together by the clustering algorithm.

As other *metric-based* approaches (e.g. [5]), our algorithm defines and applies a metric to compare the spectral stastistics at both sides of successive points of the audio signal, and hypothesizes those boundaries whose metric values exceed a given threshold. In our approach, a kind of *normalized* crossed-BIC (XBIC) [6] is applied:

$$d(X,Y) = -\log\left(\frac{P(x|\lambda_y)P(y|\lambda_x)}{P(x|\lambda_x)P(y|\lambda_y)}\right) \tag{1}$$

The algorithm considers a sliding window W of N acoustic vectors and computes the likelihood of change at the center of that window, then moves the window n vectors ahead and repeats the process until the end of the vector sequence. To compute the likelihood of change, each window is divided in two halfs, W_l and W_r, then a Gaussian distribution (with diagonal covariance matrix) is estimated for each half and finally the cross-likelihood ratio (Eq. 1) is computed and stored as likelihood of change. This yields a sequence of cross-likelihood ratios which is post-processed to get the hypothesized segment boundaries. This involves applying a threshold τ and forcing a minimum segment size δ. In practice, a boundary t is validated when its cross-likelihood ratio exceeds τ and there is no candidate boundary with greater ratio in the interval $[t-\delta, t+\delta]$. All the parameters were heuristically optimized on the development set. The optimal values were $N = 500$, $n = 10$, $\tau = 1800$ and $\delta = 0.6$ seconds.

5. Clustering of speech segments

5.1. Gaussian Mixture Models

More than 35 hours of TV broadcast speech in Spanish, Catalan, Galician and Basque, taken from the Kalaka database [7], were used to train a gender independent GMM (Universal Background Model, UBM) consisting of 256 mixture components. Again, the Sautrela toolkit was used to estimate GMM parameters, applying binary mixture splitting, orphan mixture discarding and variance flooring.

5.2. Sufficient statistics

Zero (n) and first order (x) sufficient statistics were computed for each speech segment. The one-iteration relevance-MAP adapted and normalized mean vectors $m = \frac{\mu_{map}-\mu_{UBM}}{\sigma}$ were computed according to the following expression [8, 2]:

$$m = (\tau\mathbf{I} + diag\,(n))^{-1} \cdot x$$

5.3. Dot scoring similarity measure

Linear scoring (dot-scoring) is a simple and fast technique used in speaker verification that makes use of a linearized procedure to score test segments against target models. Given a feature stream f (the target signal) and a speaker spk, the first-order Taylor-series approximation to the GMM log-likelihood is given by:

$$\log P\,(f|spk) \approx \log P\,(f|UBM) + m_{spk}^t \cdot \nabla P(f|UBM)$$

where m_{spk} denotes the normalized mean vector of speaker spk, ∇ denotes the gradient vector with regard to the standard-deviation-normalized means of the UBM, and $\nabla P(f|UBM) = x_f$ is the first order statistics vector of the target signal f. Then, the log-likelihood ratio between the target model and the UBM, used for scoring, can be expressed as follows:

$$score\,(f, spk) = \log\frac{P\,(f|spk)}{P\,(f|UBM)} \approx m_{spk}^t \cdot x_f$$

For the diarization task, the similarity $sim(a,b)$ between two segments a and b was defined as:

$$\begin{aligned} sim\,(a,b) &= min\left\{score\,(f_a, spk_b)\,, score\,(f_b, spk_a)\right\} \\ &= min\left\{m_b^t \cdot x_a, m_a^t \cdot x_b\right\} \end{aligned}$$

5.4. Score normalization

TZ normalization was applied to dot-scores. Two development sessions were used for the estimation of T-norm and Z-norm parameters. Taking into account score normalization, the similarity measure was redefined as:

$$sim\,(a,b) = min\left\{score\,(f_a, spk_b)^{TZ}, score\,(f_b, spk_a)^{TZ}\right\}$$

5.5. The clustering algorithm

The similarity measure defined above was used to perform agglomerative hierarchical clustering. Given two segments (or two clusters of segments), if they are clustered together, computation of sufficient statistics for the joint cluster is straightforward:

$$\begin{aligned} x_{a+b} &= x_a + x_b \\ n_{a+b} &= n_a + n_b \end{aligned}$$

This leads to a very simple clustering algorithm:

1. **Find** $s_{max} = \underset{\forall(a,b)}{\text{argmax}}\left\{sim\,(a,b)\right\}$
2. **If** $s_{max} < \Theta$ **then** STOP
3. **Set** $x_a = x_a + x_b$
   $n_a = n_a + n_b$
4. **Remove** cluster b
5. **Jump to** 1

Based on preliminary results on the development set, the threshold Θ was set to 3.38. Figure 1 shows system performance as a function of Θ, for the development sessions 3-16. Note that results are consistent across sessions, the optimal performance being attained for threshold values between 3 and 4.

6. Results

Table 1 shows the performance of the clustering algorithm described above on the evaluation set, using four different segmentations:

- Seg1: Reference Speaker Segmentation
- Seg2: Reference Speaker Segmentation + GTTS Acoustic Change Detection
- Seg3: Reference Acoustic Segmentation + GTTS Acoustic Change Detection
- Seg4: GTTS Acoustic Segmentation + GTTS Acoustic Change Detection

The Overall Speaker Diarization Error obtained with the Reference Speaker Segmentation (Seg1, 20.48%) would be the best performance that our clustering system could reach for the evaluation set. The difference between this result and the result obtained by the fully automated system (Seg4, 33.16%) may be explained as follows:

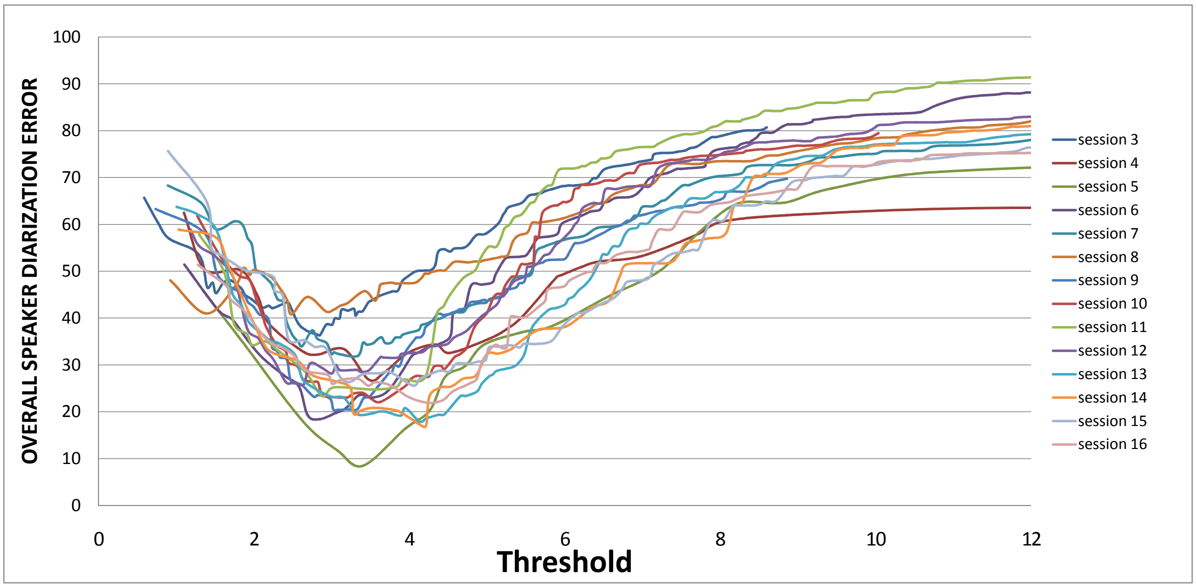


Figure 1: Overall Speaker Diarization Error as a function of the similarity threshold applied as stopping criterion in the clustering algorithm, for sessions 3-16 of the development set.

Table 1: Overall Speaker Diarization Error obtained by applying the clustering algorithm on four different segmentations of the evaluation set (see text for details).

	Seg1	**Seg2**	**Seg3**	**Seg4**
OSDErr (%)	20.48	26.14	29.61	33.16

- Difference between Seg3 and Seg4: 3.55%. Seg3 starts from a perfect audio classification, whereas Seg4 applies the GTTS audio classification system. So, the difference can be explained by the acoustic classification error.
- Difference between Seg1 and Seg2: 5.66%. Since both systems take the reference speaker segmentation as a starting point, the difference in performance can only be due to over-segmentation. Applying the acoustic change detector on the optimal speaker segmentation does not remove speaker boundaries but produces many short segments whose statistics strongly depend on local variabilities. This explains why the performance of the clustering algorithm, which is based on those statistics, degrades for short segments.
- Difference between Seg2 and Seg3: 3.47%. Seg2 includes all the speaker boundaries (plus a number of acoustic changes inside speaker turns), whereas Seg3 may be missing some of them. This explains the difference.

6.1. Processing time

Table 2 shows the CPU time (expressed as real-time factor, ×RT) employed in six separate operations: (1) feature extraction for segmentation; (2) audio segmentation; (3) acoustic change detection; (4) feature extraction for clustering; (5) computation of sufficient statistics; and (6) hierarchical clustering of speech segments, for both the reference speaker segmentation and the automatic segmentation. Note that the CPU time employed in clustering is almost four times higher for the automatic segmentation than for the reference segmentation, because of the different number of speech segments: 7.24 and 3.62 segments/minute, respectively. The total CPU time of the speaker diarization system is 0.2932×RT.

Table 2: CPU time (real-time factor, ×RT) employed by the speaker diarization system modules.

	Segmentation	
	Reference	**Automatic**
Features (segmentation)	–	0.0033
Audio segmentation	–	0.0375
Acoustic change detection	–	0.1058
Features (clustering)	0.0026	
Statistics	0.0050	
Clustering	0.038	0.139

Computations were made in two servers. The first one, devoted to acoustic classification and acoustic change detection, was a Dell PowerEdge 1950, equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 4GB of RAM. The second one, devoted to clustering, was a Dell PowerEdge R610, equipped with 2 Xeon 5550 (each featuring 4 cores) at 2.66GHz and 32GB of RAM.

7. Conclusions

This paper describes the speaker diarization system developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country for the Albayzin 2010 Speaker Diarization Evaluation. Though quite simple in its structure, based on a chain of four uncoupled modules: audio segmentation, acoustic change detection, computation of sufficient statistics and hierarchical clustering of speech segments, the proposed system attained competitive results in the evaluation.

Experiments carried out on different segmentations showed: (1) that the lowest error rate that the clustering algorithm could attain for the evaluation set was around 20%; and (2) that over-segmentation introduced by the acoustic change detector was the main source of degradation, because the lack of robustness in the estimation of statistics for short segments. Future work may try to improve the robustness of the clustering algorithm to short segments, or alternatively, to avoid over-segmentation while keeping the detection rate of speaker boundaries.

Though not analysed in this paper, we developed an extended version of the clustering algorithm that performed speaker diarization *simultaneously* on the whole set of sessions, thus producing a single set of speaker labels. In fact, we only realized that the optimal mapping of speaker labels would be done independently for each session the day before the deadline (October 16th, 2010). The extended algorithm included a refinement stage which grouped together session clusters according to the algorithm described above, applying the same similarity threshold. We found no way of evaluating this approach because a given label corresponded to different speakers in different sessions.

8. References

[1] L. J. Rodriguez, M. Penagarikano, and G. Bordel, *A Simple but Effective Approach to Speaker Tracking in Broadcast News*, vol. LCNS 4478 of *Lecture Notes in Computer Science*, pp. 48–55. Springer Verlag, Berlin Heidelberg: Pattern Recognition and Image Analysis (IbPRIA 2007), Joan Martí, José Miguel Benedí, Ana Maria Mendonça and Joan Serrat (Eds.), 2007.

[2] M. Penagarikano, A. Varona, M. Diez., L. J. Rodriguez-Fuentes, and G. Bordel, "University of the Basque Country System for NIST 2010 Speaker Recognition Evaluation," in *Proceedings of the II Iberian SLTech Workshop*, (Vigo, Spain), November 2010.

[3] M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, (San Juan, Puerto Rico), pp. 386–391, December 2005.

[4] M. Penagarikano and G. Bordel, "Layered Markov Models: A New Architectural Approach to Automatic Speech Recognition," in *Proceedings of the MLSP Workshop*, (São Luís,Brasil), pp. 305–314, October 2004.

[5] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, (Lansdowne, Virginia, USA), February 8-11, 1998.

[6] X. Anguera, J. Hernando, and J. Anguita, "XBIC: nueva medida para segmentación de locutor hacia el indexado automático de la señal de voz," in *Actas de las Terceras Jornadas en Tecnología del Habla*, (Valencia, España), pp. 237–242, 17-19 de noviembre de 2004.

[7] L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, A. Varona, and M. Diez, "KALAKA: A TV Broadcast Speech Database for the Evaluation of Language Recognition Systems," in *7th International Conference on Language Resources and Evaluation*, (Valleta, Malta), 17-23 May 2010.

[8] A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.

The UVigo-GTM Speaker Diarization System for the Albayzin'10 Evaluation

Laura Docio-Fernandez, Paula Lopez-Otero, Carmen Garcia-Mateo

Department of Signal Theory and Communications, Universidade de Vigo
ldocio@gts.uvigo.es, plopez@gts.uvigo.es, carmen@gts.uvigo.es

Abstract

In this paper, the system submitted by the UVigo-GTM for the Albayzin 2010 Speaker Diarization Evaluation is described. This system is built upon the speaker segmentation system presented in our ICASSP 2010 paper. Specifically, the system uses a poisson-based false alarm reduction strategy. Then, the speaker segmentation strategy assumes that the occurrence of changes constitute a Poisson process, so changes will be discarded with a probability that follows a Poisson cumulative density function. The speaker clustering step we use an agglomerative clustering approach in which the speech segments are merged until reaching a stopping point.

Index Terms: speaker segmentation, speaker clustering, cluto

1. Introduction

Nowadays, an emerging application area where speech technologies are involved is the field of structuring the information of multimedia (audio-visual) documents. These multimedia documents are, in general, multi-speaker audio recordings, and for some applications it may be relevant to determine "who spoke when". This task is also referred to as "speaker segmentation and clustering" or "speaker diarization" in the literature. The segmentation of the data in terms of speakers could help in efficient navigation through audio documents, such as meeting recordings or broadcast news archives. Using these segmentation clues, an interested user can directly access a particular segment of the speech spoken by a particular speaker. Other applications of the speaker segmentation task include speaker adaptation in speech recognition and speaker identification-verification-tracking.

The Albayzin 2010 Speaker Diarization Evaluation task focuses in audio broadcast news programs. The UVigo-GTM speaker diarization system follows a two-stage speaker diarization approach: a speaker segmentation stage, which detects speaker change points; and a speaker clustering stage, where the speech segments, each spoken by one speaker, are clustered using an agglomerative hierarchical strategy.

In [1], an online four-step speaker segmentation system is introduced: it first performs a coarse segmentation of the data, then refines or discards the change points, discriminates between speech and non-speech, and merges segments that are likely to be spoken by the same speaker. It was noticed that this baseline segmentation system has a high false alarm rate and tends to estimate short segments. In [2], two novel approaches for reducing the number of false alarms, in order to avoid erroneous speaker changes, were introduced, evaluated and compared with the false-alarm discard algorithm proposed in [1]. The first approach rejects, given a discard probability, those changes that are suspicious of being false alarms because of their low ΔBIC value. The second strategy assumes that the occurrence of changes constitute a Poisson process, so changes will be discarded with a probability that follows a Poisson cumulative density function. The goal of such techniques is to confirm true speaker changes and suppress erroneous speaker changes. The UVigo-GTM speaker diarization system submitted for the Albayzin 2010 Evaluation is based on the second startegy for rejecting change-points.

To accomplish the clustering task, an agglomerative hierarchical clustering method was chosen. The clustering algorithm measures the similarity between clusters based on the similarity between pairs of speech segments. The critical elements of this clustering technique are the distance or similarity metric used to compare the speech segments, and the selection of the final number of clusters.

This paper is organized as follows. Section 2 gives a brief description of the baseline speaker segmentation system. The proposed approaches to reduce the false alarm rate are presented in Section 3. In Section 4 an explanation of the experimental framework is given. The performance of the speaker segmentation system using each one of the false alarm reduction strategies is shown and discussed in Section 5. Finally, Section 6 concludes this paper and gives some ideas of future work.

2. The speaker segmentation stage

The architecture of the baseline speaker segmentation system described in [1] is depicted in Fig. 1, where it can be observed that it has four stages: first, a coarse segmentation is made with the Distance Changing Trend Segmentation algorithm (DCTS) [3], in order to detect audio change-point candidates and then a refinement or rejection of these change-point candidates is performed by the Bayesian Information Criterion (BIC) algorithm [4]. After that, the system makes a Maximum a Posteriori (MAP) adaptation of three different Gaussian Mixture Models (GMMs) to decide whether the audio segment delimited by the new change-point and the preceding one is speech, music or silence/noise. If the segment is speech, the same procedure will be employed to classify the speech in male or female speech. Finally, when the two latest segments are speech, an approach based on the Cross Likelihood Ratio (CLR) [5] test is applied in order to find out if both speech segments are spoken by the same speaker; in that case both speech segments are merged.

2.1. Poisson distributed-based false alarm rejection strategy

The proposed strategy to discard false alarms is based on the monitoring of the ΔBIC value

$$\Delta BIC(i) = L(i) - \lambda P \tag{1}$$

where P is the penalty, corresponding to the number of free parameters of the Gaussian model, and λ is a weight that increases or decreases the influence of the penalty. When λ is a small

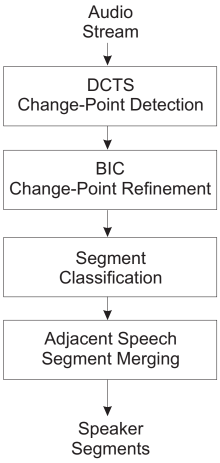


Figure 1: *Architecture of the speaker segmentation system presented in [1].*

value, less changes will be discarded by the BIC algorithm; the opposite happens when λ gets bigger.

Equation (1) has a member $L(i)$ which represents a likelihood ratio:

$$L(i) = \frac{L}{2}log|\Sigma| - \frac{L_1}{2}log|\Sigma_1| - \frac{L_2}{2}log|\Sigma_2| \qquad (2)$$

where L, L_1 and L_2 are the number of frames of segments X, X_1 and X_2 respectively; and Σ, Σ_1 and Σ_2 are the covariance matrices of the models M, M_1 and M_2 respectively. Thus, there will be a change in the audio stream when

$$\frac{L}{2}log|\Sigma| - \frac{L_1}{2}log|\Sigma_1| - \frac{L_2}{2}log|\Sigma_2| > \lambda P \qquad (3)$$

In this false alarm suppression strategy, it is assume that the occurrence times of change-points can be modeled by a Poisson process.

