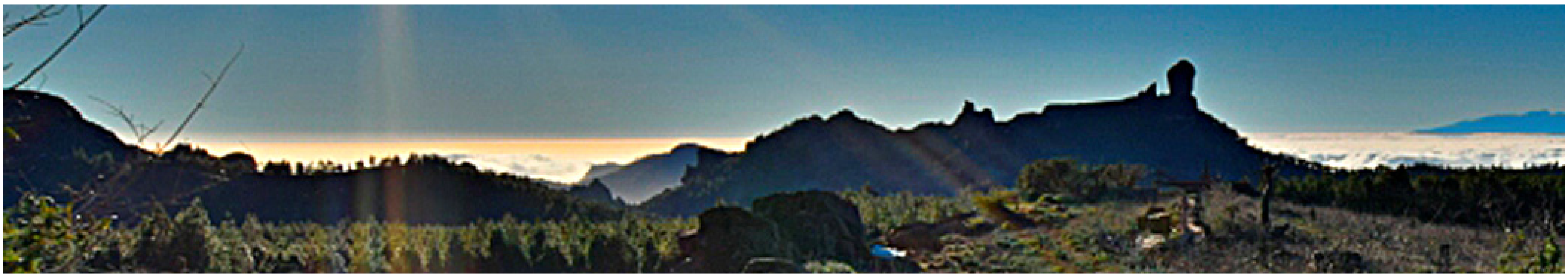
iberSPEECH2014

November 19-21 2014, Las Palmas de Gran Canaria



# Conference Proceedings

# iberSPEECH 2014

## VIII Jornadas en Tecnologías del Habla

## and IV Iberian SLTech Workshop

**Escuela de Ingeniería en Telecomunicación y Electrónica**
**Universidad de Las Palmas de Gran Canaria**
**SPAIN**

**Organized by:**

IDeTIC
Instituto para el Desarrollo Tecnológico y la Innovación en Comunicaciones

*División de Procesado Digital de Señal*

RTTH

*Spanish Thematic Network on Speech Technology (RTTH)*

ISCA

IL SIG

**Supported by:**

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

ULPGC 25 ANIVERSARIO 1989-2014

**Selected papers published by:**

Springer
Lecture Notes in Artificial Intelligence
Lecture Notes in Computer Science

LNAI Vol. 8854

PROCEEDINGS

# iberSPEECH2014

# VII Jornadas en Tecnología del Habla
and
# III Iberian SLTech Workshop

November 19-21, 2014
Escuela de Ingeniería enTelecomunicación y Electrónica
Universidad de Las Palmas de Gran Canaria
Las Palmas de Gran Canaria, Spain

*Editors: Juan Luis Navarro Mesa*
*Pedro Quintana Morales*
*Alfonso Ortega*
*António Teixeira*
*Eduardo Hernández Pérez*
*Antonio Ravelo García*
*Iván Guerra Moreno*
*Manuel Medina Molina*
*Sofía Martín González*

**Organized by:**

IDeTIC

***División de Procesado Digital de Señal***

RTTH

***Spanish Thematic Network on Speech Technology (RTTH)***

**Supported by:**

UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

ULPGC 25 ANIVERSARIO 1989-2014

ISBN: 978-84-617-2862-6

http://iberspeech2014.ulpgc.es

# WELCOME MESSAGE

The Spanish Thematic Network on Speech Technology (RTTH) and the ISCA-Special Interest Group on Iberian Languages (SIG-IL) are pleased to present the selected papers of IberSpeech 2014, Joint VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop, held in Las Palmas de Gran Canaria, Spain, on November 19-21. The Organizing Committee of IberSpeech believes and trusts that we have achieved the quality that the researchers in Advances in Speech and Language Technologies for Iberian Languages value. To ensure this quality, each article has been reviewed at least by three members of the Scientific Review Committee, who have provided feedback to improve the final version of the articles in this book.

The conference has become mature as different editions have been organized, starting in Vigo 2010 with FALA and continuing in Madrid 2012 with the new denomination: Iberspeech. This new edition becomes a step further in the support of researchers in Iberian Languages. IberSpeech is a joint event resulting from the merging of two conferences, the ”Jornadas en Tecnología del Habla” and the Iberian SLTech Workshop. The first has been organized by the “Red Temática en Tecnologías del Habla” (Spanish Speech Technology Thematic Network, http://www.rthabla.es) since 2000. This network was created in 1999 and currently includes over 200 researchers and 30 research groups in speech technology all over Spain. And the second, has been organized by the Special Interest Group on Iberian Languages (SIG-IL, http://www.il-sig.org/) of the International Speech Communication Association (ISCA), The Iberian SLTech Workshop had its first edition in Porto Salvo, Portugal, in 2009.

As a result, IberSpeech is one of the most important research meetings in the field of speech and language processing focusing on Iberian Languages, attracting many researchers (about 140 in the 2014 edition), mainly from Spain, Portugal, and from other Iberian-speaking countries in Latin America. We have also attracted the interest of several research groups from all around the world, including China, United Kingdom, France, Japan, Hungary, Israel, Norway, Czech Republic, and Germany.

Although the main focus is on Iberian Languages and the Iberian region, the conference is not restricted to them. Proof of this are the ALBAYZIN Technology Competitive Evaluations, organized in conjunction with the conference, which in this edition have attracted the interest of several research groups. The ALBAYZIN Technology Competitive Evaluations have been organized alongside with the conference since 2006, promoting the fair and transparent comparison of technology in different fields related to speech and language technology. In this edition we have two different evaluations: Audio Segmentation and Search on Speech. The organization of each one of these evaluations requires preparing development and test data, providing data along with a clear set of rules to the participants, and gathering and comparing results from participants. This organization has been carried out by different groups of researchers and is crucial for the success in participation that we are envisaging. Although results from the evaluations cannot be included in this volume due to timing restrictions, we would like to express our gratitude to the organizers and also to the participants in the evaluations.

We have had 60 submitted papers and, after a strict peer-reviewing process, only 29 have been selected for publication in this volume of the Springer Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence. This selection has been based on the scores and comments provided by our Scientific Review Committee, which includes over 79 researchers from different institutions mainly from Spain, Portugal, Latin America, USA, UK, Hungary and Czech Republic to which we also would like to express our deepest gratitude. Each article has been reviewed by at least three different reviewers and authors have had time to address the comments before submitting the camera-ready paper. The articles have been organized into four different topics:

- Speech Production, Analysis, Coding and Synthesis
- Speaker and Language Characterization
- Automatic Speech Recognition
- Speech and Language Technologies in Different Application Fields

Besides the excellent research articles included in this volume, the conference had the pleasure of having two extraordinary keynote speakers: Dr. Pedro Gómez Vilda (Departamento de Arquitectura y Tecnología de Sistemas Informáticos de la Universidad Politécnica de Madrid, Spain) and Dr. Roger K. Moore (Department of Computer Science University of Sheffield, UK).

We would also like to thank Springer, and in particular to Alfred Hoffmann, for the possibility of publishing this volume, his suggestions in order to increase the spread of the international scope of IberSpeech 2014, his help and great work in preparing it.

Finally, we would like to thank all those whose effort has made possible this conference, the members of the local organizing committee, the technical and program chairs, the reviewers and so many people who put their best to achieve a successful conference.

Juan Luis Navarro Mesa
Alfonso Ortega Giménez
António Teixeira

IberSpeech 2014 Co-Chairs

# ORGANIZATION

## General Chairs

| | |
|---|---|
| Juan Luis Navarro Mesa | Universidad de Las Palmas de Gran Canaria |
| Alfonso Ortega Giménez | Universidad de Zaragoza |
| António Teixeira | Universidade de Aveiro |

## Technical Chairs

| | |
|---|---|
| Carmen García Mateo | Universidade de Vigo |
| Doroteo Torre Toledano | Universidad Autónoma de Madrid |
| Rubén San-Segundo Hernández | Universidad Politécnica de Madrid |

## Program Chairs

| | |
|---|---|
| Eduardo Hernández Pérez | Universidad de Las Palmas de Gran Canaria |
| Pedro Quintana Morales | Universidad de Las Palmas de Gran Canaria |
| Jesús Alonso Hernández | Universidad de Las Palmas de Gran Canaria |

## Publication Chair

| | |
|---|---|
| Sofía Martín González | Universidad de Las Palmas de Gran Canaria |

## Demos Chairs

| | |
|---|---|
| Javier Hernando Pericas | Universidad Politécnica de Cataluña |
| Rubén San-Segundo Hernández | Universidad Politécnica de Madrid |
| Daniela Braga | Voicebox Technologies |

## Awards Chairs

| | |
|---|---|
| Javier Hernando Pericas | Universidad Politécnica de Cataluña |
| Inma Hernáez Rioja | Universidad del País Vasco |
| Daniela Braga | Voicebox Technologies |

## Evaluation Chairs

| | |
|---|---|
| Antonio Ravelo García | Universidad de Las Palmas de Gran Canaria |
| Carlos Travieso González | Universidad de Las Palmas de Gran Canaria |

## Local Organizing Committee

### Universidad de Las Palmas de Gran Canaria

Jesús B. Alonso Hernández
Juan M. Caballero Suárez
Fidel Cabrera Quintero
Elena García Quevedo
Iván D. Guerra Moreno
Eduardo Hernández Pérez
Sofía I. Martín González
Manuel M. Medina Molina
Juan Luis Navarro Mesa
Pedro J. Quintana Morales
Antonio G. Ravelo García
Carlos Travieso González

## Albayzin Commitee

| | |
|---|---|
| Juan Luis Navarro Mesa | Universidad de Las Palmas de Gran Canaria |
| Alfonso Ortega Giménez | Universidad de Zaragoza |
| António Teixeira | Universidade de Aveiro |
| Rubén San-Segundo Hernández | Universidad Politécnica de Madrid |

## Scientific Review Committee

| | |
|---|---|
| Alberto Abad | L2F/Spoken Language Systems Laboratory |
| Jesús B. Alonso Hernández | Universidad de Las Palmas de Gran Canaria |
| Olatz Arregi Uriarte | Euskal Herriko Unibertsitatea |
| Plínio Barbosa | Universidade Estadual de Campinas |
| Daniela Braga | Voicebox |
| Juan M. Caballero Suárez | Universidad de Las Palmas de Gran Canaria |
| Fidel Cabrera Quintero | Universidad de Las Palmas de Gran Canaria |
| Valentín Cardeñoso Payo | Universidad de Valladolid |
| Paula Carvalho | Universidade de Lisboa |
| María José Castro Bleda | Universitat Politécnica de Valencia |
| Jan Cernocky | BTU (Brno) |
| Chin Hui Lee | Georgia Tech |
| Ricardo de Córdoba Herralde | Universidad Politécnica de Madrid |
| Carmen de la Mota Górriz | Universitat Autònoma de Barcelona |
| Laura Docío Fernández | Universidade de Vigo |
| Daniel Erro Eslava | Euskal Herriko Unibertsitatea |
| David Escudero Mancebo | Universidad de Valladolid |
| Rubén Fernández Pozo | Universidad Politécnica de Madrid |
| Javier Ferreiros López | Universidad Politécnica de Madrid |
| Julián Fierrez Aguilar | Universidad Autónoma de Madrid |
| Javier Franco Pedroso | Universidad Autónoma de Madrid |
| Ascensión Gallardo Antolín | Universidad Carlos III de Madrid |
| Carmen García Mateo | Universidade de Vigo |
| Elena García Quevedo | Universidad de Las Palmas de Gran Canaria |
| Juan Ignacio Godino Llorente | Universidad Politécnica de Madrid |
| Pedro Gómez Vilda | Universidad Politécnica de Madrid |
| Iván D. Guerra Moreno | Universidad de Las Palmas de Gran Canaria |
| Annika Hamalainen | Microsoft (MLDC) |
| Inma Hernáez Rioja | Euskal Herriko Unibertsitatea |
| Eduardo Hernández Pérez | Universidad de Las Palmas de Gran Canaria |
| Francisco Javier Hernando Pericas | Universitat Politècnica de Catalunya |
| Lluís Felip Hurtado Oliver | Universitat Politécnica de Valencia |
| Eduardo Lleida Solano | Universidad de Zaragoza |
| María Teresa López Soto | Universidad de Sevilla |
| Ramón López-Cózar Delgado | Universidad de Granada |
| Jordi Luque Serrano | Telefónica I+D |
| José B. Mariño Acebal | Universitat Politècnica de Catalunya |
| Sofía I. Martín González | Universidad de Las Palmas de Gran Canaria |
| Carlos David Martínez Hinarejos | Universitat Politécnica de Valencia |
| Manuel M. Medina Molina | Universidad de Las Palmas de Gran Canaria |
| Hugo Meinedo | Microsoft (MLDC) |
| Helena Moniz | INESC |
| Juan Manuel Montero Martínez | Universidad Politécnica de Madrid |

| | |
|---|---|
| Nicolas Morales Mombiela | Nuance |
| Antonio Moreno Sandoval | Universidad Autónoma de Madrid |
| Climent Nadeu Camprubi | Universitat Politècnica de Catalunya |
| Juan Luis Navarro Mesa | Universidad de Las Palmas de Gran Canaria |
| Eva Navas Cordón | Euskal Herriko Unibertsitatea |
| Géza Nemeth | BME, Hungary |
| Juan Nolazco Flores | Tecnológico de Monterrey, Mexico |
| Alfonso Ortega Giménez | Universidad de Zaragoza |
| Antonio Miguel Peinado Herreros | Universidad de Granada |
| Carmen Peláez Moreno | Universidad Carlos III de Madrid |
| José Luis Pérez Córdoba | Universidad de Granada |
| Ferrán Pla Santamaría | Universitat Politécnica de Valencia |
| Paulo Quaresma | Universidade de Évora |
| Pedro J. Quintana Morales | Universidad de Las Palmas de Gran Canaria |
| Daniel Ramos Castro | Universidad Autónoma de Madrid |
| Andreia Rauber | Universidade Católica de Pelotas |
| Antonio G. Ravelo García | Universidad de Las Palmas de Gran Canaria |
| José Adrián Rodríguez Fonollosa | Universitat Politècnica de Catalunya |
| Eduardo Rodríguez Banga | Universidade de Vigo |
| Luis Javier Rodríguez Fuentes | Microsoft (MLDC) |
| Joan Andreu Sánchez Peiró | Universitat Politécnica de Valencia |
| Emilio Sanchís Arnal | Universitat Politécnica de Valencia |
| Rubén San-Segundo Hernández | Universidad Politécnica de Madrid |
| Diana Santos | University of Oslo |
| Kepa Sarasola Gabiola | Euskal Herriko Unibertsitatea |
| Encarnación Segarra Soriano | Universitat Politécnica de Valencia |
| Mário Silva | Universidade de Lisboa |
| Alberto Simões | ESEIG/IPP |
| Richard Stern | Carnegie Mellon University |
| António Teixeira | Universidade de Aveiro |
| Javier Tejedor Noguerales | Universidad de Alcalá de Henares |
| Doroteo Torre Toledano | Universidad Autónoma de Madrid |
| Isabel Trancoso | IST/University of Lisbon & INESC-ID |
| Carlos Travieso González | Universidad de Las Palmas de Gran Canaria |
| María Amparo Varona Fernández | Euskal Herriko Unibertsitatea |
| Aline Villavicencio | Federal University of Rio Grande do Sul |

# IberSPEECH 2014 Conference Program

| **Wednesday 19 November** | **Thursday 20 November** | **Friday 21 November** |
|---|---|---|
| 8:00 On-site registration<br>(Hall Building A) | | |
| **8:30 Opening Ceremony**<br>(Building A, Salón de Actos) | 8:45 On-site registration<br>(Hall Building A) | 8:45 On-site registration<br>(Hall Building A) |
| **9:00 O1 - Oral Session**<br>Speaker and Language Characterization<br>Chair: Santiago R. Piqueras<br>(Building A, Salón de Actos) | **9:00 O3 - Oral Session**<br>Automatic Speech Recognition I<br>Chair: Doroteo T. Toledano<br>(Building A, Salón de Actos) | **9:00 O5 - Oral Session**<br>Automatic Speech Recognition II<br>Chair: António Teixeira<br>(Building A, Salón de Actos) |
| 10:40 Coffee break | 10:40 Coffee break | 10:40 Coffee break |
| **11:00 Keynote Talk**<br>Roger K. Moore<br>Chair: Eduardo Lleida<br>(Building A, Salón de Actos) | **11:00 P2 - Poster Session**<br>Chair: Alberto Simões<br>(Hall Building A) | **11:00 Keynote Talk**<br>Pedro Gómez Vilda<br>Chair: Jesús B. Alonso<br>(Building A, Salón de Actos) |
| **12:00 P1 - Poster Session**<br>Chair: Alberto Abad<br>(Hall Building A) | | **12:00 O6 - Oral Session**<br>Speech and Language Technologies in Different Application Fields I<br>Chair: Carmen García<br>(Building A, Salón de Actos) |
| 13:30 Lunch | 13:30 Lunch | **13:40 Closing Ceremony** |
| **15:00 O2 - Oral Session**<br>Speech Production, Analysis, Coding and Synthesis<br>Chair: Javier Tejedor<br>(Building A, Salón de Actos) | **15:00 O4 - Oral Session**<br>Speech and Language Technologies in Different Application Fields II<br>Chair: Emilio Sanchís<br>(Building A, Salón de Actos) | |
| 16:40 Coffee break | 16:40 Coffee break | |
| **17:00 Albayzin Evaluations**<br>Chair: Inma Hernáez<br>(Building A, Salón de Actos & Hall) | **17:00 Projects, Demos and Thesis Session**<br>Chair: Javier Hernando<br>(Hall Building A) | |
| **18:30 RTTH Assembly**<br>(Building A, Salón de Actos) | **18:30 Social activities** | |
| **19:30 Welcome Reception** | | |
| | **21.00 Gala Dinner** | |

# Sessions included in Springer LNAI 8854

## O1 - Oral Session: Speaker and Language Characterization

Unsupervised Accent Modeling for Language Identification
*David Martínez, Jesús Villalba, Eduardo Lleida, Alfonso Ortega*

Global speaker clustering towards optimal stopping criterion in binary key speaker diarization
*Héctor Delgado, Xavier Anguera, Corinne Fredouille, and Javier Serrano*

Unsupervised Training of PLDA with Variational Bayes
*Jesús Villalba and Eduardo Lleida*

On the Use of Convolutional Neural Networks in Pairwise Language Recognition
*Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Ruben Zazo, Daniel Ramos and Joaquin Gonzalez-Rodriguez*

Global Impostor Selection for DBNs in Multi-Session i-Vector Speaker Recognition
*Omid Ghahabi and Javier Hernando*

## O2 - Oral Session: Speech Production, Analysis, Coding and Synthesis

Analysis and synthesis of emotional speech in Spanish for the chat domain
*Yesika Laplaza, Juan María Garrido*

Developing a Basque TTS for the Navarro-Lapurdian dialect
*Eva Navas, Inma Hernaez, Daniel Erro, Jasone Salaberria, Beñat Oyharçabal and Manuel Padilla*

Fine Vocoder Tuning for HMM-Based Speech Synthesis: Effect of the Analysis Window Length
*Agustin Alonso, Daniel Erro, Eva Navas, Inma Hernaez*

Quantitative Analysis of /l/ Production from RT-MRI: First Results
*Samuel Silva, Paula Martins, Catarina Oliveira and António Teixeira*

Statistical text-to-speech synthesis of Spanish sub-titles
*S. Piqueras, M. A. del-Agua, A. Giménez, J. Civera and A. Juan*

## O3 - Oral Session: Automatic Speech Recognition I

Phoneme-Lattice to Phoneme-Sequence Matching Algorithm based on Dynamic Programming
*Ciro Gracia, Xavier Anguera, Jordi Luque and Ittai Artzi*

Deep Maxout Networks applied to Noise-Robust Speech Recognition
*F. de-la-Calle-Silos, A. Gallardo-Antolín, and C. Peláez-Moreno*

A Deep Neural Network Approach for Missing-Data Mask Estimation on Dual-Microphone Smartphones: Application to Noise-Robust Speech Recognition
*Iván López-Espejo, José A. González, Angel M. Gomez and Antonio M. Peinado*