A Poisson process is an independent occurrence process where the number of occurrences in two disjoint time intervals is independent, the probability of having an occurrence is proportional to the observed interval, and the occurrences are not simultaneous.

The process we are dealing with in speaker segmentation fulfills four properties, as it is a process where arrivals (of changes) happen independently from the others and in random instants. Poisson processes have a probability density function

$$f(\mu, x) = \frac{e^{-\mu}\mu^x}{x!} \qquad (4)$$

and its cumulative density function (cdf) is the sum of the probability density function in all the points below a given value:

$$F(\mu, x) = \sum_{i=0}^{x} \frac{e^{-\mu}\mu^x}{x!} \qquad (5)$$

The parameter μ represents the mean of the distribution. In this case, it will represent the number of expected changes.

The properties of the Poisson distribution are going to be used as follows: μ occurrences are expected in a given period of time. Therefore, initially a change will be accepted with a very high probability, but as the number of accepted changes increases and gets close or over the expected number, they will be more likely to be rejected. This is easily modeled by using the cumulative density function $F(\mu, x)$ as a discard probability: this discard probability will be very low at first, and as the mean is approach or exceded, it will get bigger and bigger, until a moment where it will be close to 1 (this means that all the occurrences will be rejected). It can be seen in figure 2 how the discard probability increases as the number of accepted changes gets bigger.

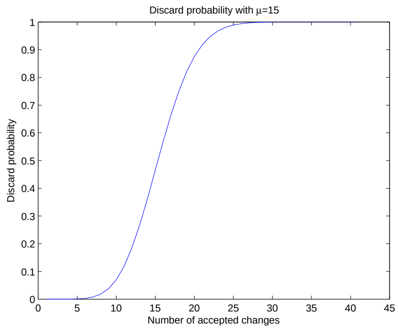


Figure 2: Discard probability on the Poisson-based discard algorithm.

3. The speaker clustering stage

The speaker clustering stage is based on an agglomerative hierarchical clustering technique and standard speaker recognition techniques based on GMM-UBM models are used.

In a training stage, a UBM-GMM model Θ, is constructed using the audio training data. In the clustering stage, first a segment model θ_i is derived by MAP adapting the Θ model parameters using the acoustic frames of the speech segment x_i. Next, a hierarchical classification algorithm is applied in three steps: 1) the first step consists in computing some distance or similarity measure between each pair of speech segments; 2) the second step consists in creating a tree structure by starting with each segment in its own cluster, and recursively merging clusters according to some distance-related criterion; 3) and the last step is to choose one of the partitions, a process called tree cutting.

Several measures of similarity between segments or clusters can be used. The first experiments were conducted using a distance measure which uses information about the likelihood score between pairs of models and speech segments. Specifically, each speech segment, x_i, is scored against all the set of trained segment models, θ_j, and the collection of those scores is used to form a N-dimensional vector X_i (N speech segments) that represents the speech segment x_i in a multidimensional vector-space:

$$X_i = \{ll(x_i|\theta_1), ll(x_i|\theta_2), \cdots, ll(x_i|\theta_N)\}$$

$ll(x_i|\theta_j)$ is the log-likelihood of the ith speech segment x_i given the model of the jth speech segment model θ_j. Thus, the similarity between two speech segments can be computed straightforwardly by using the cosine distance between the two corresponding vectors. This distance ignores absolute sizes of the measurements, and only considers their relative ones; and it is a popular distance measure for comparing documents in the information retrieval literature.

Hierarchical agglomerative methods are well documented in the literature. The aim is to pick the closest pair of clusters according to a distance matrix and merge them. This step is repeated until there is only one cluster. The distance matrix only gives the distance between pairs of single data points, so some method is required to construct a distance between clusters from distances between single data points. There are several possibilities, being most of them variants of the single-link, complete-link, group average-link and minimum variance algorithms. Among these algorithms, the single-link, complete-link and group average-link are the most popular. These algorithms differ in the way they characterize the similarity between a pair of clusters. In the single-link method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters. In the complete-link algorithm, the distance between two clusters is the maximum of all pairwise distances between patterns in the two clusters. In the group average-link approach, the distance between two clusters is the average of all pairwise distances between patterns in the two clusters. Very little is known about what qualities make a cluster distance good for clustering. The general purpose clustering toolkit, CLUTO, developed by the University of Minnesota [7], was used for this unsupervised speaker clustering stage.

In the results submitted to Albayzin 2010 Speaker Diarization Evaluation, the complete-link algorithm was selected, and the stopping criterion was based on a fixed number of clusters, specifically the number of clusters was fixed to 90 for each audio file.

4. Experimental framework

4.1. Database

The training and evaluation database consists of Catalan broadcast news data from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporaci Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, that includes around 87 hours of sound (24 files of approximately 4 hours long), was splitted into two parts: one part for training/development (2/3 of the total amount of data), and the other part for evaluation (the remaining 1/3).

The number of speakers per recording ranges from 30 up to 250. This high number of speakers is due to the domain of the data. Some speakers are common among different recordings. That is the case of the newscaster, the journalists or some voices from the commercials, etc. However, most of the speakers have short duration turns since their presence depends on the news itself.

The 16 available files to perform the training/development of the segmentation system were splitted as follows:

- Sessions 1 to 8 and 10 to 15: training of the silence, speech, music and GMM-UBM models.
- Sessions 9 and 16: selection of the parameters that achieve the best performance. The parameters to select were μ, λ, M.

After testing on the development data the selected parameters were: $\mu = 15.0$, $\lambda = 2.5$, $M = 64$.

4.2. Metric

Diarization Error Rate (DER) as defined by NIST in Rich Transcription evaluations [8] will be used to assess the submitted systems. In order to measure the performance, an optimum one-to-one mapping of reference speaker IDs to system output speaker IDs is computed. The measure of optimality will be the aggregation, over all reference speakers, of the time that is jointly attributed to both the reference speaker and the (corresponding) system output speaker to which that reference speaker is mapped. This mapping over will always be computed over all the speech, including regions of overlap . Mapping is subject to the following restrictions:

- Each reference speaker will map to at most one system output speaker, and each system output speaker will map to at most one reference speaker.
- Mapping of speakers will be computed separately for each speech data file.

Speaker detection performance will be expressed in terms of the miss and false alarm rates that result from the mapping. An overall time-based speaker diarization error score will be computed as the fraction of speaker time that is not attributed correctly to a speaker.

4.3. Acoustic features

The audio signal is characterized by 12 mel-frequency cepstral coefficients (MFCC) extracted every 10 ms using 25 ms Hamming windows. Then these cepstral features are augmented by the log-energy. The DCTS and BIC change detection stages use only the 12 MFCCs and the log-energy as features. In the speech/non-speech classification and the gender classification modules the first and second derivatives of this feature vector are also considered.

The speech, non-speech, male and female and GMM-UBM models are 64 diagonal Gaussian Mixture Models (GMM) directly trained on data extracted from the train corpus by using the Expectation-Maximization (EM) algorithm.

5. Experimental results

Table 1 provides the results obtained by the system submitted by the UVigo-GTM research group. In column 2 of Table 1 it can be observed a lack of consistency in the speaker diarization error between different speech (session) files. This fact suggest that a serious mistake was made somewhere when processing these evaluation files with the speaker diarization system. After checking the code and the file processing the mistake was corrected and the results obtained are those shown in column 3 of Table 1. These results are still far from those obtained by the other participants.

A manual inspection of the number of speakers in each evaluation audio file shows that selecting 90 as the number of clusters is not the best option for all the audio files. The wrong choice of number of clusters affects the system performance adversely. Table 2 shows the number of speakers in each audio file. Selecting the right number of clusters has not been considered in the submitted system. The approach used a fixed number of clusters for all the audio files.

Speaker Diarization experiments were conducted using the “Unweighted Pair-Groups Method Average (UPGMA)” criterion function for agglomerative clustering, which defines cluster similarity in terms of the average pairwise similarity between the segments in the two clusters. This criterion is widely used

Table 1: *Speaker diarization results on the evaluation corpus: CLINK agglomerative criterion function and 90 clusters per audio file.*

audio file	DER	
session17	40.49	40.49
session18	68.33	42.99
session19	40.11	40.11
session20	69.70	45.36
session21	42.80	42.80
session22	88.23	39.61
session23	70.57	34.87
session24	36.96	36.96
global	58.03	40.21

Table 2: *Number of speakers in each evaluation audio file.*

s17	s18	s19	s20	s21	s22	s23	s24
106	91	70	120	65	93	66	93

in text document clustering because it is more robust than other traditional agglomerative clustering approaches. The resulting SDER are shown in Table 3. Compared with the results on Table 1, the use of UPGMA gives significant improvement over the submitted results. Table 3 also shows the influence played by the number of clusters in the speaker diarization error.

Table 3: *Speaker diarization results on the evaluation corpus: UPGMA agglomerative criterion function.*

	DER						
	70	80	90	95	100	105	110
session17	43.44	42.54	37.98	37.63	37.61	**35.43**	37.20
session18	38.46	36.67	35.69	34.41	34.66	34.21	**34.01**
session19	**29.87**	32.57	32.84	32.22	32.89	33.15	33.95
session20	38.58	38.12	38.73	37.69	37.60	36.94	**34.68**
session21	**32.01**	32.59	31.91	32.09	32.95	33.73	35.27
session22	44.91	40.98	39.68	39.18	38.80	38.26	**38.07**
session23	**27.58**	28.89	28.71	28.57	28.62	28.83	29.30
session24	32.66	33.60	34.32	34.20	**32.75**	33.29	32.88
global	36.09	35.87	35.11	34.60	34.54	34.23	34.34

6. Conclusions and future directions

The speaker segmentation system submitted to Albayzin 2010 Evaluation was described in this paper. The speaker diarization task focuses in the context of broadcast news. According to the results obtained by the proposed system on the evaluation dat, it was realized that a huge mistake was made when processing some evaluation speech files.

Future work will focus on combining the traditional short-term MFCCs features with prosodic and other acoustic features in order to discriminate better between speakers. Related to the speaker clustering stage, future work will focus on:

- The analysis of methods for speaker clustering used in the state-of-art speaker diarization systems.
- The analysis of strategies to detect or to discover the number of clusters, i.e, approaches for cluster stopping.
- The use of other similarity measures between speech segments and other criteria to group speech segments.

7. Acknowledgements

The UVigo-GTM group would like to thank the organizers and coordinators of the Albayzin 2010 Speaker Diarization Evaluation for their help and for kindly providing the data corpus. The work reported here was partially sponsored by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04 “Búsqueda de Información en Contenidos Audiovisuales plurilingües” and by the Ministry of Science and Innovation under the FPI grant BES-2010-033358.

8. References

[1] L. Docio-Fernandez, P. Lopez-Otero, and C. Garcia-Mateo, “An adaptive threshold computation for unsupervised speaker segmentation,” *Proc. 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, September 2009, pp. 849–843.

[2] P. Lopez-Otero, L. Docio-Fernandez and C. Garcia-Mateo, “Novel strategies for reducing the false alarm rate in a speaker segmentation system,” *Proc. 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2010*, Dallas, USA, March 2010, pp. 1470–1473.

[3] Y. Wang, J. Han, H. Li, and T. Zheng, “A novel audio segmentation method based on changing trend of distance between audio scenes,” *Journal of Communication and Computer*, vol. 3, pp. 22–30, July 2006.

[4] Scott S. Chen and P. S. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, February 1998.

[5] D. A. Reynolds, E. Singer, B. A. Carlson, and J. J. McLaughlin, “Blind clustering of speech utterances based on speaker and language characteristics,” *in Proc. of the Int. Conf. on Spoken Language Processing*, Sydney, Australia, 1998.

[6] A. O. Allen, “Probability, Statistics, and Queueing Theory with Computer Science Applications”, 2nd ed. Academic Press, 1990.

[7] “CLUTO - Software for Clustering High-Dimensional Datasets”, http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview.

[8] “NIST Rich Transcription Evaluation”, http://nist.gov/itl/iad/mig/rt.cfm.

VIVOLAB-UZ Speaker Diarization System for the Albayzin 2010 Evaluation Campaign

Carlos Vaquero, Alfonso Ortega, Eduardo Lleida

Communications Technology Group (GTC)
Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{cvaquero,ortega,lleida}@unizar.es

Abstract

This paper describes the speaker diarization systems proposed by the VIVOLAB-UZ group for the Albayzin 2010 speaker diarization evaluation. Our approaches combine recent improvements in the field of speaker segmentation in two speaker telephone conversations, using eigenvoice modeling, with the traditional Agglomerative Hierarchical Clustering approach. We are presenting two submissions. Our first system uses a simple eigenvoice factor analysis model to extract a stream of speaker factors for every recording that enable better speaker separability. The speaker factor stream is used for speaker segmentation. Then, both the clusters obtained are agglomerated using Bayesian Information Criterion as distance metric, obtaining the speaker labels. Our second submission is exactly the same system, but it uses Viterbi resegmentation to refine speaker change points as a final step.

Index Terms: Speaker diarization, Factor Analysis, intrasession variability, Agglomerative Hierarchical Clustering, Bayesian Information Criterion

1. Introduction

The main breakthroughs in the field of speaker diarization have been introduced this decade, in part due to the NIST Rich Transcriptions (RT) evaluations. From 2002, NIST has coordinated several Rich Transcription Evaluations aiming at extracting information from audio recordings such as speaker turns or speech transcriptions. All these evaluations involved a Speaker Diarization task, that has become the framework for research and development of the state of the art speaker diarization technologies. In the beginning, the environment to evaluate speaker diarization approaches in the RT framework were telephone conversations and broadcast news. From 2005 the evaluation has focused on meetings. The current Albayzin evaluation does not differ much from those RT evaluations on broadcast news.

Most of the best performing systems presented in the RT evaluations are based on Agglomerative Hierarchical Clustering, that is, after a first segmentation, that gives a set of clusters, the system performs a bottom-up clustering until a stopping criterion is met [1]. Usually, Viterbi resegmentations are performed every time two clusters are merged, and several criteria are used for cluster merging and as stopping criterion. One of the most widespread is the Bayesian Information Criterion (BIC), that has shown to perform well for both cluster merging and stopping criterion.

On the other hand, recently, there has been a great advance in the field of speaker identification, in part motivated by the NIST Speaker Recognition Evaluations (SRE). One of the main breakthroughs of the last years has been the formulation of the Joint Factor Analysis (JFA) for speaker verification [2]. This has motivated the application of this new technique to different areas, mainly to the task of speaker segmentation in two speaker conversations. Some approaches for two speaker segmentation that make use of JFA are presented in [3], [4], [5].

VIVOLAB-UZ is submitting two systems, both based on a combination of a JFA based speaker segmentation system and a BIC based AHC system. The only difference is that the first system obtains speaker labels directly from the BIC AHC step, while the second uses these labels to perform a final Viterbi resegmentation.

This work has been partially funded by the national project TIN2008-06856-C05-04.

2. System Description

Our speaker diarization systems fuse a JFA based speaker segmentation system and a BIC based AHC system. Currently our speaker segmentation system works with a given number of speakers (it was designed for 2-speaker conversations), so firstly, after running a speech activity detector (SAD), we split every recording into 5 minute slices and every slice is processed with the speaker segmentation system separately. We force the speaker segmentation system to find 10 speakers in every slice. Once we have 10 clusters for every 5-minute slice, we perform a BIC AHC algorithm over the whole recording to merge those clusters belonging to the same speakers until a stopping criterion is met. This way we obtain th output for the first system. Our second system will use directly this output to perform a Viterbi resegmentation. The different steps of the VIVOLAB-UZ speaker diarization system are described in the following subsections.

2.1. Features

The features used for all steps in the diarization system are 18 MFCC including c0, computed every 10 ms over a 25 ms window. No normalization is used on the features.

2.2. Speech Activity Detector

The first step in most speaker diarization systems is to separate the speech segments from those segments that does not contain speech. In our case, non-speech segments may contain music, noise or silence. To obtain the speech segments we train 64 gaussian GMMs for two classes using the development data. One GMM for speech and the other one for non-speech. We run a Viterbi segmentation, modeling every class with 10 tied-states [6] that share the same GMM as the observation distribution.

2.3. Segmentation System

In the proposed speaker segmentation system, described in [5], we use a factor analysis approach to model the desired sources of variability. As a starting point we try to capture the variability present among different speakers. For this purpose, we model every speaker by a Gaussian Mixture Model (GMM) adapted from a Universal Background Model (UBM) using an eigenvoice approach [7], according to:

$$M_s = M_{UBM} + Vy. \tag{1}$$

Where M_s is the speaker GMM supervector, obtained concatenating all Gaussian means, M_{UBM} is the UBM supervector, V is the low rank eigenvoice matrix, and y is the set of speaker factors, which follows a standard normal distribution $N(y|0, I)$ a priori. This way every speaker is represented by a GMM supervector in a high dimension space, and in such space we allow the speakers to lay in the low dimension subspace generated by the column vectors of V, which point to the directions of maximum variability among speakers. We refer to this variability as inter-speaker variability and to the low rank subspace as the speaker subspace.

In our approach we use a 256 Gaussian UBM. The dimension of the speaker subspace is 20, compared to the dimension of the supervector space that is $256 \times 18 = 4608$. This way every point estimate for a given speaker is defined by a set of 20 speaker factors.

Once we have the speaker factors we apply Within Class Covariance Normalization (WCCN) to compensate intra-session variability and to ensure that the variance of the speaker factors is close to I for every speaker as in [8]. WCCN is a normalization method that allows to obtain a linear transformation for a given set of features belonging to different classes so that the within class covariance matrix S_w defined in Eq. 2 is equal to the identity matrix I. This technique assumes that all classes have the same covariance matrix.

$$S_w = \frac{1}{S-1}\sum_{s=1}^{S}\frac{1}{N^s-1}\sum_{n=1}^{N^s}(y_n^s - \mu^s)(y_n^s - \mu^s)^T \tag{2}$$

$$\mu^s = \frac{1}{N^s}\sum_{s=1}^{N^s} y_n^s \tag{3}$$

To obtain the linear transformation we first obtain S_w as shown in Eq. 2 and then we apply Cholesky decomposition, so the transformed speaker factors y' will follow this expression:

$$y' = Ry \tag{4}$$

$$S_w^{-1} = R'R \tag{5}$$

where R is the upper triangular matrix obtained by Cholesky decomposition.