Language Model Adaptation for Lecture Transcription by Document Retrieval
*A. Martínez-Villaronga, M. A. del-Agua, J. A. Silvestre-Cerdà, J. Andrés-Ferrer, and A. Juan*

Articulatory Feature Extraction from Voice and their Impact on Hybrid Acoustic Models
*Jorge LLombart, Antonio Miguel and Eduardo Lleida*

## O4 - Oral Session: Speech and Language Technologies in Different Application Fields I

Speech watermarking based on coding of the harmonic phase
*Inma Hernaez, Ibon Saratxaga, Jianpei Ye, Jon Sanchez, Daniel Erro and Eva Navas*

Bootstrapping a Portuguese WordNet from Galician, Spanish and English wordnets
*Alberto Simões and Xavier Gómez Guinovart*

ATVS-CSLT-HCTLab System for NIST 2013 Open Keyword Search Evaluation
*Javier Tejedor, Doroteo T. Toledano, and Dong Wang*

The transLectures-UPV toolkit
*M. A. del-Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan*

The AhoSR automatic speech recognition system
*Igor Odriozola, Luis Serrano, Inma Hernaez, and Eva Navas*

## O5 - Oral Session: Automatic Speech Recognition II

Flexible Stand-Alone Keyword Recognition Application Using Dynamic Time Warping
*Miquel Ferrarons, Xavier Anguera and Jordi Luque*

Confidence Measures in Automatic Speech Recognition Systems for Error Detection in Restricted Domains
*Julia Olcoz, Alfonso Ortega, Antonio Miguel and Eduardo Lleida*

Recognition of distant voice commands for home applications in Portuguese
*Miguel Matos, Alberto Abad, Ramón Astudillo, and Isabel Trancoso*

CVX-optimized Beamforming and Vector Taylor Series Compensation with German ASR employing Star-shaped Microphone Array
*Juan A. Morales-Cordovilla, Hannes Pessentheiner, Martín Hagmueller, José A. González and Gernot Kubin*

## O6 - Oral Session: Speech and Language Technologies in Different Application Fields II

Assessing the Applicability of Surface EMG to Tongue Gesture Detection,
*João Freitas, Samuel Silva, António Teixeira and Miguel Sales Dias*

Towards Cross-lingual Emotion Transplantation
*Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Junichi Yamagishi, and Juan M. Montero.*

A preliminary study of Acoustic Events Classification With Factor Analysis in Meeting Rooms
*Diego Castán, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida*

A Spoken Language Database for Research on Moderate Cognitive Impairment: Design and Preliminary Analysis
*Fernando Espinoza-Cuadros, Marlene A. Garcia-Zamora, Dania Torres-Boza, Carlos A. Ferrer-Riesgo, Ana Montero-Benavides, Eduardo Gonzalez-Moreira, Luis A. Hernandez-Gómez.*

Towards Customized Automatic Segmentation of Subtitles
*Aitor Álvarez, Haritz Arzelus and Thierry Etchegoyhen*

# TABLE OF CONTENTS

## Keynote Talks

## P1 - Poster Session

## P2 - Poster Session

## ALBAYZIN Evaluation Special Session

### Audio Segmentation Evaluation

### Search-On-Speech Evaluation

## Projects, Demos and Thesis Special Session

### Projects

## Demos

## Ph.D. Thesis

# Keynote Talk

Roger K. Moore

**Title: Towards Spoken Language Interaction with 'Intelligent' Systems: where are we, and what should we do next?**

### Abstract:

Over the past thirty or so years, the field of spoken language processing has made impressive progress from simple laboratory demonstrations to mainstream consumer products. However, the limited capabilities of commercial applications such as Siri highlight the fact that there is still some way to go before we are capable of creating Autonomous Social Agents that are truly capable of conversing effectively with their human counterparts in real-world situations. This talk will address the fundamental issues facing spoken language processing, and will highlight the need to go beyond the current fashion for using machine learning in a more-or-less blind attempt to train static models on ecologically unrealistic amounts of unrepresentative training data. Rather, the talk will focus on critical developments outside the field of speech and language - particularly in the neurosciences and in cognitive robotics - and will show how insights into the behaviour of living systems in general and human beings in particular could have a direct impact on the next generation of spoken language systems.

### Speaker Bio:

Prof. Roger K. Moore studied Computer and Communications Engineering at the University of Essex and was awarded the B.A. (Hons.) degree in 1973. He subsequently received the M.Sc. and Ph.D. degrees from the same university in 1975 and 1977 respectively, both theses being on the topic of automatic speech recognition. After a period of post-doctoral research in the Phonetics Department at University College London, Prof. Moore was head-hunted in 1980 to establish a speech recognition research team at the Royal Signals and Radar Establishment (RSRE) in Malvern.

In 1985 Prof. Moore became head of the newly created 'Speech Research Unit' (SRU) and subsequently rose to the position of Senior Fellow (Deputy Chief Scientific Officer - Individual Merit) in the 'Defence and Evaluation Research Agency' (DERA). Following the privatisation of the SRU in 1999, Prof. Moore continued to provide the technical lead as Chief Scientific Officer at 20/20 Speech Ltd. (now Aurix Ltd.) - a joint venture company between DERA (now QinetiQ) and NXT plc. In 2004 Prof. Moore was appointed Professor of Spoken Language Processing in the 'Speech and Hearing' Research Group (SPandH) at Sheffield University, where he is pioneering research that is aimed at developing computational models of spoken language processing by both mind and machine.

Prof. Moore is currently working on a unified theory of spoken language processing in the general area of 'Cognitive Informatics' called 'PRESENCE' (PREdictive SENsorimotor Control and Emulation). PRESENCE weaves together accounts from a wide variety of different disciplines concerned with the behaviour of living systems - many of them outside the normal realms of spoken language - and compiles them into a new framework that is intended to breathe life into a new generation of research into spoken language processing.

Prof. Moore has authored and co-authored over 150 scientific publications in the general area of speech technology applications, algorithms and assessment. He is a Fellow of the UK Institute of Acoustics, a Visiting Professor in the Department of Phonetics and Linguistics at University College London and a Visiting Professor at the Bristol Robotics Laboratory. He is Editor-in-Chief of 'Computer Speech and Language' and a member of the Editorial/Advisory boards for 'Speech Communication', 'Languages' and the 'International Journal of Cognitive Informatics and Natural Intelligence' (IJCiNi). He is past Chairman of the 'European Expert Advisory Group on Language Engineering Standards' (EAGLES) working party on spoken language resources, and Editor of the 'Handbook of Standards and Resources for Spoken Language Systems'.

Prof. Moore served as President of the 'European Speech Communication Association' (ESCA) and the 'International Speech Communication Association' (ISCA) from 1997 to 2001, and as President of the Permanent Council of the 'International Conferences on Spoken Language Processing' (PC-ICSLP) from 1996 to 2001. During this period he pioneered the internationalisation of ESCA, the integration of the EUROSPEECH and ICSLP conferences into an annual INTERSPEECH conference, and chaired the joint ISCA/PC-ICSLP working party which drew up the detailed recommendations for the merger.

In 1994 Prof. Moore was awarded the prestigious UK Institute of Acoustics Tyndall medal for “*distinguished work in the field of speech research and technology*” and in 1999 he was presented with the NATO RTO Scientific Achievement Award for “*repeated contribution in scientific and technological cooperation*”. In 2008 he was elected as one of the first ISCA Fellows "*in recognition of his applications of human speech perception and production models to speech technologies and his service to ISCA as President*".

Prof. Moore was General Chair for INTERSPEECH 2009.

# Keynote Talk

## Pedro Gómez Vilda

**Title: Speech as a Vehicular Tool for Neurological Disease Monitoring**

**Abstract:**

Modern Statistical Signal Processing and Machine Learning Techniques are opening a new research area of most relevance for Speech Technologies, which is the field of medical applications. Organic Larynx Pathology Detection and Grading is being successfully accomplished thanks to the advances produced during the last decade, and is becoming a reality nowadays. Neurological Disease Monitoring and Assessment is one of the emerging fields for its interest in the years to come, especially in relation with Neurodegenerative Diseases as Parkinson's, Alzheimer's, Amyotrophic Lateral Sclerosis, or other non AD Aging Dementias, among others. Neuromotor and/or cognitive degeneration behind these diseases need a systemic neuromechanical description, in terms of the different physiological organs involved in speech production, mainly in the laryngeal, naso-pharyngeal and oral subsystems. Possible strategies to collect observable acoustic correlates from the speech signal, in reference to specific biomechanical systems in larynx (vocal folds), pharynx (velopharyngeal switch) and mouth (lingual complex, jaw, lips), are described. Methodologies to associate acoustic correlates to neuromotor and neurocognitive activity by means of different Statistical Pattern Recognition Techniques are also commented. Results from different on-going studies are to be presented and discussed.

**Speaker Bio:**

Dr. Pedro Gómez Vilda was born at Burgo de Osma (Soria), Spain. He received the degrees of Communications Engineer (MSc. level), Universidad Politécnica de Madrid (1978), and Ph.D. in Computer Science, Universidad Politécnica de Madrid (1983). His professional and academic activities can be summarized as 1976-77: Scholarship from Nuclear Studies Center, Nuclear Energy Board, Spain; 1977-78: I+D Engineer, NORTRON Electronics; 1978-82: Assistant Teacher; 1982-88, and 1988-till now: Full Professor, Facultad de Informática, Universidad Politécnica de Madrid. His research lines are in Signal Processing, Speech Recognition, Biomechanical System Modeling, Bioengineering, Bioinformatics, Pattern Recognition, Neural Networks, Speech Perception and Production, Neuromorphic Brain Modeling, Forensic Sciences, and Neurological Disease Monitoring. Prof. Gómez-Vilda is author or co-author of 290 publications, including book chapters and journal articles with international referencing in ISBN and/or ISSN, and 105 conferences and lectures in different institutions and scenarios. He is currently Head of the Research Group on Informatics Applied to Signal and Image Processing, and director of the Neuromorphic Speech Processing Lab, Center for Biomedical Technology, Universidad Politécnica de Madrid. He is member of IEEE, Life Sciences Systems and Applications Technical Committee, International Speech Communication Association (ISCA), and European Association of Signal Processing (EURASIP). He is also scientific reviewer of IEEE Transactions on Circuits and Sistems, Neural Networks, Speech and Audio and Signal Processing, Speech Communication, the Journal of the Acoustical Society of America, Neurocomputing, Cognitive Computation, Computers in Medicine and Biology, Biomedical Signal Processing and Control, Electronic Letters, and Signal Processing Letters. He has been also Professeur Invité par la División de l'Enseignement Superieur et la Recherche, France, Enseignant-Chercheur invité par l'Université de Cergy-Pontoise, France, Gastwissenschaftler (Invited Professor) an der Universität Regensburg, Freeistaat Bayern, Germany, Honorary Professor of Technical University of Cluj-Napoca, Romania, and Doctor Honoris Causa by Technical University of Cluj-Napoca, Romania. He is co-author of three patents in Spain and the USA, and founding partner and scientific director of the start-up BioMetroSoft SL (www.biometrosoft.com), created in 2011 from a contest of ideas to promote technology-based companies.