To perform speaker segmentation given a sequence of feature vectors, we estimate the speaker factors for every frame over a 100 frame window, with an overlap of 990 ms, we transform the speaker factors using WCCN, and we estimate a 10-Gaussian GMM to model the stream of speaker factors obtained, after removing non-speech frames according to the SAD. Each one of these Gaussians will be assigned to a single speaker. We perform this process over 5-minute slices, obtaining 10 new speakers for every slice.

2.3.1. Initialization

We have detected that a good initialization is quite important to ensure that every Gaussian in the GMM corresponds to a single speaker. In our approach, we use prior knowledge about speaker factors proposed in [2]: A priori, speaker factors are assumed to be distributed according to the standard normal distribution $N(y|0, I)$. Since we obtain speaker factors from a small data sample, using MAP estimation, we can expect the posterior distribution of speaker factors for a single speaker to keep some properties of the prior. In addition, since we perform WCCN on the speaker factors, we will be closer to fulfill this assumption. Assuming that the posterior variance is close to I, we can perform PCA to obtain those directions of maximum variability in the speaker factor space. Then we will use 9 ($N_{spks} - 1$) directions to obtain, using K-means, a first clustering.

This strategy gives ten clusters that can be seen as a first speaker segmentation, and then K-means clustering is performed over the 20 speaker factors to reassign frames to the ten clusters and a single Gaussian is trained on each of them.

2.3.2. Core Segmentation

The 10 Gaussians previously trained serve as initial GMM of the whole recording. Then a two stage iterative process is applied until convergence: first several Expectation-Maximization (EM) iterations are used and then, every Gaussian is assigned to a single speaker and a Viterbi segmentation is performed. According to this new frame assignment, 10 Gaussian models are trained and the iterative process restarts again. Convergence is reached when the segmentation of the current iteration is identical to that obtained in the previous one.

To avoid fast speaker changes, in the Viterbi segmentation, we modify the speaker turn duration distribution using a sequence of tied-states [6] for every speaker model. This way, we avoid the state duration to follow a geometric distribution that cannot accurately model real speaker turn durations. Each speaker model is composed of 10 states that share the same observation distribution, a single Gaussian in this case. Tied-states are not considered for the silence, but a single state without an observation distribution is used, since the algorithm is forced to go through the silence state according to the SAD labels. We have observed that this way of modeling speaker turn duration yields better results than modifying the transition probability.

2.4. Clustering

Once we have a set of ten clusters for every 5-minute-slice, an AHC step is performed to obtain the final clusters that correspond to the actual speakers. For this purpose, BIC is considered both as clustering metric and as stopping criterion. Every cluster is modeled using a single full covariance gaussian using MFCC, and two hypotheses are considered for every pair of clusters: The null hypothesis, that is, assuming that both clusters belong to the same speaker, and the 2 speaker hypothesis, that is, assuming that every cluster belong to a different speaker. BIC is computed for both hypotheses, and the ΔBIC is computed as $\Delta BIC = BIC_{2spks} - BIC_{null}$. The pair of clusters having lower ΔBIC is merged. Clusters are not longer merged when $\Delta BIC > 0$. To penalize the 2 speaker hypothesis the λ parameter for model complexity penalization is set to 10.0.

This step will give the output of the first VIVOLAB-UZ submission.

2.5. Resegmentation

Only the second system performs a final Viterbi Resegmentation. For this purpose, we model every speaker with a 32 component GMM using MFCC, according to the output of the BIC AHC step. As in the core segmentatino system, we use 10 tied-states for speaker models and a single state for all silence frames.

3. Development data

As development data we have considered those recordings provided for this purpose in the evaluation (16 sessions from the Catalan Broadcast News database) and the English and Mexican Broadcast News Speech from the Hub4 database.

For training the GMMs used in the SAD the 12 first sessions from the provided development data were used. The remaining 4 sessions were used to check the SAD and to adjust the AHC parameters (λ).

For training the UBM, the Eigenvoice matrix V and the WCCN transformation the Hub4 database was considered. Performance will be degraded for using different languages and dialects to train the models to obtain the speaker factors, but we could not find any other labeled datasets in Catalan.

4. Computational cost

The proposed system runs in matlab and it is not optimized. The following table resumes the computational cost for every step of the proposed diarization system:

Operation	Computational cost, real time (rt)
Obtain speaker factors	$0.20 \times rt$
SAD	$0.02 \times rt$
Segmentation	$1.10 \times rt$
BIC AHC	$< 0.01 \times rt$
Viterbi resegmentation	$\approx 1.20 \times rt$

Table 1: *Computational cost of the speaker diarization system, step by step.*

5. Conclusions and Future Work

This is the first time that VIVOLAB-UZ group participates in a Speaker Diarization Evaluation. We have built a system combining two state-of-the-art technologies: JFA and eigenvoice modeling for speaker segmentation and BIC based AHC for speaker clustering. We believe in the potential of the eigenvoice modeling for the task of speaker segmentation, but due to the lack of training data, we do not expect this submission to obtain state-of-the-art performance.

As future work we plan to study the potential of the eigenvoice modeling using matched data to train the V matrix, and we plan to improve the speaker clustering using GMM to model every cluster instead of single gaussians, and using a bigger GMM to model the null hypothesis, canceling the complexity penalization term in the BIC computation. This should give better agglomeration and a more robust stopping criterion.

Finally we plan to introduce Bayesian approaches to determine the number of speakers. We are working in the development of a speaker diarization system that combines eigenvoice modeling for speaker segmentation and Variational Bayes for determining the number of speakers in the recording. We believe this approach can obtain competitive performance.

6. Acknowledgements

We would like to thank the organizers of Albayzin 2010 Speaker Diarization Evaluation for their effort in preparing the evaluation data, and also to all the organization of FALA 2010 for supporting the Albayzin Speaker Diarization Evaluation.

7. References

[1] Wooters, C., Huijbregts, M., “The ICSI RT07s Speaker Diarization System”, in Proc Machine Learning for Multimodal Interaction, LCNS, Springer, Heidelberg, 2007.

[2] P. Kenny et al, “A Study of Inter-Speaker Variability in Speaker Verification”, IEEE Trans. Audio, Speech Proc., 2008

[3] Castaldo, F. et al, “Stream Based Speaker Segmentation Using Speaker Factors and Eigenvoices”, in Proc ICASSP, 4133-4136, Las Vegas, NV, 2008.

[4] Reynolds, D. et al “A Study of New Approaches to Speaker Diarization”, in Proc Interspeech, 1047–1050, Brighton, UK, 2009

[5] Vaquero, C. et al “Confidence Measures for Speaker Segmentation and their Relation to Speaker Verification”, in Proc Interspeech, Makuhari, Japan, 2010.

[6] Levinson, S.E., “Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition”, Computer Speech and Language, I:29–45, 1986.

[7] R. Kuhn et al. “Rapid Speaker Adaptation in Eigenvoice Space”, IEEE Trans. Speech Audio Proc. Vol 8, no. 6, 695–707, 2000.

[8] Vaquero, C., Ortega, A., Lleida, E. “Intra-session Variability Compensation for Speaker Segmentation”, to appear in FALA, Vigo, Spain, 2010.

Speaker Diarization Using Gaussian Mixture Turns and Segment Matching

Arlindo Veiga[1,2], *Carla Lopes*[1,2], *Fernando Perdigão*[1,2]

[1] Department of Electrical and Computer Engineering, University of Coimbra
[2] Instituto de Telecomunicações – Pólo de Coimbra
{aveiga, calopes, fp}@co.it.pt

Abstract

Speaker diarization aims to detect "who spoke when" in large audio segments. It is an important task in processing of broadcast news audio, making easier the audio segments selection and indexing task. In this paper an unsupervised speaker diarization scheme is proposed using a Gaussian Mixture Model as a Universal Background Model, Bayesian Information Criterion and fingerprint detection. A decoder that outputs a mixture sequence is used with high mixture transition penalization. Homogeneous segments tend to produce sequences with only one mixture allowing speaker turns to be detected using mixture transitions. Results for the Catalan broadcast news 3/24 TV channel are reported.

Index Terms: speaker diarization, audio fingerprint, GMM, BIC

1. Introduction

Gaussian Mixture Model (GMM) has the ability to model arbitrary densities distributions. They have shown to have excellent performance in speech modelling tasks. Many speaker recognition systems use a large GMM called Universal Background Model (UBM) [1] to represent all speaker distribution features and to adapt it to speaker's models. The Expectation Maximization (EM) is a well-established algorithm that is used to estimate the GMM parameters.

Since the speaker diarization contest relies on the assumption that do not exist prior knowledge about the speakers, the use of speaker models is not allowed. However UBM can still be used for unsupervised speaker segmentation and classification considering each mixture as a cluster. A decoder that outputs a mixture sequence was used with a high mixture transition penalization. Homogeneous segments should result in a single mixture sequence in most of the time. The ALIZE/Mistral project [2] proposes a system for speaker diarization that use a similar decoder for speaker clustering. Furthermore, the so called Bayesian Information Criterion (BIC) [3] is used in most proposed approach for unsupervised speaker clustering because it requires only one training parameter estimation.

Before speaker clustering, and in order to deal with non-speech audio segments, a segmentation system is used to detect silence, and music. The fingerprint system [4] is used to ensure that similar segments have same label. Acoustic or audio fingerprinting refers to a condensed representation of an audio signal that can be used to identify an audio sample or quickly locate similar items in audio streams. A binary representation of spectral patterns computed by the convolution of spectrogram with a mask is used. With this technique we can quickly and easily discover repeated segments with high confidence. To deal with classification inconsistencies in repeated segments, we use majority vote to determine which label should be used. In speaker diarization we use also this technique to set the same speaker to similar audio segments.

Finally, the BIC algorithm is used to cluster homogeneous segments. This step performs a substantial increase in the systems' performance.

2. System description

An audio segmentation system is used to segment audio files into five classes: clean speech, music, speech with music in background, speech with noise in background and others non-speech events (silence and noises). Non-speech segments are discarded and all speech segments are labelled with different speaker identification, creating one cluster per speech segment. To merge similar clusters, a UBM with 256 components is trained with the entire audio session. Then, a model based clustering method using this model is applied by means of a mixture sequence decoding. This will be explained in more detail in section 4.

. The result of the decoding is a sequence of segments with speaker labels. The fingerprinting system is then used to search for label inconsistencies in similar segments. Each segment is used as a "jingle" to verify if there is another equal segment in the audio session. In that case the speaker labels are made equal.

A final verification is applied to the labelled segments based on BIC. A pair-wise BIC difference (ΔBIC) between each cluster is computed and the decision to merge clusters is based on this value.

Figure 1 describes the speaker diarization system.

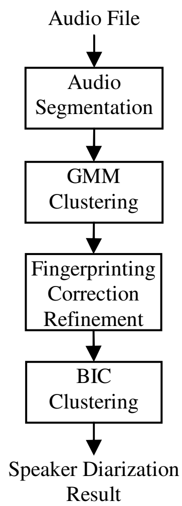


Figure 1: *Speaker diarization block diagram.*

The audio segmentation system used here is the same presented on audio segmentation contest.

3. Features

Audio features are computed every 100ms within a 200ms Hamming window, resulting on a frame-rate of 10 frames per second. The dimension of the feature vector is 48, corresponding to 24 static parameters and their first order derivatives. The static parameters consists of 16 Mel frequency cepstral coefficients with more 8 parameters shown in Table1.

Table 1. *Acoustic Feature Set used in combination with standard MFCC Features.*

Number	Feature description
1	Frame Energy in dB
2	Zero Crossing Rate
3	Spectral Centroid
4	Spectral roll-off - 90%
5	Max normalized correlation coefficient in each frame
6	frequency of the max normalized correlation coefficient
7	Harmonicity measure (proportional to the harmonic duration)
8	Spectral Flux

Audio segmentation and UBM training uses all the 48 features of the acoustic vector while BIC clustering uses only MFCC features (16). In most papers 12 MFCC have proved to be very efficient for speaker turns task. In the [5] the authors report that the use of first order derivative coefficients deteriorates the system performance. The BIC value depends of dimensions of feature vector.

Audio fingerprinting system use a binary representation of spectral patterns and is computed every 20ms within 240ms Hamming window resulting 50 frames per second. Binary representation of each frame is saved on 32 bits integer and searching similar segments is based on *xor* operations and bit counts.

4. GMM Clustering

HTK [6] tools were used to initialize and train each session's background model. The number of mixture used was 256 because the number of speakers per recording ranges from 30 up to 250. It is expected to have one dominant mixture per speaker.

A simple decoder was used in order to find the best mixture sequence, given the audio features. It corresponds to a Viterbi decoder with 256 states in parallel, each one using a single mixture Gaussian of the UBM and a high transition penalization between different states. Homogeneous segments uttered by the same speaker tend to have the same mixture Gaussian during most of the segment time. So, given two segments, it is possible to estimate if they are uttered by the same speaker analyzing the decoding results of a concatenation of these two segments. Figure 2, illustrate a pair of segments and the mixture sequence. The algorithm computes the mean mixture index in the first (fixed) segment and in all other forward segments. If the difference between them means is low, the second segment becomes with the label of the first one.

The resultant mixture sequence is highly dependent on the transition penalty used in the decoding process. Using just a simple mean criterion implies the use of a high mixture transition penalty (in this case a value of -100 was used as a logarithmic weight).

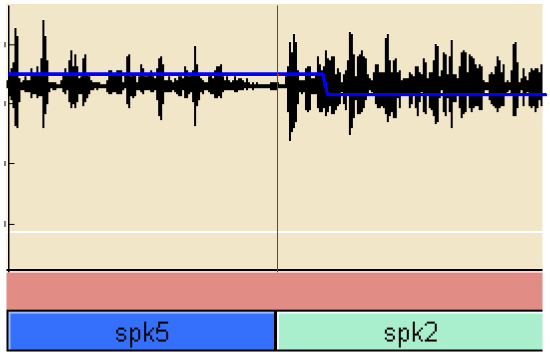


Figure 2: *Two concatenated segments and mixture sequence.*

It is possible to use different criteria to analyze the mixture sequence in each segment in order to decide if they are uttered by the same speaker. In future work a more robust criterion will be searched for.

The last step consists in finding similar segments using the fingerprinting technique. If two segments are found to be repeated, then the same label is set for both segments. This corrects inconsistencies on labelled segments.

5. BIC Clustering

To reduce the GMM clustering diarization error rate (DER), another clustering algorithm is used, based on BIC [3] computation. Using speaker turn detention approach, ΔBIC requires tuning only one parameter (λ). Considering two segments, X_1 and X_2, each one represented by a single Gaussian, $X_1 \sim N(x;\boldsymbol{\mu}_{X1},\boldsymbol{\Sigma}_{X1})$ and $X_2 \sim N(x;\boldsymbol{\mu}_{X2},\boldsymbol{\Sigma}_{X2})$, a concatenated segment X can still be represented by a single Gaussian $X \sim N(x;\boldsymbol{\mu}_X,\boldsymbol{\Sigma}_X)$ if X_1 and X_2 are similar segments (uttered by the same speaker). This results in a positive ΔBIC value, defined as [3,5]:

$$\Delta BIC = -R + \lambda P \tag{1}$$

where:

$$R = \tfrac{N_X}{2}\log\left(\left|\Sigma_X\right|\right) - \tfrac{N_{X1}}{2}\log\left(\left|\Sigma_{X1}\right|\right) - \tfrac{N_{X2}}{2}\log\left(\left|\Sigma_{X2}\right|\right), \tag{2}$$

$$P = \frac{1}{2}\left(p + \frac{1}{2}p(p+1)\right) \tag{3}$$

λ is an empiric parameter normally with a value between 0.5 and 2. In this work 0.6 was used after tuning with the some reference sessions.

6. Results

The DER on the test database of the contest is 55.84 %. The time to diarize the test sessions was around 4 hours, about half an hour per session. Most of time is spent in the Viterbi algorithm. The code was mainly implemented in Matlab (with optimizations in C++ mex) and the used machine was a Cray CX1.

7. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, vol. 10, no. 1, pp. 19-41, Jan. 2000.

[2] J.-F. Bonastre, F. Wils, and S. Meignier, “ALIZE, a Free Toolkit for Speaker Recognition,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2005), pp. 737-740, 2005.

[3] S. S. Chen and P. S. Gopalakrishnan, “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion,” in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 127-132, 1998.

[4] C. Neves, A. Veiga, L. Sá, and F. Perdigão, “Audio Fingerprinting System for Broadcast Streams,” in Proceedings of the Conference on Telecommunications (ConfTele2009), pp. 481-484, 2009.

[5] P. Delacourt and C. J. Wellekens, “DISTBIC: A Speaker-based Segmentation for Audio Data Indexing,” Speech Communication, vol. 32, no. 1, pp. 111-126, Sep. 2000.

[6] S. Young et al., The HTK Book (for HTK Version 3.4). Cambridge, U. K: Cambridge University Engineering Department, 2006.

Albayzin Evaluation: Audio Segmentation System at CEPHIS-UAB

Héctor Delgado, Javier Serrano, Jordi Carrabina

Software-Hardware Prototypes and Solutions Lab, Autonomous University of Barcelona, Spain
hector.delgado@uab.cat, javier.serrano@uab.cat, jordi.carrabina@uab.cat

Abstract

This paper describes the audio segmentation system developed at the Software-Hardware Prototypes and Solutions Lab (Autonomous University of Barcelona) for the Albayzin 2010 Evaluations at the FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop.

1. Introduction

Audio segmentation problem consists of dividing audio streams into acoustically homogeneous segments. It is usually applied to a series of applications, such as indexing and retrieval, as a previous step to improve ASR accuracy by means of adaptation techniques, and so on. Before segmenting, a categorization of the acoustic classes must be done, depending on the particular domain where it will be applied. Typically, audio streams are segmented into silence, music, background noise, clean speech and speech corrupted with some kind of noise (music or background music).

The paper is structured as follows: section 2 describes the system, particularly the training and test data, feature extraction and acoustic classes modeling. Experimental results are given in section 3. Finally, some conclusion are extracted in section 4

2. System overview

The current section describes the audio segmentation system set-up in detail, particularly the training and test data, the feature extraction and the acoustic classes modeling and configuration.