# Jitter and Shimmer Measurements for Speaker Diarization

*Abraham Woubie*[1], *Jordi Luque*[2], *and Javier Hernando*[1]

[1] Universitat Politecnica de Catalunya, BarcelonaTech, Barcelona, Spain
[2]Telefonica Research, Edificio Telefonica-Diagonal, Barcelona, Spain
abraham.woubie.zewoudie@upc.edu,jls@tid.es,javier.hernando@upc.edu

**Abstract.** Jitter and shimmer voice quality features have been successfully used to characterize speaker voice traits and detect voice pathologies. Jitter and shimmer measure variations in the fundamental frequency and amplitude of speaker's voice, respectively. Due to their nature, they can be used to assess differences between speakers. In this paper, we investigate the usefulness of these voice quality features in the task of speaker diarization. The combination of voice quality features with the conventional spectral features, Mel-Frequency Cepstral Coefficients (MFCC), is addressed in the framework of Augmented Multiparty Interaction (AMI) corpus, a multi-party and spontaneous speech set of recordings. Both sets of features are independently modeled using mixture of Gaussians and fused together at the score likelihood level. The experiments carried out on the AMI corpus show that incorporating jitter and shimmer measurements to the baseline spectral features decreases the diarization error rate in most of the recordings.

**Keywords:** speaker diarization, spectral features, jitter, shimmer, fusion

## 1 Introduction

Speaker diarization is the process of segmenting and clustering a speech recording into homogeneous regions and answers the question "Who spoke when" without any prior knowledge about the speakers [1]. A typical diarization system performs three basic tasks: first, it discriminates speech segments from the non-speech ones; second, it detects speaker change points to segment the audio data and finally, it groups these segmented regions into speaker homogeneous clusters. Speaker diarization can be used in different applications such as speaker tracking and speech recognition [1].

The performance of a speaker diarization system largely depends on successful extraction of relevant speaker independent features. Although short-term spectral features are the most widely used ones for different speech applications, the authors in [2] show that long term features can be employed to reveal individual differences which can not be captured by short-term spectral features. The current state-of-the-art speaker diarization systems usually make use of short-term spectral features as representation of speaker traits[3]. However, the work of [4] and [5] show that the performance of the state-of-the-art speaker diarization systems can be improved by combining spectral features with prosodic and other long-term features.

Jitter and shimmer measure fundamental frequency and amplitude variations, respectively. Previous studies have shown that these voice quality features have been successfully used in speaker recognition and emotion classification tasks. The work of [6] shows that adding jitter and shimmer voice quality features to both spectral and prosodic features improves the performance of a speaker verification system. The work of [7] also reports that fusion of voice quality features together with the spectral ones improves the classification accuracy of different speaking styles and conveys information that discriminates the different animal arousal levels. Furthermore, these voice quality features are more robust to acoustic degradation and noise channel effects [8].

Based on these studies, we propose the use of jitter and shimmer voice quality features for speaker diarization task as they can add complementary information to the baseline spectral features. The main contribution of this work is the extraction of jitter and shimmer voice quality features and their fusion with the spectral ones in the framework of speaker diarization task. The experiments are tested on AMI corpus [9], a multi-party and spontaneous speech set of recordings, and assessed in terms of speaker diarization error (DER).

This paper is organized as follows. An overview of voice quality features used in this work is presented in Section 2. Section 3 provides an overview of agglomerative hierarchical clustering of speakers followed by fusion of spectral and voice quality features in Section 4. Experimental results are presented in Section 5 and finally, conclusions of the experiments are given in section 6.

## 2 Voice-quality features

Although the dominant features for speaker diarization are MFCC, studies such as [4] and [5] show that long term features such as prosody can also be used in speaker diarization systems. Long term features are able to acquire phonetic, prosodic and lexical information which cannot be captured by spectral ones.

Jitter measures variations of the pitch in voice whereas shimmer describes variation of the loudness. Studies show that these voice quality features can be used to detect voice pathologies [10]. They are normally used to measure long sustained vowels where measured values above a certain threshold are considered as pathological voices. Studies show that voice quality features have been successfully used in speaker recognition and other speech technology researches. For example, the work of [10] reports that jitter and shimmer measurements provide significant differences between different speaking styles.

Although different estimations of jitter and shimmer measurements can be found in the literature, we focus only on the following three measurements called absolute jitter, absolute shimmer and shimmer apq3 encouraged by previous work of [6]. The work of [6] has shown that these three measurements provided better results in speaker recognition tasks than the other jitter and shimmer estimations. The three voice quality measurements are extracted over 30ms frame length at 10ms rate by means of Praat software [11]. Then, we calculate the mean of each of the three measurements over a window length of 500ms at 10ms step to smooth out fundamental frequency estimation and synchronize with the short-term spectral features.

### 2.1 Jitter measurement

Jitter (absolute) is a cycle-to-cycle perturbation in the fundamental frequency of the voice , i.e. the average absolute difference between consecutive periods, expressed as:

$$\text{Jitter (absolute)} = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}| \tag{1}$$

where $T_i$ are the extracted pitch period lengths and $N$ is the number of extracted pitch periods.

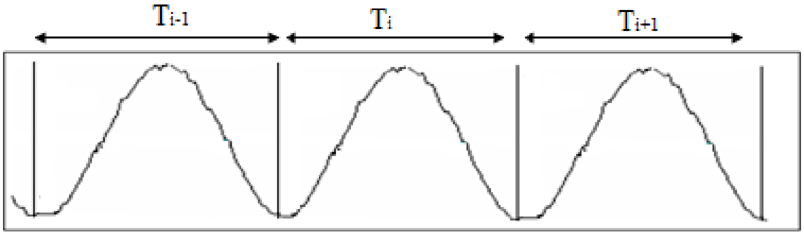


**Fig. 1.** *Jitter measurement for $N = 3$ pitch periods*

### 2.2 Shimmer measurement

- Shimmer (absolute) is the average absolute logarithm of the ratio between amplitudes of consecutive periods expressed as:

$$\text{Shimmer (absolute)} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left|20 \log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{2}$$

  where $A_i$ are the extracted peak-to-peak amplitude data and $N$ is the number of extracted pitch periods.
- Shimmer (apq3) is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude expressed as:

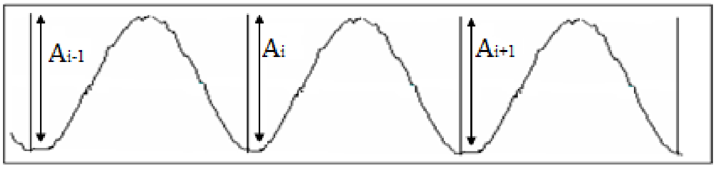


**Fig. 2.** *Absolute shimmer measurement for $N = 3$ pitch periods*

$$\text{Shimmer (apq3)} = \frac{1}{N-2}\sum_{i=2}^{N-1}|A_i - (\frac{A_{i-1}+A_i+A_{i+1}}{3})| \tag{3}$$

where $A_i$ are the extracted peak-to-peak amplitude data and $N$ is the number of extracted pitch periods.

## 3 Agglomerative hierarchical clustering of speakers

For this work, speaker diarization is performed on a mono-channel audio recording. Our approach is based on the bottom-up version of agglomerative hierarchical clustering(AHC). AHC-based diarization has been shown as one of the most successful approaches to address the problem of speaker clustering [12, 13]. Algorithm 1 highlights the main steps of the AHC popular technique. Input features $\{\mathbf{x}_i\}$ are partitioned in a set of segments $\hat{C}_i$, dividing the whole feature set. The clusters in the first iteration are initialized through previous segments and a model is built on them. Next, distances $d(\hat{C}_k, \hat{C}_l)$ among cluster models are computed in a pairwise comparison which aims to group similar regions. The initial clustering is iterated and the clusters are merged and aligned until some condition is fulfilled, e.g., a threshold on the previous distance matrix. Finally, each remaining cluster is expected to represent an ensemble of the data based on the selected distance measure.

**Algorithm 1** Agglomerative Hierarchical Clustering (AHC), bottom-up alternative.

```
Require: $\{\mathbf{x}_i\},\ i = 1\ldots,\hat{n}$ : speech segments
         $\hat{C}_i,\ i = 1,\ldots,\hat{n}$ : initial clusters
Ensure:  $C_i,\ i = 1,\ldots,n$ : finally remaining clusters
$\hat{C}_i \leftarrow \{\mathbf{x}_i\},\ i = 1,\ldots,\hat{n}$
repeat
    $i, j \leftarrow$ argmin $d(\hat{C}_k, \hat{C}_l),\ k, l = 1,\ldots,\hat{n}, k \neq l$
    merge $\hat{C}_i$ and $\hat{C}_j$
    $\hat{n} \leftarrow \hat{n} - 1$
until no more cluster merging is needed
return $C_i,\ i = 1,\ldots,n$
```

In Figure 3, it is depicted a more detailed scheme of the AHC-based speaker clustering. The previous high level steps are adapted to the speaker diarization task jointly with the key idea that each cluster $C_i$ should be composed exclusively by speech from the same speaker.

*Speech activity detection*(SAD): We have used Oracle SAD (the reference speech/non speech annotations) as our speech activity detection.