2.1. Training data

It consists of a Catalan broadcast news database from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporació Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, that includes around 87 hours of sound (24 files of approximately 4 hours long), will be split into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). The distribution of classes within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3%. The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

2.2. Feature extraction configuration

PLP (further explanation in [1]) method has been chosen for the current task. PLP features have been empirically proven to be beneficial for audio segmentation tasks [2]. Firstly, a speech signal processing is made. A 0.97 coefficient pre-emphasis filter is applied, and a 25 ms Hamming window that scrolls each 10 ms is used to obtain signal frames. Then, a feature vector of 12 PLP coefficients is obtained from each frame using a 50 channel filter bank. Finally, the energy coefficient, delta and delta-delta features (time derivatives) are added to the feature vectors.

2.3. Acoustic classes modeling

An HMM for each acoustic class is created. Each HMM has three states in a left-to-right topology. Only the central state has a self-transition and a diagonal covariance matrix single-gaussian mixture model as emitting probability density function.

Then, the single-gaussian models are consecutively split into 2, 4, 8, 16, 32 and 64 mixture gaussians, re-estimating the model parameters using the Baum-Welch algorithm (HERest tool in HTK [3]) before doing each split.

Once the model set is obtained, the audio segmentation is performed through the Viterbi algorithm (HVite tool). Part of the training data has been used to tune some parameters in order to find a compromise between accuracy and computation time.

3. Experimental results

3.1. The metric

The proposed metric is inspired on the NIST metric for speaker diarization. The metric is defined as a relative error averaged over all acoustic classes:

$$Error = average_i(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}) \qquad (1)$$

where

- $dur(miss_i)$ is the total duration of deletion errors (misses) for the *i*th acoustic class.
- $dur(fa_i)$ is total duration of all insertion errors (false alarms) for the *ith* acoustic class.
- $dur(ref_i)$ is the total duration of all the *i*th acoustic class instances according to the reference file.

A forgiveness collar of 1 sec (both + and -) will not be scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an acoustic class begins/ends.

Table 1: Segmentation error

	Music	Speech	Speech over music	Speech over noise	Average
Error (%)	23.65	45.07	36.95	45.21	**37.72**

3.2. Results

Experimental results are shown in table 1. A results discussion is given in section 4

3.3. Execution time

The experiment has been carried out in a Intel Core2 Duo 6420, 2.13 GHz CPU, 3 GByte RAM system. The operating system is Linux (Ubuntu 10.04).

The total execution time is shown next (output of the 'time' UNIX command).

```
real    22m1.533s
user    20m26.270s
sys     0m7.870s
```

4. Conclusions

Generally, it can be observed that the resulting error rates are considerably high. The used technique has been successfully applied on speaker segmentation tasks. However, other methods should be explored for audio segmentation purposes.

The system presents a better error when classifying music. On the other hand, the error rate increases considerably when the system tries to distinguish between the different kinds of speech. It indicates that the proposed models and feature extraction techniques are not totally suitable in order to classify speech in under different acoustic conditions. Possible improvements would be the use of other kind of acoustic class-based feature extraction instead of traditional features like MFCC. Other further improvement could be the application of a hierarchical architecture instead of classifying each acoustic class independently.

5. Acknowledgements

This work has been partly founded by the Spanish Ministerio de Ciencia y Innovacion, Project Reference TEC2008-03835/TEC. This article is supported by the Catalan Government Grant Agency Ref. 2009SGR700

6. References

[1] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001.

[2] H. Delgado, "Segmentación de Vídeo mediante Reconocimiento de Locutores," Master's thesis, Universitat Autónoma de Barcelona, June 2009.

[3] S. Young and G. Evermann, *The HTK Book*. Cambridge University Engineering Department, March 2009.

ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation

Javier Franco-Pedroso, Ignacio Lopez-Moreno, Doroteo T. Toledano, and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
{javier.franco, ignacio.lopez, doroteo.torre, joaquin.gonzalez}@uam.es

Abstract

This paper describes the ATVS-UAM systems submitted to the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation. The ATVS-UAM audio segmentation system is based on a 5-GMM-MMI-state HMM model. Testing utterances are aligned with the model by means of the Viterbi algorithm. Spurious changes in the state sequence were removed by mode-filtering step. Finally, too sort segments were removed. The ATVS-UAM speaker diarization system is a novelty approach based on the cosine distance clustering of the Total Variability speech factors -the so-called iVectors- performed in two steps, followed by a Viterbi decodification of the probabilities based on the distances between the candidate speaker centroids and the iVectors stream.

Index Terms: audio segmentation, speaker diarization, viterbi, factor analysis, maximum mutual information.

1. Introduction

In the recent years the speaker and language recognition community dedicates special attention to the *real conditions* challenge. This challenge involves audio recordings preceding from different sources in addition a single speaker, such as noise, channel effects, speech or music. Speaker turns in a conversation also causes significant degradation in performance for poor segmentations. Such challenge motivates the ATVS-UAM participation in Albayzin 2010. Recently, Factor Analysis (FA) methods have shown excellent results facing some of these problems such as the compensation of the channel and speaker variability. Moreover, FA is currently the state-of-the-art technology for speaker and language recognition, with promising results in other fields such as speech recognition. A successfully FA scheme for speaker diarization was firstly proposed by Castalado [1] in 2008 and later extended in [2]. Castaldo uses low dimensional speaker vectors that are obtained over highly overlapped windows of one-second length. Thus FA generalizes as a secondary parameterization of the input speech stream. This new short-term speaker-factors space shows excellent results when classical speaker diarization techniques are applied on it. In [3] Najim and Kenny enhances the classical FA scheme by: a) Modeling together speaker and channel variability, in what is called total variability. Additional improvements can be achieved with a discriminative training of the target classes such as Linear Discriminant Analysis (LDA) [4] and b) Estimating the posterior probabilities of a speaker participating in the conversation as the cosine distance between the averaged iVectors over the training and testing utterances [4].

Other concerns that have been addressed during the design of the ATVS-UAM Audio Segmentation System were the use of features that includes information of the time dependency structure of the speech, such as Shifted Delta Cepstral coefficients (SDC) [5] and the usage of Maximum Mutual Information (MMI) [6] to improve the discrimination rate while maximizing the mutual information between acoustic classes. In multi-class problems such as Language Recognition or even Speech Recognition, GMM-MMI and HMM-MMI models have shown notable discrimination improvements, also motivating their usage for this submission.

The rest of the paper is organized as follows. Section 2 describes feature extraction for each system. Then, we describe system details for each evaluation task, audio segmentation (Section 3) and speaker diarization (Section 4). Finally, conclusions are presented in Section 5.

2. Feature extraction

2.1. Audio segmentation

Audio Segmentation parameterization consists in 7 MFCC with CMN-Rasta-Warping concatenated to their 7-1-3-7 Shifted Delta Coefficients (SDC).

SDC features have been widely used in Language Recognition due to the fact that they capture the time dependency structure of the language better than the speed or acceleration coefficients (also known as delta and delta-delta). Similarly, SDC features are expected to distinguish the time dependency of the speech over the music or noise.

2.2. Speaker diarization

The front-end parameterization for speaker diarization is illustrated in the Figure 1. It follows a classical Speaker Recognition recipe: 19 MFCC coefficients concatenated to their deltas and followed by Cepstral Mean Normalization (CMN), RASTA filtering and feature warping.

All the training data labelled as ‘speech’, ‘speech with noise in background’ and ‘speech with music in background’ is used to train a 1024-mixtures UBM model. Given this UBM, sufficient stats are extracted for every labeled segment. The total variability subspace is then modeled following the FA recipe. The next step is to compute a LDA matrix that discriminates among speakers. Such matrix is trained with the speaker labels provided and compensated statistics, called iVectors.

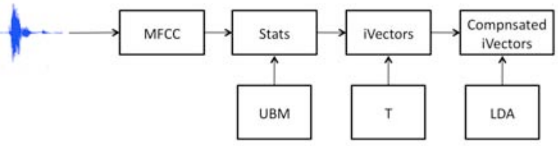


Figure 1: Schematic diagram of the feature extraction scheme for speaker diarization.

As in [1] our back-end parameterization computes iVectors every 20ms over a one second length window. Resulting iVectors are projected over the space defined by the LDA matrix.

3. Audio segmentation system

The ATVS-UAM-UAM audio segmentation system is illustrated in the Figure 2. It is based on the Viterbi alignment of the audio stream using a five-state HMM. One for each target acoustic class: 'speech', 'speech with noise in background', 'speech with music in background' 'music' and 'others'.

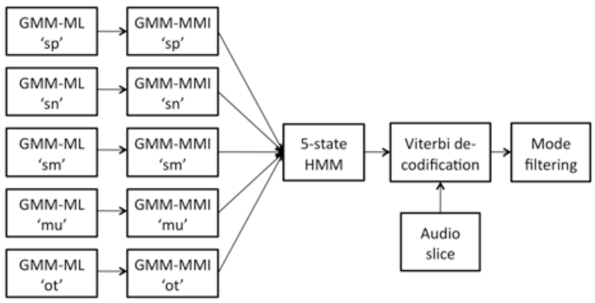


Figure 2: Schematic diagram of the ATVS-UAM audio segmentation system

Each HMM state consists in a 1024 mixtures GMM, previously trained by means of 5 iterations of the Maximum-Likelihood criterion, and enhanced later by means of 18 iterations of the Maximum Mutual Information criterion. This latter step were carried out using the HMM Toolkit STK software from BUT Speech@FIT (Brno University of Technology, Faculty of Information Technology) [7]. All development data provided for the evaluation were used to train these GMMs and no additional data were used.

The SDC features stream is previously divided into 60 seconds length audio slices that are independently processed. Initial 2 seconds of each slice are overlapped with the previous one.

Viterbi alignment is performed using the HMM Toolbox for Matlab by Kevin Murphy [8].

After the Viterbi decodification, a mode-filtering step over a 700 ms sliding window is used to avoid spurious changes between states. Finally, for each class, very short segments were removed –those ones with length smaller than around 3 seconds.

Table 1 summarizes ATVS-UAM audio segmentation system testing timing.

Table 1: *Breakdown timing for ATVS-UAM audio segmentation system.*

Testing (per 4 hours session file)	
Feature extraction	14 minutes
Viterbi decodification + mode-filtering	20 hours

4. Speaker diarization system

ATVS-UAM speaker diarization system (Figure 3) is based on the previous works [2] and [3].

The MFCC features stream is firstly divided into 90 seconds length audio slices –contiguous windows are 33% overlaped-. Compensated iVectors in each slice are clusterized based on their cosine distance. The number of clusters is controlled by maximum allowed distance between the vectors to the centroid of the cluster. In our implementation we used as centroid the averaged vector in each cluster and it represents a candidate speakers model. Candidate speaker models are accumulated over all the slices in the test session, together with the frequency of appearance of their cluster. Since speakers are expected to appear in more than one slice, a secondary clustering is used to merge the first iteration centroids, obtaining then an enhanced set of candidate speakers. A prior probability is assigned to each of the candidate speakers based on its relative frequency of appearance in the entire session.

In a second pass over the slices we compute the probability of each candidate speaker with the stream of iVectors. Such probability is estimated using the cosine distance and normalized with the prior probability of each candidate speaker. The final diarization labels are obtained with a Viterbi decodification of these scores.

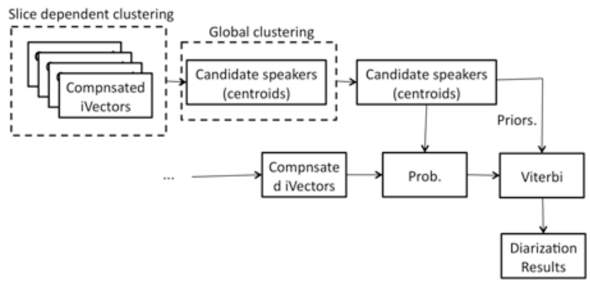


Figure 3.Schematic diagram of the ATVS-UAM speaker diarization system.

Table 2 summarizes ATVS-UAM speaker diarization system testing timing.

Table 2: *Breakdown timing for ATVS-UAM speaker diarization system.*

Testing (per 4 hours session file)	
Feature extraction	40 minutes
iVectors computation	32 hours
iVectors clustering + Viterbi decodification	15 minutes

5. Conclusions

This paper summarizes the ATVS-UAM participation in Albayzin 2010 Evaluations. ATVS-UAM submits results for two of the four proposed evaluations: Audio Segementation and Speaker Diarization. In the latest case we present a novelty approach based on FA to model the total variability subspace. The so-computed iVectors are clustered based on an estimation of the likelihood using cosine distance. Thus, centroids to each cluster can be considered candidate speakers. Likelihoods for each candidate speakers are computed in a second pass over the iVector stream. The final sequence of decisions is computed using the Viterbi algorithm. The ATVS-UAM Audio Segmentation system submitted is based on a five states HMM, each of them trained independently with a 1024 gaussians GMM using MMI. The final sequence of decisions is obtained as an enhanced Viterbi decodification.

6. References

[1] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-based Speaker Segmentation Using Speaker Factors and Eigenvoices," in Proc. ICASSP, Las Vegas, Nevada, Mar. 2008, pp. 4133 – 4136.

[2] P. Kenny, D. Reynolds, F. Castaldo "Diarization on Telephone Conversation using Factor Analysis". IEEE Journal on Selected Topics In Signal Processing. 2010.

[3] Najim Dehak, Patrick Kenny, Rda Dehak, Pierre Ouellet, and Pierre Dumouchel, "Front end Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech and Language Processing, 2010.

[4] Najim Dehak, Reda Dehak, James Glass, Douglas Reynolds, and Patrick Kenny, "Cosine similarity scoring without score normalization techniques," in Odyssey, 2010

[5] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller, Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features." *Proc. ICSLP 2002*, Sept. 2002, pp. 89-92.

[6] P. Matejka, L. Burget, P. Sckwarz, J Cernocky. "Brno University of Technology System for NIST 2005 Language Recognition Evaluation", in Proceedings of Odyssey 2006. Puerto Rico.

[7] http://speech.fit.vutbr.cz/

[8] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.htm

GTTS Systems for the Albayzin 2010 Audio Segmentation Evaluation

Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
luisjavier.rodriguez@ehu.es

Abstract

This paper briefly describes the audio segmentation systems developed by the Software Technology Working Group (http://gtts.ehu.es) at the University of the Basque Country (EHU), for the Albayzin 2010 Audio Segmentation Evaluation. The primary system consists of five Gaussian Mixture Models estimated independently on the reference segmentations provided for development, and applied on a frame-by-frame basis to get a sequence of smoothed log-likelihoods. The class yielding the maximum likelihood is chosen at each frame, and finally a mode filter is applied to smooth the sequence of decisions. The contrastive system (used as speech/non-speeh detector in the GTTS submission to the Albayzin 2010 Speaker Diarization Evaluation) consists of an ergodic Continuous Hidden Markov Model with 5 states (one per class) and 512 mixtures per state. Independent sets of segments (extracted from the reference segmentations provided for development) are used to estimate the emission distributions corresponding to the HMM states, transition probabilities being heuristically fixed. Given an input signal, this model produces an optimal decoding (and segmentation) according to the maximum likelihood criterion.

Index Terms: Audio Segmentation, Gaussian Mixture Models, Hidden Markov Models

1. Introduction

Our participation in this evaluation was motivated by our participation in the Albayzin 2010 Speaker Diarization Evaluation, since speaker diarization requires a speech/non-speech detector to discard non-speech segments (containing music, silence, noise, etc.), so that clustering is performed only on speech segments. Therefore, we have not optimized our systems for the classification task proposed in the evaluation, but for a speech/non-speech detection setup. We used the reference segmentations provided for development to estimate five acoustic models, and then applied two simple classification approaches, with two main concerns: rapid development and low computational cost. Our first (and quite obvious) approach consisted in estimating a 5-class ergodic HMM and applying maximum-likelihood Viterbi decoding. This approach yielded quite good performance in the speech/non-speech classification task. However, with the aim to improve performance on the 4-class segmentation task proposed in this evaluation, a second system was developed. The second system, based on GMM frame-by-frame scoring, yielded better results on the development set and is presented as the GTTS primary system. All speech processing, HMM/GMM estimation, Viterbi decoding and GMM likelihood computations were performed with the Sautrela toolkit [1]. Text processing and file manipulation were done with UNIX utilities and applications (awk, SoX, etc.).

This work has been supported by the Government of the Basque Country, under program SAIOTEK (project S-PE09UN47), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

2. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features. The choice of MFCC is based on the fact that historically there have been no features specifically designed for audio segmentation, and the MFCC are the most commonly used parameters for speech processing applications.

The audio was analysed in frames of 32 milliseconds (512 samples) at intervals of 10 milliseconds. A Hamming window was applied and a 512-point FFT computed. The FFT amplitudes were then averaged in 24 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. A Discrete Cosine Transform was finally applied to the logarithm of the filter amplitudes, obtaining 13 Mel-Frequency Cepstral Coefficients (MFCC), including the zero (energy) coefficient. Cepstral Mean Subtraction was not applied, in order to keep channel and background information that may be relevant for audio classification.

3. Audio segmentation based on HMM decoding (contrastive system)

Development data were organized as follows: 4 sessions (3, 7, 11 and 13) were used for tuning purposes; the remaining 12 sessions were used to estimate model parameters. In fact, these latter sessions were splitted (using SoX) into five subsets of segments, according to reference segmentations provided with development data, for the five acoustic classes: (1) music, (2) clean speech, (3) speech with music in the background, (4) speech with noise in the background and (5) other (noise, long silence fragments, etc.).

A single-state HMM was estimated for each class, using the Baum-Welch algorithm on the corresponding set of segments. An ergodic Continuous Hidden Markov Model was built by composing the five single-state HMMs under the Layered Markov Model framework defined in Sautrela [2]. Given an input sequence of feature vectors, the optimal decoding (and segmentation) was obtained by applying the Viterbi algorithm to get the optimal sequence of states in the ergodic HMM.

The number of mixtures per state (512) and the transition probabilities (auto-transitions fixed to 0.999999, transitions between states and final state transitions fixed to $2 \cdot 10^{-7}$) were optimized on audio segmentation experiments over the 4 tuning sessions. Though system performance was quite poor for the 4-class setup defined in the evaluation, when considering a 2-class speech/non-speech classification setup, the false alarm error rate was 1.16% and the miss error rate was 1.78% for the speech class (gathering the three speech sub-classes mentioned above). This system was used as speech/non-speech detector in the GTTS submission to the Albayzin 2010 Speaker Diarization Evaluation.

4. Audio segmentation based on frame-by-frame GMM scoring (primary system)

Development data were organized the same way as for the HMM-based system (12 sessions for training, 4 sessions for tuning). A GMM was estimated for each class, starting from the corresponding subset of training segments. Given an input sequence of feature vectors, the set of GMMs was applied to compute frame-by-frame log-likelihoods. A smoothing window of length N was then applied, so that each log-likelihood was replaced by the arithmetic mean computed in that window, as follows:

$$\hat{ll}(i,t) = \frac{1}{N} \sum_{k=-N/2}^{N/2} ll(i, t+k)$$

At each frame, the class yielding the highest smoothed likelihood was chosen, and a frame-level sequence of class labels was produced. Finally, a mode filter of length M was applied to smooth the sequence of decisions. The number of mixtures of the GMMs (1024), the length of the score smoothing window ($N = 100$) and the length of the mode filter ($M = 200$) were optimized on audio segmentation experiments over the 4 tuning sessions. This system yielded better results than the HMM-based system for the 4-class setup defined in the evaluation. When considering a 2-class speech/non-speech classification setup, the false alarm error rate was 1.14% and the miss error rate was 1.32% for the speech class.

5. Results

Tables 1 and 2 show the performance of the two audio segmentation systems described above on the development and evaluation sets, respectively. Besides the average segmentation error used to rank systems, miss and false alarm error rates in speech detection are shown too.

Table 1: Performance of the primary and alternative GTTS audio segmentation systems on a development set consisting of sessions 3, 7, 11 and 13.

	primary	**contrastive**
%error (AS)	43.48	48.08
%miss error (speech)	1.32	1.78
%fa error (speech)	1.14	1.16

Experiments were carried out on a Dell PowerEdge 1950, equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 4GB of RAM. CPU times (in terms of real-time factor, $\times$RT) are shown in Table 3, considering three separate operations: (1) feature extraction, (2) model estimation and (3) audio segmentation. In the latter case, I/O operations and all the secondary computations needed to carry out the 4-class audio segmentation task are counted. Note that the contrastive system employs more time than the primary system for model estimation, but is faster for audio segmentation, providing only slightly worse performance in the speech/non-speech segmentation task. The total CPU time, computed by adding CPU times for feature extraction and audio segmentation, falls below 0.05$\times$RT in both cases.

Table 2: Performance of the primary and alternative GTTS audio segmentation systems on the evaluation set (sessions 17-24).

	primary	**contrastive**
%error (AS)	45.10	48.50
%miss error (speech)	1.23	1.55
%fa error (speech)	0.90	0.86

Table 3: CPU time (real-time factor, $\times$RT) employed in feature extraction, model estimation and audio segmentation for the primary and contrastive GTTS systems.

	primary	**contrastive**
Feature extraction	0.0033	
Model estimation	0.1205	0.4819
Audio segmentation	0.0458	0.0375

6. Conclusions

Two naive audio segmentation systems have been developed and evaluated: a primary system based on frame-by-frame GMM scoring and subsequent mode filtering; and a contrastive system based on a five-class ergodic HMM which outputs the optimal Viterbi-based sequence of states (classes) given an input signal. Though their performance on the 4-class audio segmentation task proposed in this evaluation was quite poor, they provided miss and false alarm error rates of around 1% in speech detection. This makes them suitable as speech/non-speech detectors for a speaker diarization system (as we actually intended). Systems have been built and evaluated in two weeks and their CPU time requirements fall below 0.05$\times$RT.

7. References

[1] M. Penagarikano and G. Bordel, “Sautrela: A Highly Modular Open Source Speech Recognition Framework,” in *Proceedings of the ASRU Workshop*, (San Juan, Puerto Rico), pp. 386–391, December 2005.

[2] M. Penagarikano and G. Bordel, “Layered markov models: A new architectural approach to automatic speech recognition,” in *Proceedings of the MLSP Workshop*, (São Luís,Brasil), pp. 305–314, October 2004.

UPM-UC3M system for music and speech segmentation

Ascensión Gallardo Antolín

Dept. of Signal Theory and Communications
Universidad Carlos III de Madrid
gallardo@tsc.uc3m.es

Rubén San Segundo Hernández

Speech Technology Group
Universidad Politécnica de Madrid
lapiz@die.upm.es

Abstract

This paper describes the UPM-UC3M system for the Albayzín evaluation 2010 on Audio Segmentation. This evaluation task consists of segmenting a broadcast news audio document into clean speech, music, speech with noise in background and speech with music in background. The UPM-UC3M system is based on Hidden Markov Models (HMMs), including a 3-state HMM for every acoustic class. The number of states and the number of Gaussian per state have been tuned for this evaluation. The main analysis during system development has been focused on feature selection. Also, two different architectures have been tested: the first one corresponds to an one-step system whereas the second one is a hierarchical system in which different features have been used for segmenting the different audio classes. For both systems, we have considered long term statistics of MFCC (Mel Frequency Ceptral Coefficients), spectral entropy and CHROMA coefficients. For the best configuration of the one-step system, we have obtained a 25.3% average error rate and 18.7% diarization error (using the NIST tool) and a 23.9% average error rate and 17.9% diarization error for the hierarchical one.

Index Terms: music and speech segmentation, Chroma features, HMMs

1. Introduction

The problem of distinguishing speech signals from other audio signals (e.g., music) has become increasingly important as automatic speech recognition (ASR) systems are applied to more realworld multimedia domains, such as the automatic transcription of broadcast news, in which speech is typically interspersed with segments of music and other background noise. A pre-processing stage that segments the signal into periods of speech and non-speech is very important for improving recognition accuracy.

In another way, automatically detecting music parts from audio signals in TV or radio broadcasts is becoming a basic and important task to meet the increasing demands for multi-media indexing systems and music copyright management systems. In such audio signals, music is often overlapped by narration, conversation, or other environmental sounds.

Previous works on speech and music segmentation have been focused on features analysis or system architecture. About feature analysis, we can remark [1] where authors combine Mel Frequency Cepstral Coefficients (MFCCs) with other features like 4-Hz modulation energy, percentage of low energy frames, spectral centroid, spectral roll-off point, spectral flux, zero-crossing rate and spectral edge. In [2] histogram equalization-based features are proposed for speech, music, and song discrimination. In [3] an artificial neural network (ANN) trained on clean speech only (as used in a standard large vocabulary speech recognition system) is used as a channel model at the output of which the entropy and ‘‘dynamism’’ is measured every 10 ms. These features are then integrated over time through an ergodic 2-state (speech and non-speech) hidden Markov model (HMM) with minimum duration constraints on each HMM state. Finally, in [4], authors propose root mean square (RMS), and zero-crossings features for speech/music discrimination.

In respect of system architecture, [5] proposes a decision-tree-based algorithm for speech/music segmentation. [6] presents a comparison between two different techniques for speech/music discrimination. The first method is based on zero crossing rate and Bayesian classification. The second method uses more features and is based on neural networks (specifically a multi-layer Perceptron).

In work [7], authors propose a hierarchical system segmenting Broadcast News audio files. Finally, in [8] a system that segments an audio signal as speech and music by using posterior probability based features is proposed and implemented using Sphinx. This system uses Hidden-Markov-Model based acoustic models that are trained in Sphinx for posterior probability calculations. Acoustic models are trained with the HMM-states that are associated with the context-independent phones.

This paper presents the UPM-UC3M system for the Albayzín evaluation 2010 on Audio Segmentation. This system is based on Hidden Markov Models (HMMs), including a 3-state HMM for every acoustic class. For feature extraction, we have considered long term statistics of MFCC (Mel Frequency Ceptral Coefficients), spectral entropy [9] and CHROMA coefficients. Chroma features are a powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave [10].

2. Evaluation on Audio segmentation

The proposed evaluation task consists of segmenting a broadcast news audio document into a few acoustic classes (ACs):

- Speech [sp]. Clean speech in studio from a close microphone.
- Music [mu]. Music is understood in a general sense.
- Speech with noise in background [sn]. Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).
- Speech with music in background [sm]. Overlapping of speech and music classes or speech with noise in background and music classes.

There is another class that is not evaluated: Other [ot]. This class refers to any type of audio signal (including noises) that does not correspond to any other class.

2.1. Database description

The database consists of a Catalan broadcast news database from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporació Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, that includes around 87 hours of sound (24 files of approximately 4 hours long), has been split into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). The distribution of classes within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3%. The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

During system training and development, we used 16 files (sessions) that were divided randomly into two sets: 14 files for HMM training and 2 files (sessions 4 and 14) for testing the system performance.

2.2. Evaluation metrics

The proposed metric is inspired on the NIST metric for speaker diarization. The metric is defined as a relative error averaged over all ACs:

$$Error = \underset{i}{average}\left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}\right)$$

where

- *dur(miss$_i$)* – the total duration of all deletion errors (misses) for the ith AC.
- *dur(fa$_i$)* – the total duration of all insertion errors (false alarms) for the ith AC.
- *dur(ref$_i$)* – the total duration of all the ith AC instances according to the reference file.

Note that incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. In the case when the system output is Other (non-Other) and the corresponding reference label is non-Other (Other), the audio segment is computed as a deletion (insertion) error for only non-Other AC. A forgiveness collar of 1 sec (both + and -) will not be scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an acoustic class begins/ends.

In this paper, we also report, for all the experiments, the NIST metric for speaker diarization. The NIST metric is similar to the error computed with the previous formula but doing a weight average: considering every AC duration for weighting each error.

3. Baseline

The baseline is a one-step system based on HMM. In particular, we have considered a 3-state HMM model for each acoustic class, considering 16 Gaussians per state. The HMM topology can be seen in Figure 1. The number of states has been adjusted from preliminary experiments.

The features considered in this system have been statistics over 1 second window (with an overlapping of 0.5 seconds) of the 15-MFCCs (Mel Frequency Cepstral Coefficients) and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and standard deviation. In total, there are 96 features every 0.5 seconds. In preliminary experiments, several windows lengths and overlapping were considered.

For all the experiments, we have used HTK software [11] for training and testing the HMMs. For feature extraction, we have considered the OpenSMILE tool [12].

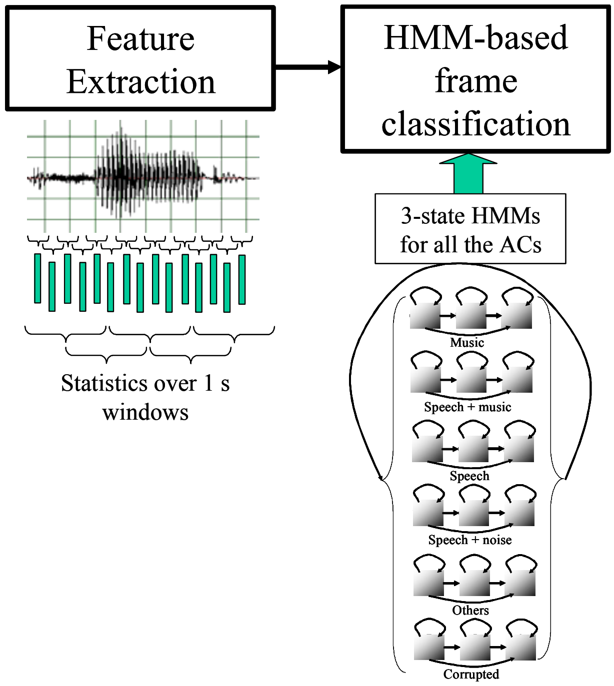


Figure 1. System diagram with details about feature extraction and HMM topology

4. Feature analysis

The main analysis during system development has been focused on feature selection. During system development, we have evaluated an important amount of features used in speech and speaker recognition. The best features for this task have been:

- MFCC15_E_D_A (mean+var): 15-MFCCs and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and variance computed along a 1 second with 0.5s overlapping.
- MFCC15_E_D_A (mean+std). Similar to the previous one but considering as statistics: mean and standard deviation.
- MFCC15_E_D_A (mean+std+skew): Similar to the previous one but considering as statistics: mean, standard deviation and skewness.
- MFCC15_E_D_A (mean+std+skew+kurt): adding kurtosis as a new statistic.
- MFCC15_E_D_A (mean+std+kurt): same that previous one removing skewness.
- PLP7_E_D_A (mean+std): 7-PLP (Perceptual Linear Prediction Coefficients) and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and standard deviation along a 1 second with 0.5s overlapping
- PLP5_E_D_A (mean+std): Same to previous one but considering 5 PLP instead of 7.

- PLP5_E_D_A (mean+std+skew+kurt): adding skewness and kurtosis as new statistics along 1 second window.
- MFCC15CHR_E_D_A (mean+std): 15-MFCCs, local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta and 12 CHROMA coefficients computed every 50 ms. Statistics are mean and standard deviation along a 1 second with 0.5s overlapping.
- MFCC15CHR+SpectralFeatures_E_D_A (mean+std): same to the previous one adding the statistics (mean and standard deviation) of several spectral features computed at 50ms frames (flux, centroid, entropy and band energies).
- MFCC15CHR+Entropy_E_D_A (mean+std): same to the previous one adding only the mean and standard deviation of the spectral entropy.

Table 1 presents the results for different features: FALA error for each AC, the average (AVG) and the NIST error.

Features	FALA error					NIST
	mu	sm	sp	sn	avg	
MFCC15_E_D_A (mean+var)	16.8	34.6	44.1	54.5	37.5	26.3
MFCC15_E_D_A (mean+std)	18.6	33.7	32.1	45.8	32.6	23.2
MFCC15_E_D_A (mean+std+skew)	14.7	33.0	38.0	51.8	34.4	24.8
MFCC15_E_D_A (mean+std+skew +kurt)	13.6	27.5	40.1	47.0	32.1	22.8
MFCC15_E_D_A (mean+std+ kurt)	15.2	31.8	42.5	49.8	34.8	24.6
PLP7_E_D_A (mean+std)	17.6	36.1	39.8	52.4	36.5	25.6
PLP5_E_D_A (mean+std)	13.6	38.1	40.9	53.5	36.5	26.3
PLP5_E_D_A (mean+std+skew +kurt)	12.3	34.2	41.6	50.4	34.6	25.1
MFCC15CHR_E_D_A (mean+std)	19.1	24.9	29.4	41.3	28.7	20.9
MFCC15CHR+SpectralFeatures_E_D_A (mean+std)	19.1	31.0	30.4	44.6	31.3	22.1
MFCC15CHR+ Entropy_E_D_A (mean+std)	15.7	27.7	28.0	38.9	**27.5**	**20.1**

Table 1. Results for different features: FALA error for each AC and the average (AVG) and the NIST error

From these experiments, the main conclusions are:

- Including CHROMA coefficients allows reducing significantly the error for all ACs from 32.6% to 28.7%.
- Including several spectral features frames (flux, centroid, entropy and band energies) do not improve the results. However, when only the spectral entropy is considered, the average error is reduced to 27.5%.
- For music segmentation, PLP features perform better than MFCC (considering even less number of features: 5 instead of 15) obtaining a 12.3% error.

5. Increasing the number of Gaussians per state

The feature study (in previous sections) was performed considering 16 Gaussians per state in order to reduce the time needed for the experiments. After this analysis, we decided to increase the number of Gaussians per state obtaining a good number around 128 or 256. Because of this, we decide to repeat the experiment with the most promising features but, in this case, considering 128 Gaussians per state, except in the case of MFCC15_E_D_A (mean+std) for which better results were achieved with 256 Gaussians. The considered features were:

- MFCC15_E_D_A (mean+std): 15-MFCCs and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean and standard deviation along a 1 second with 0.5s overlapping.
- MFCC15_E_D_A (mean+std+skew+kurt): adding skewness and kurtosis as a new statistics.
- MFCC15CHR+Entropy_E_D_A (mean+std): 15-MFCCs, local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta, 12 CHROMA coefficients computed every 50 ms, and the 50ms frame spectral entropy. Statistics are mean and standard deviation along a 1 second with 0.5s overlapping.
- PLP5_E_D_A (mean+std+skew+kurt): 5-PLP and local energy computed in 25ms windows (with an overlapping of 15ms), their delta and double delta. The statistics are mean, standard deviation, skewness and kurtosis along a 1 second with 0.5s overlapping.

Table 2 presents the results for different features: FALA error for each AC, the average (AVG) and the NIST error.

Features	FALA error					NIST
	mu	sm	sp	sn	avg	
MFCC15_E_D_A (mean+std)	20.7	33.9	24.2	32.5	27.8	19.9
MFCC15_E_D_A (mean+std+skew+ kur)	17.3	29.7	26.6	37.9	27.9	19.8
PLP5_E_D_A (mean+std+skew+ kurt)	15.6	34.3	31.4	40.2	30.4	22.4
MFCC15CHR+ Entropy_E_D_A (mean+std)	14.7	25.7	26.8	34.0	**25.3**	**18.7**

Table 2. Results for different features and 128 or 256 Gaussians per state: FALA error for each AC and the average (AVG) and the NIST error

The main conclusions are:

- In all cases, when increasing the number of Gaussians the results are better. Best results are obtained when using MFCC+CHROMA+Entropy features.
- For music segmentation, PLP and MFCC+CHROMA+Entropy features perform better than only MFCC, whereas MFCC+CHROMA+Entropy are the best features for extracting speechmusic segments. However, speech and speechnoise are better segmented using only MFCC. Following these observations we consider the possibility to implement a hierarchical system in which different features are used for segmenting different acoustic classes. This system is described in next section.

6. Hierarchical system

According to the results presented in the previous section we designed a hierarchical system whose structure is shown in Figure 2. Two alternatives were considered:

- HS1: in this case, music and speechmusic are segmented using, respectively, PLP and MFCC+CHROMA

+Entropy features and finally, speech and speechnoise segments are extracted using MFCC parameters.
- HS2: MFCC+CHROMA+Entropy features are used for extracting music and speechmusic whereas MFCC are considered for segmenting speech and speechnoise.

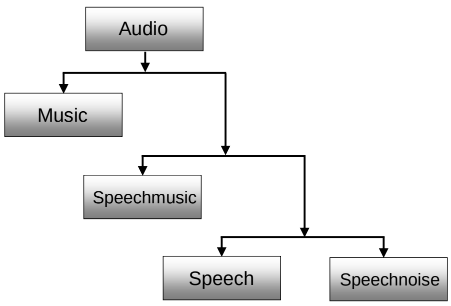


Figure 2. Block diagram of the hierarchical system

As it can be observed in Table3, results achieved by HS1 and HS2 systems are very similar. In comparison with the one-step system (with MFCC+CHROMA+Entropy features), the hierarchical one improves the results, yielding an error reduction of about 1.4% absolute.