*Cluster initialization*: An initial segmentation is performed based on the homogeneous partition along time of the speech-only features, see (Fig. 3 block A). The number of initial clusters is selected automatically depending on the meeting duration but constrained to the range [35,65] clusters. It aims to deal with the trade-off between having a significant number of samples for modeling and avoiding common issues of AHC, such as overclustering and its high computational cost. So the number $K_{\text{init}}$ of initial clusters is defined as

$$K_{\text{init}} = \frac{N}{G_{\text{init}}\, R_{\text{CC}}}, \tag{4}$$

where $N$ stands for the number of features available per cluster and $G_{\text{init}}$ is the number of Gaussians initially assigned to each cluster. The complexity ratio $R_{\text{CC}}$, the minimum number of frames per Gaussian, is fixed to 7 and the $G_{\text{init}}$ to 5 Gaussians. Despite of its simplicity, this regular partition of the data allows the creation of "pure" enough initial cluster which is a key point in the algorithm [14, 15].

*Acoustic modeling*: Each set of acoustic features related to a cluster is independently modeled using HMM/GMM which is iteratively refined, (Fig. 3 block B). It is done in each clustering iteration through a two step training and decoding process. Each state of the HMM is composed by a mixture of Gaussians, fitting the probability distribution of the features by the classical expectation-maximization (EM) algorithm. Note that two independent HMM models are estimated per each feature stream but their log likelihoods given a feature are weighted as explained in Section 4. The number of mixtures is chosen as a function of the available seconds of speech per cluster in the MFCC features and fixed for the shimmer and jitter features. A time constraint, as in [16], is also imposed on the HMM topology which enforces the minimum duration of the speaker turn to be greater than 3 seconds.

*Agglomerative distance* is based on the Bayesian Information Criterion (BIC) as a metric among clusters. Furthermore, the stopping criterion, or ending point of the algorithm, is also driven by a threshold on the same matrix of distances, (Fig. 3 block C). A modified BIC-based metric [16] is employed to select the set of cluster-pairs candidates with smallest distances among them. Cluster-pair $(i, j)$ is merged depending on whether its $BIC_{ij}$ fulfills

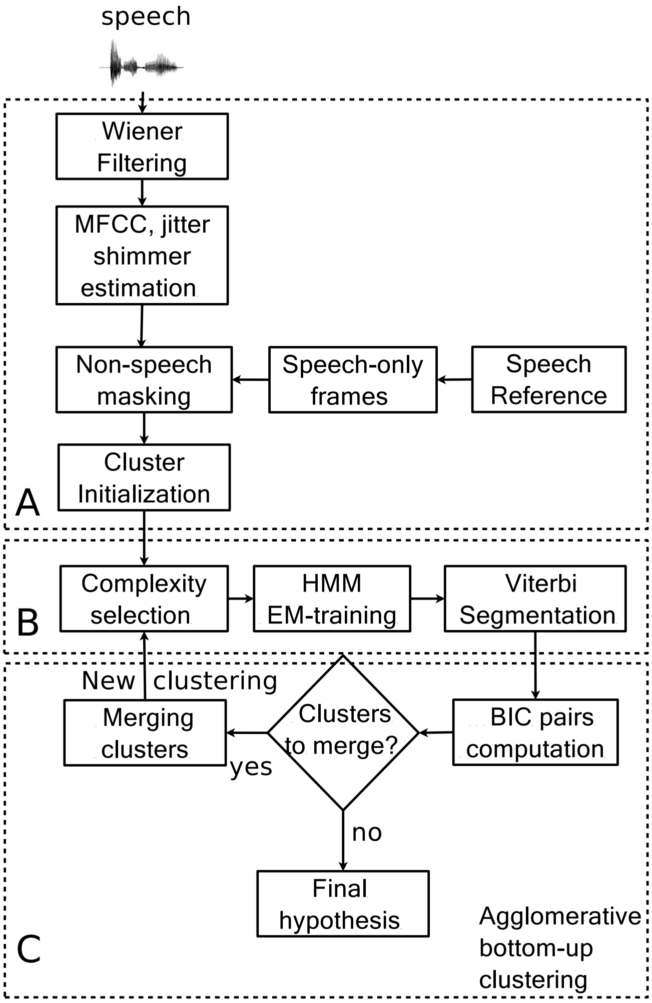


**Fig. 3.** *Speaker diarization scheme based on Agglomerative Hierarchical Clustering with automatic complexity selection.*

$$BIC_{ij} > \max\big(\gamma, BIC_{\mu} + \frac{3}{2}BIC_{\sigma}\big), \tag{5}$$

where $BIC_{ij}$ is the $BIC$ estimation between the clusters $i$ and $j$ performed as in [16] and $\gamma$ is a threshold tuned on development data. The $BIC_{\mu}$ is the mean of $BIC_{ij}$ for $i \neq j$ and the $BIC_{\sigma}$ stands for the standard deviation of the same BIC set. Once clusters are merged, a two-step training and decoding iteration is performed again to refine the model statistics and align them with the speech recording, block B (see Fig. 3). The model complexity $M_i^j$, the number of mixtures composing the model associated to cluster $i$ at iteration $j$, is updated according to the $R_{\text{CC}}$ value but only for the MFCC stream. In the case of voice quality features, Gaussian complexity is fixed manually and different values are explored. The automatic selection of the model complexity for MFCC features has shown a successful performance while it avoids the use of the penalty term in the classical BIC formulation [17, 12]. It is done by the following equation

$$M_i^j = \left\lfloor (\frac{N_i^j}{R_{\text{CC}}}) + \frac{1}{2} \right\rfloor, \tag{6}$$

where $N_i^j$ is the number of frames belonging to the cluster $i$. A more detailed description of the system can be found in [13, 18].

# 4 Fusion of spectral and voice quality features

Since the spectral and voice quality features have different dimensions and use different number of Gaussians per model, two independent HMM models have been estimated per each feature stream. The spectral features are used in parallel with voice quality features both in segmentation and clustering. The segmentation process uses the joint log likelihood ratio of both feature sets of the best path to create a segmentation hypothesis and the agglomerative clustering uses $\Delta$BIC of fused Gaussian mixture mode scores to decide cluster merging. Given a set of input features vectors $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$, MFCC and jitter/shimmer respectively, which belong to same cluster, the log-likelihood is computed as a joint likelihood of both feature distributions as follows:

$$\log P(\mathbf{x}, \mathbf{y}|\theta_{ix}, \theta_{iy}) = \alpha \log P(\mathbf{x}|\theta_{ix}) + (1 - \alpha) \log P(\mathbf{y}|\theta_{iy}), \tag{7}$$

where $\theta_{ix}$ is the model of cluster $i$ using spectral feature vectors $\{\mathbf{x}\}$, and $\theta_{iy}$ is the model of the same cluster $i$ using jitter and shimmer feature vectors $\{\mathbf{y}\}$. The weight of the spectral feature vector is $\alpha$ and, consequently, $(1 - \alpha)$ is the weight of jitter and shimmer voice quality features.

# 5 Experiments

## 5.1 Database and experimental setup

The experiments are tested on AMI meeting corpus, a multi-party and spontaneous speech set of recordings, which consists of roughly 100 hours of speech. We have selected the 11 evaluation sets of the corpus to evaluate the diarization error rate of our approach. The average duration per meeting is around 27 minutes.

First of all, any noise of the input audio signal is minimized using Wiener filtering and we then apply speech activity detection algorithm to detect the speech segments and discard the non-speech ones.

**Table 1.** *Average DER results of the AMI corpus for different weighted combinations of MFCC, and Jitter and Shimmer features(JS) using 2 number of Gaussians for the JS.*

| **Feature set** | **Weight of MFCC** | **Weight of JS** | **DER** |
|---|---|---|---|
| MFCC (Baseline) | 1 | 0 | 24.76% |
| MFCC + JS | 0.95 | 0.05 | **21.45**% |
| MFCC + JS | 0.925 | 0.075 | 22.76% |
| MFCC + JS | 0.9 | 0.1 | 22.23% |

The raw speech waveforms are then parameterized into sequences of MFCC using Fast Fourier Transform based log spectra with 30ms frame length and 10ms frame shift. The total number of coefficients extracted for the spectral features are 20. The extracted MFCC do not include deltas. The extraction of the three voice quality features is done as explained in Section 2. Fusion of the two set of features is done at the score likelihood level as explained in Section 4.

### 5.2 Experimental results

The performance of a speaker diarization system is evaluated using diarization error rate (DER) which represents the error contribution of missed speech, false alarm and speaker error. [1] We have used the reference speech/non speech annotations as our speech activity detection. The reason for using the reference speech/non speech annotations is that we are only interested to investigate the usefulness of voice quality features in reducing DER. The use of another speech activity detection may complicate the task and create more confusion. Therefore, the false alarms and missed speech have zero values in our experimental results.

As shown in Table 1, we have applied different weights for both features sets to find out the optimum set of weight values that provide us with the best results in terms of DER. The baseline system, which relies on spectral features, shows a DER of 24.76%. Weighting the MFCC by 0.95 and the voice quality features by 0.05 gives us a DER of 21.45%. It represents a 13.37% relative improvement compared to the baseline. We have observed that incorporatingg jitter and shimmer measurements to the baseline spectral features decreases the diarization error rate in nine of the eleven AMI recordings. Table 1 also shows that using different weight values for the jitter and shimmer features shows DER values better than the baseline.

We have also carried out an experiment to find out the best number of Gaussians for the voice quality features when its weight value is 0.05. The best DER result is found when we use 2 Gaussians as shown in Figure 4 which gives us a DER of 21.45% . The figure also shows that using one, three and five Gaussians provide better DER values than the baseline.The standard deviations of DER values in Figure 4 show the DER variations among the recordings.

[1] The scoring tool is the NIST RT scoring used as: ./md-eval-v21.pl -1 -nafc -o -R reference.rttm -S system_hypothesis.rttm

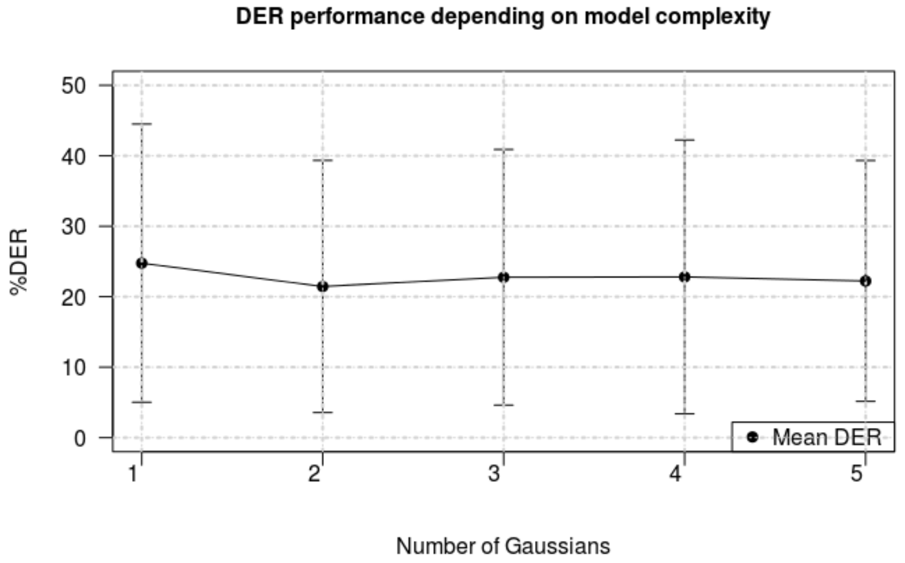


**Fig. 4.** *Diarization Error Rate (DER) and standard deviation as per the number of Gaussians for the JS with 0.95 and 0.05 weight values for MFCC and JS, respectively.*

## 6 Conclusions

We have proposed the use of jitter and shimmer voice quality features for speaker diarization experiment as these features add complementary information to the conventional baseline MFCC features. Jitter and shimmer voice quality features are first extracted from the fundamental frequency contour, and are then fused together with the baseline MFCC features. The fusion of the two streams is done at the score likelihood level by a weighted linear combination of the output log-likelihoods of each model. Our experiments show that fusing jitter and shimmer voice quality features with the baseline spectral features shows a 13.37% relative DER improvement.

## 7 Acknowledgements

This work has been partially funded by the Spanish Government projects TEC2010-21040-C02-01 and PCIN-2013-067.

## References

1. Tranter, S. and Reynolds, D.: An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on* 14, 1557 –1565 (2006)

2. Farrús, M., Garde, A., Ejarque, P., Luque, J., and Hernando, J.: On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *9th International Conference on Spoken Language Processing, ICSLP*, 2106–2109 (2006)
3. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O.: Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing* (2011)
4. Friedland, G., Vinyals, O., Huang, Y., and Müller, C.: Prosodic and other Long-Term Features for Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing* (2009)
5. Zelenák, M. and Hernando, J.: The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization. In *INTERSPEECH*, 1041–1044 (2011)
6. Farrús, M., Hernando, J., and Ejarque, P.: Jitter and Shimmer Measurements for Speaker Recognition. In *INTERSPEECH* (2007)
7. Li, X., Tao, J., Johnson, M., Soltis, J., Savage, A., Leong, K., and Newman, J.: Stress and Emotion Classification using Jitter and Shimmer Features. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, IV–1081–IV–1084 (2007)
8. Carey, M., Parris, E., Lloyd-Thomas, H., and Bennett, S.: Robust prosodic features for speaker identification. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, 1800–1803 vol.3 (1996)
9. The Augumented Multi-party Interaction project,AMI meeting corpus. Website, `http://corpus.amiproject.org` (2011)
10. Kreiman, J. and Gerratt, B. R.: Perception of aperiodicity in pathological voice. *J Acoust Soc Am* 117 (2005)
11. Boersma, P. and Weenink, D.: Praat: doing phonetics by computer, Version 5.3.69. `http://www.praat.org/`
12. Fiscus, J. and et al.: The rich transcription evaluation project. Website, `http://www.nist.gov/speech/tests/rt/`
13. Luque, J. and Hernando, J.: Robust Speaker Identification for Meetings: UPC CLEAR-07 Meeting Room Evaluation System. In *Lecture Notes on Computer Science, LNCS*, vol. 4625. Springer-Verlag (2008)
14. Imseng, D. and Friedland, G.: Tuning-Robust Initialization Methods for Speaker Diarization. *Audio, Speech, and Language Processing, IEEE Transactions on* 18, 2028 –2037 (2010)
15. Luque, J., Segura, C., and Hernando, J.: Clustering initialization based on spatial information for speaker diarization of meetings. In *International Conference on Spoken Language Processing, ICSLP*, 383–386. Brisbane, Australia (2008)
16. Ajmera, J. and Wooters, C.: A robust speaker clustering algorithm . In *Proceedings of IEEE Speech Recognition and Understanding Workshop*. St. Thomas, U.S. Virgin Islands (2003)
17. Anguera, X., Wooters, C., and Hernando, J.: Robust Speaker Diarization for Meetings: ICSI RT06s evaluation system. In *International Conference on Spoken Language Processing, ICSLP* (2006)
18. Luque, J. and Hernando, J.: On the use of Agglomerative and Spectral Clustering in Speaker Diarization of Meetings. In *Odyssey 2012-The Speaker and Language Recognition Workshop* (2012)

# iVectors for Continuous Emotion Recognition

Paula Lopez-Otero, Laura Docio-Fernandez, and Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{plopez,ldocio,carmen,cardenal}@gts.uvigo.es

**Abstract.** This work proposes the use of the iVectors paradigm for performing continuous emotion recognition. To do so, a segmentation of the audio stream with a fixed-length sliding window is performed, in order to obtain a temporal context which is enough for capturing the emotional information of the speech. These segments are projected into the iVectors space, and the continuous emotion labels are learnt by canonical correlation analysis. A voice activity detection strategy is incorporated to the system in order to ignore the non-speech segments, which do not provide any information about the emotional state of the speaker, and to recreate a real-world scenario. Results on the framework of the Audiovisual Emotion Challenge (AVEC) 2013 show the potential of this approach for the emotion recognition task, obtaining promising results as well as using low-dimensional data representation.

**Keywords:** iVectors, continuous emotion recognition, arousal, valence

## 1 Introduction

Emotion recognition is a task consisting on extracting information about the speaker's emotional state. The emotion recognition field is gaining interest for different real-world applications. The emotional aspects of human-computer interaction started to receive some attention in the last years, becoming a huge concern [15]; this interest is caused by the importance of expressivity when dealing with a computer interface, as the interface should be able to detect the emotional state of the user and adapt its behaviour according to it [4]. The use of emotion recognition for monitoring users' reaction to advertisement is nowadays a reality, and it also has paramount importance in the entertainment industry, either for the development of toys or videogames [14] or for the development of serious games for aiding people with problems to understand social signs [1].

Emotion recognition can be carried out in two different ways: one is the discrete emotion recognition task, that consists on detecting a set of given emotions, and the other one is the continuous emotion recognition task, in which the continuous values of the affect variables have to be estimated [9]. The two problems are closely related, as a discrete emotion has a correspondence with concrete values of emotional dimensions; in [16], it is said that the different dimensions of affect lie in different angles of a circle, and the angle inside this circle depends on two different emotional dimensions, namely arousal, which measures the degree of excitation, and valence, which measures the degree of pleasantness.

Emotion recognition systems can be usually split in three stages, namely feature extraction, modelling and prediction of the emotional level. On the first stage, the features used in emotion recognition can be divided in two big groups according to their time span: low level descriptors (LLD) or instantaneous descriptors are extracted for each time frame, such as Mel-frequency cepstral coefficients, loudness, zero crossing rate, jitter or shimmer; and functionals or global descriptors are computed using the LLD extracted for the whole audio signal or for an audio segment covering several audio frames, such as the mean, standard deviation, quartile, flatness or skewness, among others.

The modelling stage of an emotion recognition system must obtain a representation of the speech that reflects the emotional information. Depending on the features used, different modelling approaches can be found in the literature. When using functionals, it is common to model the speech using those features directly or applying feature selection strategies [8]. When dealing with LLD, different techniques can be borrowed from other speech recognition tasks, such as supervised and unsupervised subspace learning techniques. The use of such techniques in discrete emotion recognition is straightforward, but their application to continuous emotion recognition has some issues. First, there is not a discrete number of emotions, so the training stage of a supervised learning strategy cannot be carried out directly; this issue can be partially solved by quantizing the emotional dimensions [17]. Moreover, an instantaneous value of the emotion must be estimated, but it is not possible to apply such learning techniques at a frame rate, as the modelling of speech through these techniques usually requires a context longer than a few milliseconds. A windowing of the speech can be done before the modelling stage in order to obtain a longer context [17].

The prediction of the emotional level is usually carried out using machine learning techniques such as support vector machines (SVM) or random forests in the discrete case, while in the continuous case strategies such as support vector regression (SVR) or canonical correlation analysis [10] can be used.

In this work, we present a continuous emotion recognition system as a continuation of our previous research on speech modelling in continuous emotion recognition. In [17], we successfully applied the eigen-voices approach to emotion recognition, and in this work we propose the use of the iVector modelling technique. The iVector paradigm is considered state-of-art in different tasks such as speaker recognition and verification [5] or language identification [6] due to its potential for representing audio in a speaker and channel independent way in a low dimensional subspace, which are desirable qualities in emotion recognition. The iVector representation was used in [19] for discrete emotion recognition, but in this work we present a system that permits the use of this representation for the continuous task. The validity of this approach is assessed in the framework of the AVEC 2013 affect recognition sub-challenge [13], which consisted on estimating the continuous levels of arousal and valence in a set of recordings featuring different speakers and channel conditions.

The rest of this paper is organized as follows: Section 2 presents the proposed system for performing continuous emotion recognition; Section 3 describes the experimental framework used to assess the validity of the proposed system; the experimental settings are described in Section 4; experimental results are discussed in Section 5; and Section 6 summarizes the conclusions and future work.

## 2 Proposed continuous emotion recognition system

Figure 1 presents an overview of the proposed approach for performing continuous emotion recognition, whose different blocks are described in detail in the rest of this Section. The emotional levels on $E$ different dimensions have to be estimated by this approach, which, as mentioned in the introduction, can be divided in three stages: feature extraction, accompanied by a segmentation step, iVector modelling and estimation of the emotional level. This system is similar to the one presented in [17], as the feature extraction and segmentation procedures, as well as the approach for predicting the emotional level, are almost the same; these two systems differ in the modelling of the speech segments, which in this case is carried out using the iVector paradigm.