System	FALA error					NIST
	mu	sm	sp	sn	AVG	
HS1	13.8	26.1	23.9	31.1	23.7	18.0
HS2	14.7	25.7	23.9	31.2	**23.9**	**17.9**

Table 3. Results for different hierarchical systems: FALA error for each AC and the average (AVG) and the NIST error

7. Final results

For the final evaluation, we presented two systems. The main system considering the best non-hierarchical system: using MFCC15CHR+ Entropy_E_D_A (mean+std) and the best hierarchical system HS1 described in previous section. The results are presented in Table 4.

System	FALA error				
	mu	sm	sp	sn	AVG
Main	19.2	25.0	39.5	37.2	30.2
Alternative	19.2	25.1	39.3	37.2	**30.2**

Table 4. Final results with the evaluation set: FALA error for each AC and the average (AVG)

8. Conclusions

This paper describes the UPM-UC3M system for the Albayzín evaluation 2010 on Audio Segmentation. The proposed system is based on Hidden Markov Models (HMMs), including a 3-state HMM for every acoustic class. The number of states and the number of Gaussian per state have been tuned for this evaluation.

The main analysis during system development has been focused on feature selection. From the experiments, we can conclude that MFCC are better complemented with CHROMA coefficients and spectral entropy than with other spectral features like flux, centroid and band energies in a one-step system.

Also a hierarchical system has been investigated obtaining slightly improvements over the one-step one when using MFCC features for segmenting speech and speechnoise and MFCC+CHROMA+Entropy for music and speechmusic.

In summary, for the best configuration of the one-step system, we have obtained a 25.3% average error rate and 18.7% diarization error (using the NIST tool) and a 23.9% average error rate and 17.9% diarization error for the hierarchical one.

9. Acknowledgements

This work has been supported by MA2VICMR (2010-2013), INAPRA (DPI2010-21247-C02-02), SD-TEAM (MEC ref: TIN2008-06856-C05-03), and TEC2008-06382 projects. Authors also want to thank all the comments from their colleagues at GTH-UPM and UC3M.

10. References

[1] Tomonori Izumitani, Ryo Mukai, and Kunio Kashino. ”A background music detection method based on robust feature extraction”. Proc ICASSP 2008.

[2] Ascensión Gallardo-Antolín and Juan M. Montero, “Histogram Equalization-Based Features for Speech, Music, and Song Discrimination”. IEEE Signal processing letters, Vol. 17, No. 7, 2010.

[3] Jitendra Ajmera, Iain McCowan, Herve Bourlard. “Speech/music segmentation using entropy and dynamism features in a HMM classification framework”. Speech Communication 40 (2003) 351–363.

[4] Costas Panagiotakis and George Tziritas, “A Speech/Music Discriminator Based on RMS and Zero-Crossings”. IEEE Trans. On Multimedia, Vol. 7, No. 1, Feb 2005.

[5] Yizhar Lavner1 and Dima Ruinskiy, “A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation”. EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[6] Alessandro Bugatti, Alessandra Flammini, PierangeloMigliorati “Audio Classification in Speech and Music: A Comparison Between a Statistical and a Neural Approach”. EURASIP Journal on Applied Signal Processing 2002:4, 372–378.

[7] Mateu Aguilo, Taras Butko, Andrey Temko, Climent Nadeu “A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task”. Proc I SLtech 2009. Lisbon.

[8] Cemil Demir; Erdem Ünal, Mehmet Ugur Dogan, “A Sphinx Based Speech-Music Segmentation Front-End For Improving The Performance Of An Automatic Speech Recognition System In Turkish”. CMU Sphinx Workshop 2010.

[9] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky, “Spectral Entropy Based Feature for Robust ASR”, Proc. ICASSP 2004, pp. 193–196.

[10] M. A. Bartsch & G. H. Wakefield, “To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing,” WASPAA, Mohonk, 2001.

[11] S. Young et al., HTK-Hidden Markov Model Toolkit (Ver 3.2). Cambridge, MA: Cambridge Univ., 2002.

[12] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Firenze, Italy, 25.-29.10.2010.

Using Fingerprinting to Aid Audio Segmentation

Carla Lopes[1,2], *Arlindo Veiga*[1,2], *Fernando Perdigão*[1,2]

[1] Department of Electrical and Computer Engineering, University of Coimbra
[2] Instituto de Telecomunicações – Pólo de Coimbra
{calopes, aveiga, fp}@co.it.pt

Abstract

Audio segmentation is an important preliminary task in audio description systems. In this paper a three-phase audio segmentation scheme is proposed, where the first phase detects silence events, the second phase detects audio repetitions trough fingerprinting and the last phase uses a hybrid HMM/ANN system to classify the remaining segments. Fingerprinting is an important aspect of audio segmentation in broadcast audio, due to the omnipresence of advertisements, jingles and even repeated programs. Detecting repetitions can be extremely fast and accurate and also permits to enhance the consistence of the annotations. Results using the Catalan broadcast 3/24 TV channel are reported.

Index Terms: audio segmentation, fingerprint, jingle detection

1. Introduction

As the amount of multimedia published data grows, the problem of managing the information contained in this data becomes more and more difficult. Tasks like: finding a roar of laughter to reuse it when editing our own audio or video; counting the number of times that a publicity spot occur or transcript broadcast news are practically impossible. Indexing and content-based retrieval are then crucial to handle the large amounts of audio and multimedia data that is becoming available on the web. Manual annotation is useful in some applications and can provide accurate description of the content. An example of this is when an upload is made in YouTube.com and the user insert keywords to describe the content. Nevertheless, in the main situations manual indexing is extremely time consuming, subjective, tedious and expensive. In these cases audio segmentation and classification plays an important role.

The purpose of audio segmentation is to divide an audio clip into several segments so that each segment contains only one class of audio. Audio signals which include speech, music and environmental sounds are important types of media. The problem of distinguishing audio signals into these different audio types is thus becoming increasingly significant. Although there are many approaches to audio segmentation they are focused on a narrow type of audio such as speech/music separation, speaker recognition and music structure extraction.

In this paper we propose to index broadcast news audio documents in five broad-classes: speech, music, speech with music in background, speech with noise in background and other. After a signal processing section, audio indexing is made up of two main sections: a segmentation section and a classification section. The first one uses silence information and fingerprint information to segment the original audio signal while the second performs a classification in term of five classed within each unlabeled segment. The classification is made by means of a hybrid ANN/HMM system.

2. Hierarchical Audio Segmentation

Broadcast news data are usually stored in long files almost impossible to analyze as a whole. When Viterbi decoding is involved, as in the present work, attention must be paid to the length of the sequence to decode, because the decoding tree grows in such a way that becomes impossible to handle. In view of this we propose a two-phase audio segmentation where the first phase detects silence events in the audio signal in order to limit the length of the segments to classify. We also use make a fingerprint of all the audio session in order to find repetitions in it. The next subsections describe the both detectors.

2.1. Silence Detector

The silence detector is a simple one, based only on the energy of the signal based on a window of 200ms with a shift of 100ms. Energy is computed in dB and when there are segments larger than 1 second below a threshold a silence event is annotated.

2.2. Repetitions detector

Audio fingerprinting refers to a condensed representation of an audio signal that can be used to identify an audio sample or quickly locate similar items in audio streams. We use a fingerprinting system where a 32-bit binary pattern is computed for each frame of about 200ms. The frame rate is 50 frames per second, allowing enough time resolution. The signal is first down sampled to 8 kHz and a spectral analysis is performed with a mel filterbank with 33 channels. The resulting spectrogram is binarized into 32 bits per frame, with a 1-bit, essentially, when there is a spectral peak [4].

The searching method is very simple. It corresponds to count the matching bits between the signature and audio binary patterns, in each frame, when the signature pattern slides over the audio pattern, in order to compute the mean bit error rate. When the bit error rate decreases below a threshold, a match is encountered. Modern processors have a special instruction to count bits, which turns this technique even faster.

In the present case we have used every reference label of the training database as a possible pattern that may repeat. We found thousands of repetitions with this method. However, as the labels of the repeated segments are not always the same, we implemented a voting method to give to that segment the label most referenced.

The information of the two detectors (silence and repetitions) are combined and the output is the original audio signal with silence and music events annotated. In Figure 1 it is showed an example where the upper sequence correspond to the reference sequence annotated in terms of the five classes and the bottom sequence the output of the hierarchical audio segmentation. The grey segments remain unclassified and each one of these segments will be given to an audio classification system in order to annotate the entire audio file. Section 3 describes the audio classification system.

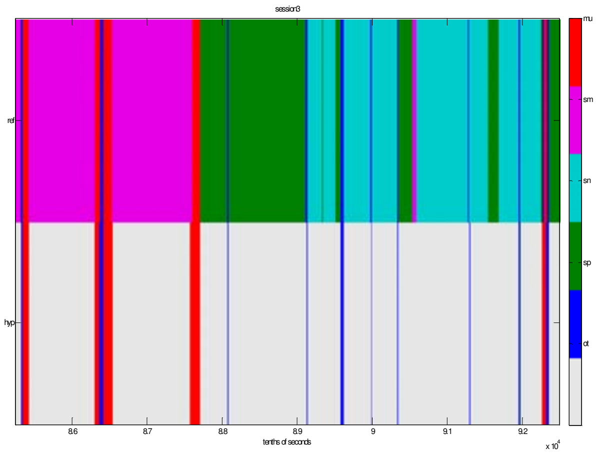


Figure 1: *Hierarchical audio segmentation example.The color coding is the following: mu - music:red; sm – speech with music: magenta; sn – speech with noise: cyan; ot – other: blue.*

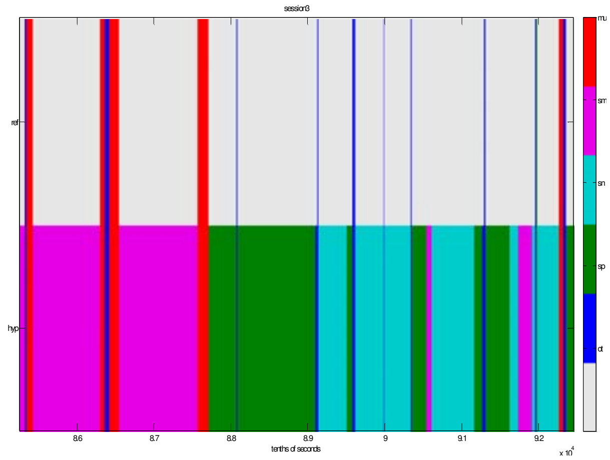


Figure 2: *Audio Classification.*

3. Audio Classification

Audio classification refers to not only the attribution of the correct sequence of labels but also to the labels' boundaries. The audio classification is based on a Viterbi decoding which is applied to each unlabeled segment (grey segments of Figure 1 and 2). The used approach uses a hybrid MLP/HMM system.

3.1. Hybrid MLP/HMM

An MLP network consisting of an input layer (with 200 hidden nodes) and an output layer with five nodes, one for each classe to classify. The 48 parameters described in Section 4.1 were used as standard input features, and a context window of 10 frames in the left and in the right of each frame was considered in the input layer. The softmax function was used as the activation function of the output layer, so that the output values are interpreted as a posterior probability of each class. All the weights and bias of the network are adjusted using batch training with a resilient back-propagation (RP) algorithm [1], so as to minimize the minimum-cross-entropy error between network output and the target values.

In the proposed hybrid approach we considered that the output predictions of the MLP correspond to class posterior probabilities for the input features and we use them as local probabilities in HMM. HMM acoustic models were built for all classes by using HTK 3.4 **Erro! A origem da referência não foi encontrada.**. Each event was modeled by a 10-state left-to-right HMM and each state shared the same MLP output. The HMM also shares the transition probabilities, which were adjusted in order to the model have the same mean duration as the corresponding event. We used HTK with some changes in order to replace the usual Gaussian mixture models by the normalized MLP outputs values and class priors.

In Figure 2 shows an example where the upper sequence corresponds to the output of the hierarchical audio segmentation system and the bottoms sequence the output of the audio classifier. Comparing this figure with Figure 1 (with the reference sequence in the upper part), we can see that most of the events were correctly recognized.

4. Experiments

The Catalan broadcast news database from the 3/24 TV channel was used for both training and testing the segmentation and classification systems, in the context of the Albayzin 2010 Evaluation Campaign [1]. It includes recorded PCM audio at 16-bit resolution and 16 kHz sampling frequency. The training set consists of 16 audio files and the test set of 8 audio files. The training material is labeled in terms of five classes:

1. ***Speech*** [sp].
Clean speech in studio from a close microphone.

2. ***Music*** [mu].
Music is understood in a general sense.

3. ***Speech with music*** in background.
Overlapping between speech and music classes.

4. ***Speech with noise*** [sn] in background.
Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).

5. ***Other***.
This class refers to any type of audio signal (including noises) that doesn't correspond to the other four classes

There is a great irregularity within the classes' distribution as depicted in Figure 3.

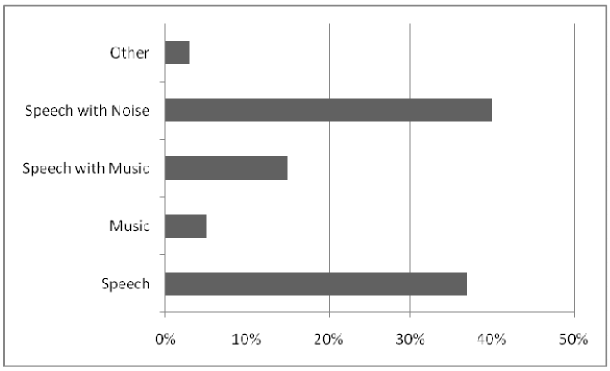


Figure 3: *Distribution of classes within the Catalan broadcast news database.*

The proposed evaluation task consists of segmenting the broadcast news audio files into the referred to five classes.
The performance of the segmentation system is evaluated by means of the metric proposed by the Albayzin 2010 Evaluation Campaign which is based on the NIST speaker diarization task. The metric computes for each class (except *other*) the amount time incorrectly identified (deletion and insertion errors) with respect to the total duration of the respective class. Similar to other metrics used in speech segmentation like *Agreement*¸ a tolerance in a window of ±1seg is given because manual alignments are prone to subjectivity. Considering *Dur* as duration, the metric may be defined as:

$$Error = \\ = \frac{1}{4}\sum_{i=1}^{4}\frac{Dur\left(deletions_{Class_i}\right)+Dur\left(insertions_{Class_i}\right)}{Dur\left(Class_i\right)} \cdot \qquad (1)$$

4.1. Features

Neural Networks are capable of incorporating all kinds of input features and adjust itself in such a way that the optimal combination of these features is found for classification. Exploiting this potentiality, input features derived from two different parameterization algorithms are combined: standard MFCC and an additional set. Since music requires more frequency resolution than speech, we considered 16 cepstral coefficients, including c_0, and not the usual 12 coefficients. Table 1 shows the additional set of features used. The feature vector comprises 48 features (16 cepstral coefficients + additional set plus first order derivative). Speech is analyzed every 100ms with a 200ms Hamming window. The classifier gives a sequence of labels, with a minimum duration of one second. The decoder is applied to the segments between the pre-determined silence/music events and not to the entire session.

Table 1. *Acoustic Feature Set used in combination with standard MFCC Features.*

Number	Feature description
1	Frame Energy in dB
2	Zero Crossing Rate
3	Spectral Centroid
4	Spectral roll-off - 90%
5	Max normalized correlation coefficient in each frame (harmonicity measure)
6	Frequency corresponding to the max normalized correlation coefficient
7	Harmonicity measure (proportional to the harmonic duration)
8	Spectral Flux

4.2. Results

A final classification error of 20.68% was obtained in the 16 sessions of the training part of the database. Table 2 shows the results per class. The class music achieved the best performance while speech with noise achieved the worse.
Quite different is the final classification using the audio test material. The final error rate almost doubled! Music kept the best class performance nevertheless it degraded 8% in the error rate. One reason for this is that the total time of repetitions and silence encountered in test sessions is only about 12% compared with the 65% on the reference sessions. In fact, we found 4427 non-overlapping segments in the reference sessions that repeat at least twice in all sessions (that we have called jingles). This corresponds to 65% of the total reference session's time.
The evaluation corresponds to finding the silences and jingles, and to the Viterbi decoding of the segments between silences and/or jingles. The code was mainly implemented in Matlab and the used machine was a Cray CX1.

Table 2. *Classification error results using training material.*

```
=======================================
The error-rate of the class mu : 13.62%
The error-rate of the class sp :  22.78%
The error-rate of the class sm :  20.10%
The error-rate of the class sn :  26.21%
=======================================
The final error-rate : 20.68%
=======================================
```

Table 3. *Classification error results using test material.*

```
=======================================
The error-rate of the class mu :  21.43%
The error-rate of the class sp :  48.03%
The error-rate of the class sm :  51.66%
The error-rate of the class sn :  48.49%
=======================================
The final error-rate : 42.40%
=======================================
```

5. Conclusions

Although the results were lower than we expected, the method of finding repetitions with fingerprinting is important in audio segmentation of broadcast audio where repetitions are always present. The observed differences from the reference and test results can be explained to overtraining but also to the segmentation method, which relies on the repetitions of audio segments. However, the method exploits the inconsistencies that exist in the annotations, which contributes also to the observed errors.

6. Acknowledgments

Carla Lopes would like to thank the Portuguese foundation: Fundação para a Ciência e a Tecnologia for the PhD Grant (SFRH/BD/27966/2006).

7. References

[1] http://fala2010.uvigo.es/index.php?option=com_content&view=article&id=60%3Aaass&catid=36&Itemid=65&lang=pt#_ftn1

[2] Riedmiller, M. and Braun, H., "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in Proc. ICNN, San Francisco, CA, 1993, pp. 586–591.

[3] Young, S. *et al*, The HTK book. Revised for HTK version 3.4, Cambridge University Engineering Department, Cambridge, December 2006.

[4] C. Neves, A. Veiga, L. Sá, and F. Perdigão, "Audio Fingerprinting System for Broadcast Streams," in Proceedings of the Conference on Telecommunications (ConfTele2009), pp. 481-484, 2009.

A Hierarchical Architecture with Feature Selection for Audio Segmentation in a Broadcast News Domain

Taras Butko and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya, Barcelona, Spain
taras.butko@upc.edu, climent.nadeu@upc.edu

Abstract

This work presents a hierarchical HMM-based audio segmentation system with feature selection designed for the Albayzin 2010 Evaluations. We propose an architecture that combines the outputs of individual binary detectors which were trained with a specific class-dependent feature set adapted to the characteristics of each class. A fast one-pass-training wrapper-based technique was used to perform a feature selection and an improvement in average accuracy with respect to using the whole set of features is reported.