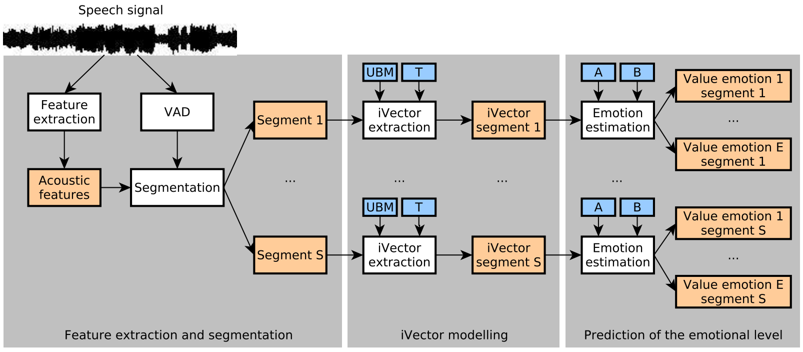


**Fig. 1.** Continuous emotion recognition system

### 2.1 Feature extraction and segmentation

The first step consists on extracting vectors of $F$ features from the audio signal. Acoustic features represent a temporal context which usually ranges from 20 to 100 ms; it is not possible to identify an emotion with such a short time span, a bigger context is necessary. Thus, to obtain vectors that represent a bigger context, the audio is segmented using a fixed-length sliding window, obtaining a set of $S$ segments. In this way, the segments of audio can be represented by means of Gaussian mixture models (GMMs).

It must be noted that the audio signals may contain non-speech parts, which act as a nuisance in the continuous emotion recognition system, i.e. the non-speech parts contribute in a negative way to this procedure, as they do not hold any information about the speaker's emotional state. Hence, it is important to perform voice activity detection (VAD) on the audio signals. To do so, the approach described in [2] was used, which uses a linked-HMM architecture and robust features that are independent of the signal energy, making this strategy robust to noise. An audio segment is considered to contain speech if at least the 50% of its duration was labelled as speech by the VAD strategy.

### 2.2 iVector modelling

The procedure performed at the previous stage of this system allows the use of iVectors for modelling the acoustic information in a low-dimensionality space. Given a Universal Background Model (UBM) with $N$ mixtures, this UBM is adapted to the segments extracted from the training files using Maximum a Posteriori (MAP) adaptation, and the means of the resulting Gaussian Mixture Model (GMM) are concatenated in order to obtain a Gaussian mean supervector for each segment. As we want to avoid the effects of speaker and channel variability, the iVector technique is applied to the Gaussian mean supervectors. This technique defines a low-dimensional space, named total variability space, in which the speech segments are represented by a vector of total factors, namely iVector [5]. A Gaussian mean supervector $\mathbf{M}$ is decomposed as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where $\mathbf{m}$ is the speaker and channel independent supervector, $\mathbf{T}$ is a low-rank total variability matrix, and $\mathbf{w}$ is the iVector corresponding to the Gaussian mean supervector. In this training stage, the matrix $\mathbf{T}$ is trained as described in [11].

Once the total variability matrix $\mathbf{T}$ is obtained, this matrix and the UBM can be used to extract iVectors from the acoustic features corresponding to the different speech segments.

### 2.3 Prediction of the emotional level

The iVector modelling strategy described above results on an iVector for each speech segment, which represents all the emotional dimensions at the same time. Estimated values of the emotional levels of the different dimensions must be extracted from these vectors; to do so, canonical correlation analysis is applied. This technique finds basis vectors for two sets of variables (on the one hand, the iVectors, and on the other hand, the groundtruth emotional levels) such that the correlations between the projection of the variables onto these basis vectors are mutually maximized [10]. After training the canonical correlation projection, two matrices are obtained which are used to project the iVectors, obtaining as a result the estimated emotional levels.

Two different strategies can be followed at this point: a projection can be learnt for each emotional dimension, or a single projection for all the emotional dimensions can be obtained. The first approach does not take into account the correlation between the different emotional dimensions, but the second approach takes advantage of this correlation, which might be due to the emotions themselves or might be caused by the rater that labelled the different emotional dimensions.

It must be noted that, as commented in Section 2.1, the non-speech segments are not used either for training or testing. Thus, there are parts of the audio files whose emotional level is not estimated, as they do not have a corresponding iVector, but it might be necessary to assign them a value for evaluation purposes. To solve this situation, the mean value of the emotional dimension, computed over the training labels, is assigned to the non-speech segments.

## 3 Experimental Framework

The framework of the AVEC 2013 affect recognition sub-challenge (ASC) was used to evaluate the proposed technique for continuous emotion recognition. This task consists on the continuous recognition of the emotional dimensions valence and arousal, and these dimensions have to be predicted for every time instant of the recording. Both arousal and valence values range from -1 to 1.

The data used in these experiments is a subset of the audio-visual depressive language corpus (AVDLC) [13]. The speakers were recorded in diverse quiet locations using a laptop and a headset. The database is organized in three partitions of 50 recordings each, which are summarized in Table 1. Each recording features one speaker (either male or female), and there can be several recordings per speaker, with a time separation of two weeks between different recordings. The subjects' age ranged between 18 and 63 years (mean 31.5).

**Table 1.** Summary of the datasets used in the experiments.

| Set | Total duration | Min duration | Max duration |
|---|---|---|---|
| Training | 13 h 17 min | 8 min 5 s | 27 min 20 s |
| Development | 13 h 5 min | 14 min 20 s | 23 min 55 s |
| Testing | 12 h 59 min | 5 min 15 s | 23 min 57 s |

The recordings were power-point guided, indicating the speaker what to do at each moment. These tasks consisted in reading excerpts of novels and fables, singing, telling stories from the speaker's past, making up a story applying the Thematic Aperception Test (TAT), and sustained vowel phonation. These tasks try to provoke different feelings on the speakers such as happiness, by talking about their best present ever, or sadness, by talking about their saddest memory of their childhood.

The recordings included in the database were manually labelled by a team of 23 raters. Each recording was annotated by a single rater, which used a joystick to instantaneously register the level of arousal or valence (the two dimensions were annotated separately, not at the same time). In order to address the intra-annotator variability, all the raters were asked to annotate a reference video, and these annotations were used to create models that compensated that variability. The annotations were binned in temporal units of time of the same duration, which in this case was 1/30 seconds (i.e. equal to the video frame rate).

Figure 2 shows the distribution of arousal and valence in the training and development datasets (the testing dataset is not included due to the unavailability of its corresponding groundtruth labels). As shown in this Figure, the most probable value of arousal and valence is 0, which means neutral arousal or valence. This is due to the fact that, during the recordings, there are long silence periods which the raters labelled as neutral. It can also be observed in this Figure that values close to -1 and 1 are not very likely to appear.

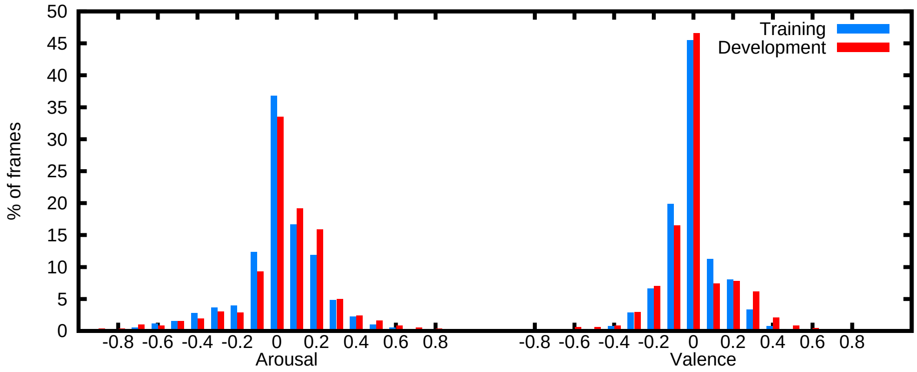


**Fig. 2.** Distribution of arousal and valence values in the training and development datasets

The evaluation metric used to measure the performance of continuous emotion recognition approaches in this framework is the Pearson's correlation coefficient between the groundtruth labels and the estimated labels, averaged over all sessions and dimensions. It must be noted that the absolute value of the correlation coefficient is used, as a negative correlation is a correlation anyway. This fact was not mentioned in the description of the AVEC 2013 ASC challenge [13], but was mentioned afterwards when the evaluation scripts were released.

## 4 Experimental settings

In this continuous emotion recognition system, 16 MFCCs plus delta and derivatives are used, as these features are a common choice for emotional speech analysis [18], leading to feature vectors of dimension 48. The use of MFCCs is also supported by previous research in [17], where the best emotion recognition results were obtained using MFCCs. These features were extracted every 10 ms using a 40 ms Hamming window and mean and variance normalization is applied. The window used to perform the audio segmentation has a length of 3 s with 1 s of overlap. The number of mixtures of the GMM was set to 256 after observing that the influence of this value in the experimental results was negligible. Feature extraction was done using the OpenSMILE feature extraction software [7], and all the tasks that involved iVector training and extraction were performed using the ALIZE 3.0 toolkit [12].

The training partition of the AVEC 2013 ASC database was used to perform the training of the system (UBM, total variability matrix and canonical correlation projection) as well as to adjust the free parameters of the system, i.e. the dimension of the iVectors. The development partition was used to test the performance of the proposed system, and the testing data was discarded due to the unavailability of their corresponding groundtruth labels.

The validity of the iVector approach for emotion recognition was assessed by comparing the proposed system with the same system without applying the iVectors paradigm, i.e. a GMM supervector (GSV) approach [3]. In this GSV approach, the audio segments were represented by mean supervectors $\mathbf{m}$ of dimension $N \cdot F = 256 \cdot 48 = 12288$, according to the notation in Section 2.2.

It must be noted that the rate of the groundtruth labels is higher than the rate of the iVectors (there is one iVector every two seconds while there is a groundtruth label every 1/30 seconds); in order to make them have the same rate, the mean value of the emotional dimension on the time span of the iVector is used as the groundtruth label to train the canonical correlation basis vectors. In the test data, as an emotional level is obtained every 2 seconds, the obtained level is replicated so it fits the rate of the groundtruth labels (i.e. every label is repeated $\frac{2}{1/30} = 60$ times).