Index Terms: audio segmentation, broadcast news, international evaluations

1. Introduction

Audio segmentation is the task of segmenting a continuous audio stream in terms of acoustically homogenous regions. Recently, the audio segmentation has received increasing attention for its applications in automatic indexing, subtitling, content analysis and information retrieval. Many research works address the problem of audio segmentation in different scenarios. In [1] the authors propose a method for robust speech, music, environment noise and silence segmentation of the audio recorded in different conditions such as TV studio, telephone etc. In [2] the audio stream from broadcast news domain is segmented into 5 different types including speech, commercials, environmental sound, physical violence and silence. The content based retrieval using TV programs is considered in [3], where 7 similar classes are defined.

Besides, several speech technologies can benefit from audio segmentation done at early steps. A previous identification of speech segments facilitates the task of speech recognition or speaker diarization. Furthermore audio segmentation is widely used to make online adaptation of ASR models or generating a set of acoustic cues for speech recognition to improve overall system performance [4].

In the context of the Albayzín-2010 evaluation campaign, which is an internationally-open set of evaluations organized by the Spanish network of speech technologies, an audio segmentation task was proposed and organized by the authors. For this evaluation we propose a system that uses a hierarchical architecture with HMM-GMM-based binary detectors. Each detector, one per class, uses a specific feature set, which is designed by adapting a feature selection technique recently introduced by the authors. In the experimental part, the results obtained by using the training data of the evaluation for both training, development and testing are presented. When compared with a one-step multi-class system, our system shows a 25% average relative improvement.

2. Albayzin 2010 audio segmentation evaluation

2.1. The database

The database used for evaluations consists of a Catalan broadcast news database from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was manually annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project. The database includes around 87 hours of annotated audio (24 files of approximately 4 hours long). According to this material five different audio classes were defined (Table 1). The distribution of classes within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3%. The database was splitted into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

The Corporació Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development.

Table 1. *The 5 acoustic classes defined for evaluations.*

Class	*Description*
Speech [sp]	Clean speech in studio from a close microphone
Music [mu]	Music is understood in a general sense
Speech over music [sm]	Overlapping of speech and music classes or speech with noise in background and music classes
Speech over noise [sn]	Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation)
Other [ot]*	This class refers to any type of audio signal (including noises) that doesn't correspond to the other four classes

* Not evaluated in final tests

2.2. Metric

The metric is defined as a relative error averaged over all acoustic classes (ACs):

$$Error = \underset{i}{average}\left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}\right) \qquad (1)$$

where

$dur(miss_i)$ – the total duration of all deletion errors (misses) for the i*th* AC.

$dur(fa_i)$ – the total duration of all insertion errors (false alarms) for the i*th* AC.

$dur(ref_i)$ – the total duration of all the i*th* AC instances according to the reference file.

The incorrectly classified audio segment (a substitution) is computed both as a deletion error for one AC and an insertion error for another. A forgiveness collar of 1 sec (both + and -) is not scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an AC begins/ends.

3. The UPC audio segmentation system

3.1. Features

A set of audio spectro-temporal features, like those used in automatic speech recognition, is extracted to describe every audio frame. It consists of 16 frequency-filtered (FF) log filter-bank energies with their first time derivatives [5], which represent the spectral envelope of the audio waveform within a frame, as well as its temporal evolution. In total, a 32-dimensional feature vector is used. The FF feature extraction scheme consists in calculating a log filter-bank energy vector for each signal frame (in our experiments the frame length is 30 ms with 20 ms shift, Hamming window is applied) and then applying a FIR filter *h(k)* on this vector along the frequency axis. We use the *h(k)={1, 0, -1}* filter in our approach. The end-points are taken into account which represent the absolute energies of the first and the last filter banks.

3.2. The system architecture

Our hierarchical system architecture is a group of detectors (called modules), where each module is responsible for detection of one acoustic class of interest [6]. As input it uses the output of the preceding module and has 2 outputs: the first corresponds to audio segments detected as corresponding class of interest, and the other is the rest of the input stream. One of the most important decisions when using this kind of architecture is to put the modules in the best order in terms of information flow, since some modules may benefit greatly from the previous detection of certain classes. For instance, previous detection of the classes that show high confusion with subsequent classes potentially can improve the overall performance.

On the other hand, in this type of architecture, it is not necessary to have the same classifier, feature set and/or topology for different detectors. Tuning of parameters is done in each the system independently, and the two-class detection can be done in a fast and easy way. Given the modules, the detection accuracy can be computed individually and a priori. Those modules with best accuracies are then placed in the early stages to facilitate the subsequent detection of the classes with worst individual accuracies.

In our implementation, each binary detector consists of 2 HMMs: “Class” and “non-Class”. Using the training approach known as one-against-all method [7], all the classes different from “Class” are used to train the “non-Class” model. Both HMMs have 3-states (with only 1 emitting state) and the observation distributions are Gaussian mixtures with continuous densities, and consist of 64 components with diagonal covariance matrices. The HTK[8] toolkit is used to perform training and the final segmentation.

The flow diagram of our hierarchical architecture is presented in Figure 1. The whole detection system consists of 5 binary detectors. Each binary detector (except silence) is trained using the features which were selected during the fast selection procedure (described in the next section).

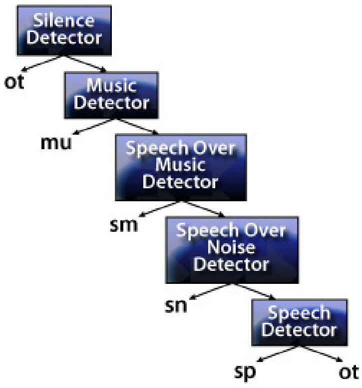


Figure 1: *Flow diagram of the hierarchical architecture.*

4. Feature selection

Actually, feature selection plays a central role in the tasks of classification and data mining, since redundant and irrelevant features often degrade the performance of classification algorithms [9]. In this paper, we use a fast one-pass-training feature selection technique [10] that avoids retraining of acoustic models during each evaluation of the candidate feature set.

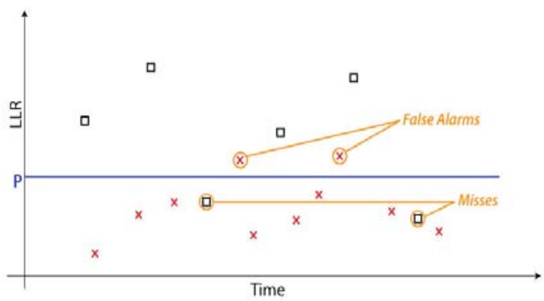


Figure 2: *Log-likelihood ratios Δi of the “Class” (squares) and the “non- Class” (crosses) segments.*

In order to perform feature selection, the database was divided into 2 parts: training and development. The fast one-pass-training feature selection technique when applied to the audio segmentation task consists in following steps:

1. Perform an initial training of "Class" and "non-Class" HMMs using the whole set of 32 FF features using the training part of the database.

2. Cut the development database into the short segments of 10 sec. Each such a segment belongs to either "Class" or "non-Class" (according to the ground truth labels).

3. Compute the log-likelihood ratios (LLRs) Δ_i of each such a segment given "Class" and "non-Class" models estimated on the step 1.

As an example, in Figure 2 we display the LLRs for all such segments in the labeled development database. Squares correspond to the "Class" instances while crosses correspond to "non-Class" instances. We consider the parameter P as a threshold (the horizontal line in Figure 2). We assume that the ith instance is detected as "Class" if its LLR ΔL_i is above P, otherwise it is detected as "non-Class". Thus, all "Class" instances (squares in Figure 2) below the P line are misses, and all "non-Class" instances (crosses) are false alarms. The parameter P is selected in such a way that the numbers of misses and false alarms are equal (equal error rate). The total number of errors (misses plus false alarms) is used as an objective function Ω for feature selection.

4. The LLR Δ_i of each segment is decomposed into a sum of "contributions" [10] coming from each feature j $(j \in 1..32)$

$$\Delta_i = \delta_{i,1} + \delta_{i,2} + \ldots + \delta_{i,32}, \qquad (2)$$

Using the sequential forward selection (SFS) approach, we iteratively select features (that correspond to the terms in (2)) that maximize the objective function Ω.

5. Experimental results

There are 16 sessions available for designing the audio segmentation system according to the evaluation plan. Half of the sessions we decided to use for training/development and the other half for testing.

First we select the appropriate number of Gaussians per HMM model for each binary detector. Actually, this number is a trade-off between the improvement in performance and the execution time needed to train the models with corresponding number of Gaussians. With 256 Gaussians we got the acceptable results. Fig. 3 demonstrates the mean error-rate obtained with increasing of the number of Gaussian mixtures per model.

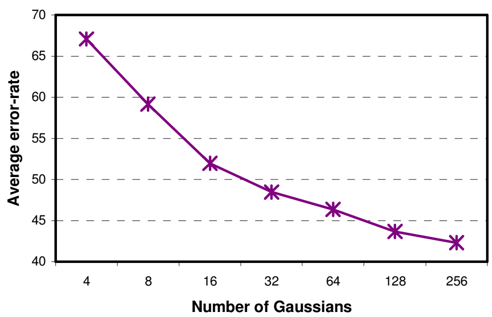


Figure 3: *Relation between mean error-rate and the number of mixtures per each GMM model.*

In Figure 4 we compare different system architectures. The "One-step multi-class" system corresponds to the HMM audio segmentation performed in one step. The "Hierarchical" architecture is described in sub-section 3.2. Finally, the system "Hierarchical + FS" is the same as previous but uses the feature selection.

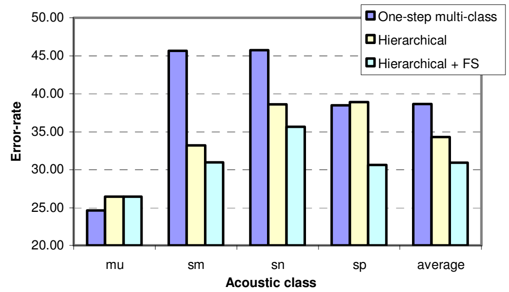


Figure 4: *Comparison of different detection systems.*

According to results from Figure 4, the hierarchical structure of audio segmentation outperforms the one-step multi-class detection system (about 8% of absolute error-rate reduction in average). Feature selection further improves the results for all classes (except music) and in average the absolute improvement is about 3%. Besides of that improvement we reduced the size of feature vector from 32 to 24 features in average.

The confusion matrix that corresponds to the "Hierarchical + FS" segmentation system is presented in Table 2, which shows the percentage of hypothesized AEs (rows) that are associated to the reference AEs (columns), so that all the numbers out of the main diagonal correspond to confusions.

Table 2: *The confusion matrix of acoustic classes.*

	mu	sp	sm	sn
mu	90	0	6	3
sp	0	83	1	16
sm	2	2	88	8
sn	0	10	7	83

As we see, the most common errors are confusions between "Music" and "Speech over music" classes and also among "Speech", "Speech over noise" and "Speech over music". Indeed, these classes have very similar acoustic content. Besides in many cases the ground truth labeling of audio is based on subjective reasoning of the annotator.

Table 3 we summarize the final results on testing database.

Table 3. *Final results on testing database.*

Database	Error-rate				
	mu	sp	sm	sn	Average
Result1	**26.40**	**44.20**	**33.88**	**41.52**	**36.50**
Result2	24.55	41.82	32.01	40.92	34.82

Where the *Result1* is obtained by the date of the presentation of final results. For that system we used the Gaussian models with only 64 mixtures and only 33% of training data were used to train the models. The *Result2* was obtained using Gaussian models with only 256 mixtures and 100% of training data.

The CPU time employed to perform testing is described below:
Feature extraction: **546 sec;**
Viterbi segmentation: **3329 sec;**
Total: **3845 sec**.

This processes were executed on PC with Intel Core 2 CPU, 2.13 GHz, 1Gb of RAM.

6. Conclusions

In this work we proposed a hierarchical HMM-based system for broadcast news audio segmentation designed for Albayzin-2010 evaluation campaign. The main advantage of such a system is that each binary detector is placed in such order that previous detections improve the results of subsequent detector.

By using a fast one-pass-training feature selection approach we have selected the subset of features that shows the best detection rate for each acoustic class, observing an improvement in average accuracy with respect to using the whole set of features. The dimension of feature vector was reduced to 24 features (in average). Such a fast technique is a good alternative to the conventional SFS hill-climbing approach when the amount of data used for training the acoustic models is large.

When compared with a one-step multi-class system, our system shows a 25% average relative improvement.

7. Acknowledgements

The authors wish to thank our colleagues at GTTS, GTC-VIVOLAB, GSI, CEPHIS-UAB, ATVS-UAM, GTM, GTH-UPM for their participation in the evaluation. Also, the authors are very grateful to Henrik Schulz for managing the collection of the database and helping for its annotation. This work has been funded by the Spanish project SAPIRE (TEC2007-65470). The first author is partially supported by a grant from the Catalan autonomous government.

8. References

[1] L. Lie, J. Hao and Z. HongJiang, "A robust audio classification and segmentation method", Proc. 9th *ACM conference on Multimedia*, p. 203-211, 2001

[2] T. L. Nwe H. Li, "Broadcast news segmentation by audio type analysis", in Proc. *ICASSP*, vol. 2, pp. 1065-1068, 2005

[3] D. Li, I.K. Sethi, N. Dimitrova, T. McGee, "Classification of general audio data for content-based retrieval", in *Pattern Recognition Letters*, v. 22, pp. 533-544, 2001

[4] H. Meinedo, J. Neto, "Audio Segmentation, Classification And Clustering in a Broadcast News Task", Proc. *ICASSP*, vol. 2, pp. 5-8, 2003

[5] C. Nadeu, J. Hernando, M. Gorricho, "On the decorrelation of filter-bank energies in speech recognition", in Proc. *European Speech Processing Conference*, pp. 1381–1384, 1995

[6] M. Aguilo, T. Butko, A. Temko, C. Nadeu, "A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task", In Proc. *I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, 2009

[7] R. Rifkin, A. Klautau, "In defense of One-Vs-All Classification", *Journal of Machine learning Research,* vol. 5, pp.101-141, 2004

[8] S.J. Young et al., "The HTK Book (for HTK Version 3.2)", Cambridge University, 2002

[9] R. Kohavi, G. John, "Wrappers for feature subset selection", *Artificial Intelligence, Spec. Issue on Relevance*, vol. 97, pp. 273-324, 1997

[10] T. Butko, C. Nadeu, "A Fast One-Pass-Training Feature Selection Technique for GMM-based Acoustic Event Detection with Audio-Visual Data", in proc. *Interspeech*, 2010

A BIC-POISSON-SVM Segmentation System for the Albayzin'10 Audio Segmentation Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

Department of Signal Theory and Communications, Universidade de Vigo
plopez@gts.uvigo.es, ldocio@gts.uvigo.es, carmen@gts.uvigo.es

Abstract

In this paper, the system submitted by the UVigo-GTM for the Albayzin 2010 Audio Segmentation Evaluation is described. The task is performed in two different stages, a segmentation stage and a classification stage. The segmentation stage employs an approach called BIC-POISSON, which consists in a modification of the BIC algorithm oriented to reduce the false alarm rate in the acoustic change detection stage. The classification is performed with Support Vector Machines, using the segments obtained in the segmentation stage.

Index Terms: audio segmentation, BIC, poisson, GMM-UBM, SVM

1. Introduction

The increasing amount of multimedia information that is available nowadays makes it necessary to develop techniques to structure and classify this kind of information. These multimedia contents include music, video and spoken documents, so speech technologies become involved. In monitoring spoken documents it might be necessary to know the background acoustic condition in order to implement an online adaptation of the monitoring/recognition system to the acoustic conditions in order to improve the overall system performance. The work presented in this paper focuses in the segmentation of audio documents into homogeneous regions according to the background conditions.

Common audio and speech segmentation techniques are based on the Bayesian information criterion (BIC) algorithm [2] [4], which seems to be the fastest and the best performing algorithm when it is compared to techniques such as the adaptation and comparison of Gaussian mixture models (GMM) [3]. One of the problems encountered when segmenting with BIC algorithm is the high number of false alarms (i.e. change-points that are detected, although they do not exist), because it degrades the quality of the results.

The Albayzin 2010 Audio Segmentation Evaluation consists in the segmentation of audio broadcast news programs into homogeneous regions according to the backgroud conditions: clean speech, music, speech with background noise, speech with background music, and other. The segmentation strategy chosen by the UVigo-GTM group for this task was the BIC-Poisson [1] one. The basic idea in this approach is to discard change-points found by the BIC algorithm when they are likely to be false alarms by assuming that changes occurring in the audio stream constitute a Poisson process, so changes that have a low ΔBIC value will be discarded with a probability that follows a Poisson cumulative density function (cdf).

The classification stage is performed by using Support Vector Machines (SVMs), representing the audio segments as supervectors composed by the stacked means of Gaussian Mixture Models (GMMs). The one-against-one technique for multiclass classification with SVMs was employed in this work.

The structure of this paper is as follows: section 2 makes an overview of the whole audio segmentation strategy; in section 3 the segmentation algorithm and the BIC-Poisson technique are described; section 4 explains how the classification of the different segments was performed; section 5 describes the experimental framework used for the experiments and its purpose; section 6 shows the results of the whole system; and section 7 concludes the work and gives prospects on future work.

2. System overview

The developed system performs the segmentation of the data and the classification of the segments independently: first the data is segmented, and after that the segments are classified into the different classes. It is possible that, after classifying the segments, adjacent segments assigned to the same class appear, so these adjacent segments are merged, deleting the acoustic change-point between them.

The segmentation of the data is performed using an approach called BIC-Poisson [1], specially designed to reduce the number of false alarms in audio and speaker segmentation tasks. The classification is done by using Support Vector Machines (SVM). The whole procedure is summarized in figure 1.

3. Segmentation

3.1. Segmentation strategy

The segmentation strategy employed in these experiments is similar to the one described in [2]. It has a coarse segmentation stage where a window of data is segmented with a low resolution. If an acoustic change-point is found, a fine segmentation stage is performed, using a higher resolution. In the first stage, if there is no acoustic change-point, the window grows until it gets to a fixed maximum size, and then it slides. When a change-point is found, the window returns to its initial size and slides. In the second stage, a window of a fixed size is centered in the change-point found in the first stage and BIC is applied again using a higher resolution. If the hypothesis test indicates that there is a change-point in the window, this change-point is confirmed.

Subsequently, in the proposed segmentation technique a change-point rejection strategy is applied. The ΔBIC value is assessed in order to evaluate the significance of the change-point detected. if this is located above a given threshold no action is taken; otherwise the change-point is discarded with a certain probability.

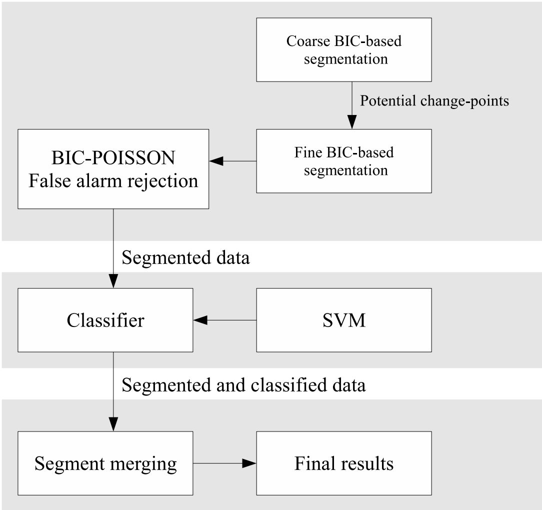


Figure 1: *Summary of the audio segmentation and classification system.*

3.2. BIC-Poisson: Poisson distribution-based change rejection

In this false alarm suppression strategy, it is assumed that the change-point instants follow a Poisson process. A Poisson process is an independent occurrence process where the number of occurrences in two disjoint time intervals is independent, the probability of an occurrence is proportional to the observed interval and occurrences are not simultaneous [5].

The speaker segmentation process fulfills those properties, since it is a process where arrivals (change-points) occur independently and in random instants. Poisson processes have a probability density function (pdf):

$$f(\mu, x) = \frac{e^{-\mu}\mu^x}{x!} \tag{1}$$

Its cumulative density function (cdf) is the sum of the probability density function in all points below a given value:

$$F(\mu, x) = \sum_{i=0}^{x} f(\mu, i) = \sum_{i=0}^{x} \frac{e^{-\mu}\mu^i}{i!} \tag{2}$$

The parameter μ, which represents the mean of the distribution, will in this case represent the expected number of changes.

The properties of the Poisson distribution will be used as follows: μ occurrences are expected over a given period of time, so initially a change is accepted with a very high probability. However, as the number of accepted changes increases and approaches or exceeds the expected number of changes, they are more likely to be rejected. This process is easily modeled using the cumulative density function $F(\mu, x)$ as a discard probability. The discard probability will be very low at first, but as the mean is approached or exceeded, it will steadily approach 1 (meaning that all the occurrences will be rejected).

Figure 2 shows graphically the increase of the discard probability with the number of accepted changes. As the amount of accepted changes reaches the value of 2μ approximately, the discard probability becomes close to one, i.e. almost every change-point will be discarded.

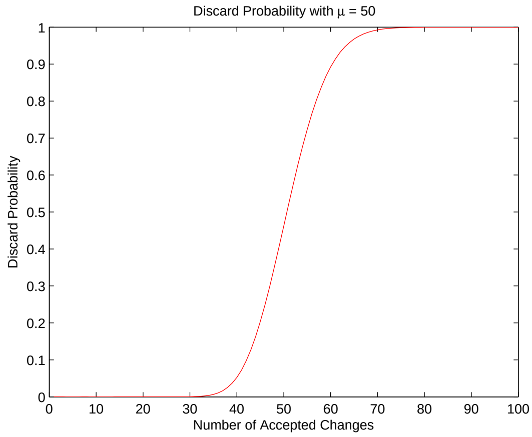


Figure 2: *Discard probability for the Poisson-based rejection algorithm.*

4. Classification

After segmenting the audio data, each audio segment has to be classified into one of the five pre-defined audio classes:

- Speech [sp]. Clean speech in studio from a close microphone.
- Music [mu]. Music is understood in a general sense.
- Speech with noise in background [sn]. Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).
- Speech with music in background [sm]. Overlapping of speech and music classes or speech with noise in background and music classes.
- Other [ot]. This class refers to any type of audio signal (including noises) that doesnt correspond to the other four classes.

The classification scheme chosen in this study is based on Support Vector Machines (SVM). The SVM is a supervised learning method which, given a set of labeled samples, apply a non-linear transform through a kernel function on the samples into a higher dimension space where the two classes are linearly separated under the margin maximization constraint.

The classification that has to be performed implies splitting the data into six different classes and SVMs, basically, can only solve binary classification problems. Nevertheless, it is possible to reduce a multiclass problem to a binary one. There are two main different approaches to solve this multiclass problem with SVMs: the one-against-one technique and the one-against-the-rest technique. To allow for multiclass classification the library for Support Vector Machines called libSVM [6], which was used in these experiments, uses the one-against-one technique by fitting all binary subclassifiers and finding the correct class by a voting mechanism.

Choosing accurate model parameters is very important to SVM training. Radial basis function (RBF) was used as kernel function in the SVMs.

$$K(x_i, x_j) = exp(-\sigma||x_i - x_j||^2) \qquad \sigma > 0$$

Model selection in this class of SVM involves two hyper-parameters: the penalty parameter C and the kernel width σ. The σ in the RBF kernel controls the shape of the kernel and C controls the tradeoff between margin maximization and error minimization. A grid-search on C and σ using 5 fold cross-validation is performed.

The SVM considers the sample data as points in a space of a given dimensionality, so a way to represent audio segments as sample data for the SVM training has to be chosen. In this case, first a Universal Background Model based on a Gaussian Mixture Model (GMM-UBM) of M mixture components is trained using data from all classes. Then, given an audio segment represented by N feature vectors of dimension D, the GMM-UBM is adapted using MAP adaptation to that audio segment [3]. By stacking the resultant means, a supervector of dimension $M \cdot D$ is obtained, being this supervector the representation of the audio segment as a point in a space of dimension $M \cdot D$.

5. Experimental framework

5.1. Database

The training and evaluation database consists of Catalan broadcast news data from the 3/24 TV channel that was recorded by the TALP Research Center from the UPC, and was annotated by Verbio Technologies. Its production took place in 2009 under the Tecnoparla research project, funded by the Generalitat de Catalunya. The Corporaci Catalana de Mitjans Audiovisuals, owner of the multimedia content, allows its use for technology research and development. The database, that includes around 87 hours of sound (24 files of approximately 4 hours long), was splitted into 2 parts: one part for training/development (2/3 of the total amount of data), and the other part for evaluation (the remaining 1/3).

The distribution of classes within the database is the following: Clean speech: 37

The 16 available files to perform the training/development of the segmentation system were splitted as follows:

- Sessions 1 to 8 and 10 to 15: training of the GMM-UBM and the multiclass SVM.
- Sessions 9 and 16: selection of the parameters that give the best performance. The parameters to select were μ, λ, M, σ and C.

After testing on the development data the selected parameters were: $\mu = 20.0$, $\lambda = 3.5$, $M = 64$, $\sigma = 0.0078$ and $C = 2$.

5.2. Metric

The metric used to evaluate the system performance is defined as a relative error averaged over all acoustic classes:

$$Error = average_i \frac{dur(miss_i)dur(fa_i)}{dur(ref_i)}$$

where

- $dur(miss_i)$ is the total duration of all deletion errors (misses) for the ith acoustic class.
- $dur(fa_i)$ is the total duration of all insertion errors (false alarms) for the ith acoustic class.
- $dur(ref_i)$ is the total duration of all the ith acoustic class instances according to the reference file.

It is worth noting that an incorrectly classified audio segment (a substitution) is computed both as a deletion error for one acoustic class and an insertion error for another. In the case when the system output is Other (non-Other) and the corresponding reference label is non-Other (Other), the audio segment is computed as a deletion (insertion) error only for the non-Other acoustic class.

A forgiveness collar of 1 sec (both + and -) will not be scored around each reference boundary. This accounts for both the inconsistent human annotation and the uncertainty about when an acoustic class begins/ends.

5.3. Acoustic features

The data to perform the evaluation is given as a set of waveforms, and these are going to be represented by 12 Mel-frequency Cepstral coefficients (MFCC), extracted using a 25 ms Hamming window at a rate of 10 ms per frame. In the segmentation stage, these features are augmented by 0-th order cepstral coefficient. In the classification stage, this 13-dimensional feature vector is augmented with first and second order dynamic coefficients resulting in a 39-dimensional feature vector. Cepstral mean and variance normalization is also applied.

6. Experimental results

Table 1 provides the results obtained by the system submitted by the UVigo-GTM research group. It can be observed that the error in pure music is low. Indeed the main cause of error is, by far, the confusion between speech with background music, speech with background noise and clean speech.

Table 1: *Audio segmentation results on the development (session9 and session16) and evaluation corpus.*

Corpus	Error				
	Tot	music	speech	speech-music	speech-noise
Dev	35.95	18.73	50.48	34.40	40.20
Eval	33.15	22.41	41.80	27.47	40.93

7. Discussion and future directions

The audio segmentation and classification system submitted to Albayzin 2010 Evaluation was described in this paper. The audio segmentation task focuses in the context of broadcast news. According to the results obtained by the proposed system on the evaluation data, the validity of the combination of the BIC-poisson based audio segmentation approach with the SVM-based multiclass classification is confirmed.

Future work will focus on combining the traditional short-term MFCC features with prosodic and other acoustic features in order to discriminate better between speech with background music, speech with background noise and clean speech. Related to the classification stage, future work will focus on the analysis of other strategies for multi-class SVM based classification.

8. Acknowledgements

The UVigo-GTM group would like to thank the organizers and coordinators of the Albayzin 2010 Audio Segmentation Evaluation for their help and for kindly providing the data corpus.

The work reported here was partially sponsored by the Spanish government and ERDF funds under the project TEC2009-14094-C04-04 “Búsqueda de Información en Contenidos Audiovisuales plurilingües” and by the Ministry of Science and Innovation under the FPI grant BES-2010-033358.

9. References

[1] P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, “Reducing the False Alarm Rate in a Speaker Segmentation System”, In preparation.

[2] M. Cettolo and M. Vescovi and R. Rizzi, “Evaluation of BIC-based algorithms for audio segmentation”, Computer Speech and Language, vol. 19, 147–170, April 2005

[3] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, January 2000.

[4] P. Delacourt and C. J. Wellekens, “DISTBIC: A speaker based segmentation for audio data indexing,” *Speech Communication*, vol. 32, pp. 111–126, September 2000.

[5] A. O. Allen, “Probability, Statistics, and Queueing Theory with Computer Science Applications”, 2nd ed. Academic Press, 1990.

[6] C.-C. Chang, C.-J. Lin, “LIBSVM: a Library for Support Vector Machines”, Online: `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

VIVOLAB-UZ Audio Segmentation System for Albayzín Evaluation 2010

Diego Castán, Alfonso Ortega, Carlos Vaquero, Antonio Miguel, Eduardo Lleida

Aragon Institute of Engineering Research (I3A)
University of Zaragoza (UZ)
(dcastan | ortega | cvaquero | amiguel | lleida)@unizar.es

Abstract

This paper presents a method for audio segmentation that separates broadcast news audio files into five acoustic classes for Albayzín Audio Segmentation 2010 [1]. The proposed system makes use of a presegmentation stage based on the Bayesian Information Criterion (BIC), a music/speech classifier based on a combination of GMMs and a binary decision tree, and finally a speech/speech with music/speech with noise classifier based on GMMs.

Index Terms: Audio segmentation, Bayesian Information Criterion, Gaussian Mixture Models, C4.5 Tree

1. Introduction

Segmentation plays an important role in audio processing applications, such as content-based audio retrieval recognition and classification, and audio database management. Audio segmentation is a process that divides an audio file into different classes. Each segment or clip should consist of a single class that is acoustically different from other classes of the audio file. A good segmentation should be able to delimit the boundaries between two classes to group segments into homogeneous classes.

There are many different tasks in audio segmentation. Several methods are focus on speech/music segmentation, speaker recognition or acoustic events detection. Albayzín 2010 Audio Segmentation database is composed of 87 hours of sound (24 files of approximately 4 hours long), will be splitted into 2 parts: for training/development (2/3 of the total amount of data), and testing (the remaining 1/3). This evaluation contains five classes:

1. Music (MU): Music is understood in a general sense.
2. Speech (SP): Clean speech in studio from a close microphone.
3. Speech with music in background (SM): Overlapping of speech and music classes or speech with noise in background and music classes.
4. Speech with noise in background (SN): Speech which is not recorded in studio conditions, or it is overlapped with some type of noise (applause, traffic noise, etc.), or includes several simultaneous voices (for instance, synchronous translation).
5. Other (OT): This class refers to any type of audio signal (including noises) that doesnât correspond to the other four classes.

This work has been partially supported by the national project TIN2008-06856-C05-04

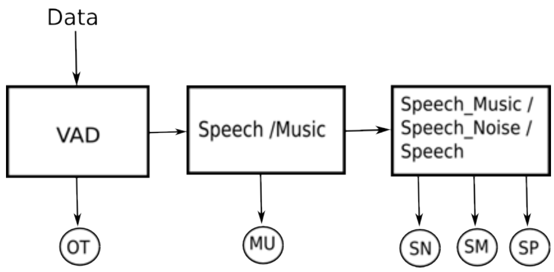


Figure 1: Classification System

The distribution of classes within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3%. The audio signals are provided in pcm format, mono, 16 bit resolution, and sampling frequency 16 kHz.

The proposed metric is inspired on the NIST metric for speaker diarization. The metric is defined as a relative error averaged over all acoustic classes:

$$Error = average_i \frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}$$

The remainder of the paper is organized as follows: section 2 discusses the set of features and masks used in the system, section 3 describes the classification algorithm, section 4 provides an evaluation of the system and is followed by the conclusion in section 5.

2. Features and masks

The features used in the proposed system are 13 MFCCs plus their derivative and second derivative (delta and delta delta).
In order to obtain more discriminative speech-music classification system, the mean, the variance and the skewness of the first MFCC is calculated.

A presegmentation is made by using a sliding window BIC based algorithm to delimit boundaries between heterogeneous segments. At the end, the segmentation system proposed in this work will assign just with one label for each segment obtained by this presegmentation. Long silence segments (with more than 3.5 s) are detected and marked as Others since those segments do not correspond to any of the predefined classes.

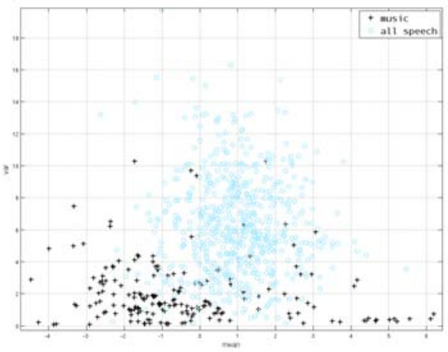


Figure 2: Behavior of the C1-Average VS C1-Variance with speech signal (circles) and music signal (crosses)

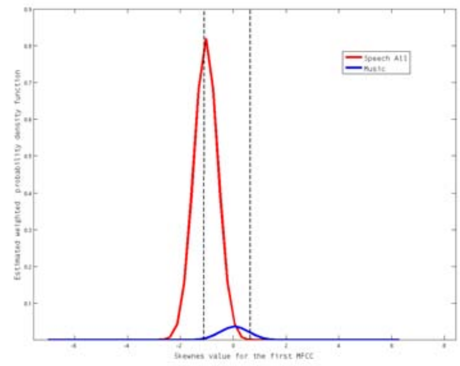


Figure 3: Estimated weighted probability density function of C1-Skewness

3. System Description

The classification of the segments reported by the BIC based segmentation system is made in two steps: a music/speech classification subsystem and a speech/speech with noise/speech with music classification subsystem presented in Fig. 1.

3.1. Music/Speech Classification

This approach is based on two Gaussian Mixture Models (GMM) with 128 components each model. The first model is trained with the music class and the second model is trained with the speech, speech with music and speech with noise classes. The accumulated likelihoods have a penalty based on the a priori knowledge of the distributions of classes within the database.

The GMMs decisions are combined with a binary decision tree trained with the C4.5 algorithm [3]. The tree is trained with the mean, the variance and the skewness calculated every 3 seconds of the first cepstral coefficient that provides information about the distribution of the energy between low and high frequency in a frame. We can see a scatter plot of the C1-mean and the C1-variance in Fig. 2. Also, in Fig. 3 the estimated probability density function of C1-skewness, weighted by the prior probability of speech and music is shown for both classes. It has been evaluated considering constant length non-overlapping windows of 3 seconds. Along with the pdf two lines are plotted dividing the x axis into 3 regions: the one on the right side, contains the points representing the segments that would be correctly classified as music. The one on the left side, contains the points representing the segments that would be correctly classified as speech. In between, the points representing the segments that would need the help of other features to decide if they correspond whether to speech, whether to music.

The combination of both classifiers is made by weighted addition of the GMM likelihood and the tree error prediction of the music class frame by frame.

3.2. Speech, Speech With Music and Speech With Noise Classification

The Speech, Speech with Music and Speech with Noise Classification approach is based on Gaussian Mixture Models (GMM) with 128 components. The first model is trained with the speech class, the second model is trained with the speech with music and the third model is trained with speech with noise class. As in the previous music/speech classifier, these likelihoods have a penalty based on the a priori knowledge of the distributions of classes within the database.

4. Results

The system has been tested over 8 files (around 32 hours of audio material). These results are presented in table 1.

Class	Accuracy
MU	28.14%
SP	51.06%
SM	48.78%
SN	51.51%
Total	44.87%

Table 1: Results on the test files

5. Future lines

Further work must be done to improve the performance of the proposed system along the following lines of research:

- Seek new ways of modelling the temporal behaviour of the detection problem under study. Along this line we think about improving the boundaries estimation substituting the BIC based approach by an HMM based solution.
- Study new ways of representing music and speech in a more dicriminative way. Provided that informal tests seem to show that using statistical moments analysis makes possible to segregate music from speech as shown in Fig. 2 and 3.

- Keep working on the hierarchical structure of our system extending it also for speech classes, and try to explode in a more discriminative way the information extracted from the binary tree classifier in order to fuse it with the information coming from the GMMs.

6. References

[1] Butko, T , "Albayzín Evaluations 2010: Audio Segmentation",

[2] Quatieri T.F., "Discrete-Time Speech Signal Processing", Prentice-Hall, Englewood Cliffs, NJ,USA, 2001.

[3] Quinlan, R., "C4.5: Programs for Machine Learning", in Morgan Kaufmann Publishers Springer Netherlands, 1993.