## 5 Experimental results

The starting point of the experimental validation of the proposed strategy consisted on performing continuous emotion recognition using different dimensions of the iVectors. To do so, the Pearson's correlation coefficient between the groundtruth training labels and the estimated ones was computed for different iVector dimensions, as presented in Table 2. This Table shows that the highest correlation was obtained when using iVectors of dimension 25. It must be noted that these results were obtained when using the manual VAD of the audio signals, in order to avoid the nuisance generated by errors on the VAD stage.

A straightforward experiment to prove the validity of the iVector technique consists on comparing the results when applying this approach and when not

**Table 2.** Pearson's correlation coefficient obtained on the training dataset with different iVector dimensions using manual VAD.

| Dimension | Arousal | Valence | Average |
|---|---|---|---|
| 25 | **0.1963** | 0.1862 | **0.1912** |
| 50 | 0.1859 | **0.1905** | 0.1882 |
| 100 | 0.1847 | 0.1892 | 0.1870 |
| 150 | 0.1846 | 0.1868 | 0.1857 |
| 200 | 0.1863 | 0.1852 | 0.1857 |

applying it. Thus, results obtained with the GSV approach were compared with those obtained when using the iVector approach. Table 3 supports the validity of the iVector modelling, as the Pearson's correlation coefficient of the estimation of arousal and valence obtained with the iVectors representation is higher than that obtained when modelling the speech segments with the GSV approach. Another issue that must be noticed is the dramatic reduction of the dimensionality of the vectors: while the dimension of the iVectors was 25, the GSV approach used vectors of dimension 12288.

**Table 3.** Pearson's correlation coefficient obtained on the development dataset with and without using the iVector modelling.

| VAD | Approach | Dimension | Arousal | Valence | Average |
|---|---|---|---|---|---|
| Manual | iVectors | 25 | **0.2041** | **0.1697** | **0.1869** |
|  | GSV | 12288 | 0.1717 | 0.1299 | 0.1508 |
| Automatic | iVectors | 25 | **0.1846** | **0.1575** | **0.1711** |
|  | GSV | 12288 | 0.1621 | 0.1264 | 0.1443 |
| - | Eigen-emotions | 50 | 0.1721 | 0.1404 | 0.1562 |

The last step of this experimental validation of the iVector modelling for emotion recognition was to apply an automatic VAD strategy to the development segments, in order to observe the impact of making errors in the detection of speech. The VAD strategy described in Section 2.1 was used for that purpose; comparing the manual VAD to the automatic VAD results obtained with such strategy, a missed speaker time of 18.6% and a false alarm speaker time of 1.8%, with respect to the scored speaker time, were achieved. Table 3 shows that the system presents some sensitivity to the VAD errors, as the Pearson's correlation coefficient is reduced by 0.015 when using iVectors and by 0.007 when using the GSV approach. These errors are due to the missed speaker time, as there are speech segments whose emotional dimensions are not being estimated because they were labelled as non-speech.

Table 3 also shows the Pearson's correlation coefficient achieved with another subspace projection-based approach available in the literature, which used the same experimental framework. This approach, namely eigen-emotions [17], shows similar results to those obtained with the GSV technique, but it is outperformed by the iVectors representation.

## 6 Conclusions and future work

This work proposed the use of iVectors on the continuous emotion recognition task, due to the ability of this paradigm to get rid of the speaker and channel variabilities. The experimental results obtained in the framework of the AVEC 2013 affect recognition sub-challenge showed an improvement on the emotion recognition results when using the iVector paradigm, as well as a huge dimensionality reduction of the feature vectors used to represent the speech segments. Hence, the experimental results suggest that the success achieved by the iVector representation in different speech technologies tasks is extensible to the emotion recognition field as well. Nevertheless, due to the temporary unavailability of the groundtruth labels of the test data used in the AVEC 2013 evaluation, it was not possible to compare these results to those obtained on the test data using other systems, but a more extensive analysis of this approach will be performed whenever these labels are available.

An automatic strategy to discriminate speech and non-speech was applied in order to assess the performance of the proposed emotion recognition approach in a realistic scenario; this procedure resulted in a slight reduction of the Pearson's correlation coefficient due to the errors on the voice activity detection module, specially to the missed speech errors. Better strategies for voice activity detection must be developed and incorporated to this emotion recognition system in order to overcome this reduction of performance.

## Acknowledgements

This work has been supported by the European Regional Development Fund, the Galician Regional Government (CN2011/019, 'Consolidation of Research Units: AtlantTIC Project' CN2012/160), the Spanish Government (FPI grant BES-2010-033358 and 'SpeechTech4All Project' TEC2012-38939-C03-01) and 'TecAnDAli Research Network' of the Consellería de Educación e Ordenación Universitaria, Xunta de Galicia.

## References

1. Barakova, E.I., Lourens, T.: Expressing and interpreting emotional movements in social games with robots. Personal Ubiquitous Computing 14(5), 457–467 (2010)
2. Basu, S.: A linked-HMM model for robust voicing and speech detection (2003)
3. Chen, Y., Xie, J.: Emotional speech recognition based on SVM with GMM supervector. Journal of Electronics (China) 29(3), 339–344 (2012)
4. Cowie, R., Douglas-Cowie, E., Karpouzis, K., Caridakis, G., Wallace, M., Kollias, S.: Recognition of emotional states in natural human-computer interaction. Multimodal User Interfaces pp. 119–153 (2008)

5. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing (2010)
6. Dehak, N., Torres-Carrasquillo, P., Reynolds, D., Dehak, R.: Language recognition via i-vectors and dimensionality reduction. In: Proceedings of Interspeech. pp. 857–860 (2011)
7. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE - the Munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM Multimedia (MM). pp. 1459–1462 (2010)
8. Gosztolya, G., Busa-Fekete, R., Tth, L.: Detecting autism, emotions and social signals using AdaBoost. In: INTERSPEECH. pp. 220–224 (2013)
9. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions 1(1), 68–99 (2010)
10. Hardoon, D.R., Szedmak, S., Szedmak, O., Shawe-taylor, J.: Canonical correlation analysis; an overview with application to learning methods. Tech. rep. (2007)
11. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing 13(3), 345–354 (2005)
12. Larcher, A., Bonastre, J., Fauve, B., Lee, K., Levy, C., Li, H., Mason, J., Parfait, J.: ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition. In: Proceedings of Interspeech. pp. 2768–2772 (2013)
13. M.Valstar, Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013 - the continuous audio/visual emotion and depression recognition challange. In: Proceedings of the 3rd International Audio/Visual Emotion Challenge and Workshop (AVEC'13) (2013)
14. Oudeyer, P.: The production and recognition of emotions in speech: features and algorithms. International Jorunal of Human-Computer Studies 59, 157–183 (2003)
15. Picard, R.W.: Toward computers that recognize and respond to user emotion. IBM Syst. J. 39(3-4), 705–719 (Jul 2000)
16. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
17. Sánchez-Lozano, E., Lopez-Otero, P., Docio-Fernandez, L., Argones-Rúa, E., Alba-Castro, J.L.: Audiovisual three-level fusion for continuous estimation of Russell's emotion circumplex. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. pp. 31–40. AVEC '13, ACM (2013)
18. Sato, N., Obuchi, Y.: Emotion recognition using Mel-frequency cepstral coefficients. Information and Media Technologies 2(3), 835–848 (2007)
19. Xia, R., Liu, Y.: Using i-vector space model for emotion recognition. In: INTERSPEECH. ISCA (2012)

# Ensemble modeling adaptation for different scenarios in speaker verification

Leibny Paola García-Perera[1], Juan Arturo Nolazco Flores[1]⋆, and Eduardo Lleida Solano[2]

[1] Tecnológico de Monterrey, Campus Monterrey
Computer Science Department
Monterrey Nuevo León, México
{paola.garcia,jnolazco}@itesm.mx
[2] Aragon Institute for Engineering Research (I3A) University of Zaragoza
Spain
lleida@unizar.es

**Abstract.** Adapting a trained model to a new scenario is a challenging issue in speaker verification. In this research, we show an *ensemble modeling* procedure able to adapt previous trained models to different conditions such as noise type or new database. The starting point of this algorithm is a set of primary models from a general dataset constructed by partitioning the data space according to a specific attribute. The regions depend on a chosen attribute. For example, for a noisy condition the partitions associate each SNR with a specific cluster; for a speaker space, the clusters are computed linking the data with similar acoustic characteristics (similarity measure among speakers). The adaptation to a new scenario is performed as a discriminative optimization, where the model parameters update using the information of the positive and negative samples of the new data. The results obtained showed that the ensemble adaptation is a suitable solution for both: noisy and clean condition. Our methodology outperforms traditional methods in $\sim 30\%$ for clean speech and $\sim 40\%$ for noisy speech.

**Keywords:** speaker verification, noise condition, ensemble approach, discriminative training

## 1 Introduction

Speaker Verification (SV) is a detection task which outcome is an acceptance or rejection decision. The core of the current systems relies, in part or completely, on the correct model parameter estimation of two classes: targets and imposters. These ideas have evolved from GMM-UBM (Gaussian Mixture Model - Universal Background Model) approach [1], to the state-of-the-art factor analysis approaches [2]. However, the training of the models is still computationally expensive and depends on the amount of data available.

⋆ Thanks to Cátedra de Seguridad de la Información, Tecnologico de Monterrey, Campus Monterrey.

A suitable solution, in this sense, is to find adaptation schemes that may update the already trained models to new scenarios with minimum effort. These new scenarios are, for example, a new database with and without noise. For a clean speech situation, the actual state-of-the-art methods have shown to be competitive [2]. By using the *factor analysis* approaches, the loading matrices handle the variabilities among the target speakers, even when the data belong to different databases. The noisy conditions are more challenging. In this respect, there are just few studies that deal with noise from an entire SV point of view. PRISM [3], was one of the first attempts to analyze the effect of the noise in the current database. Several solutions with different flavors were explored afterwards. Examples of methods that treat the problem using factor analysis are described in [4, 5]. Lately, techniques based on VTS (Vector Taylor Series) has shown to be successful [6, 7].

Our research focuses on an alternative method to deal with new data. The study is motivated by the fact that there are imposters that are more "likely" to be confused with target speakers. Searching for that imposter-specific data set might be exhaustive and time consuming. An ingenious manner to find those imposters that can help to improve the target model is to partition the data space. Computing specific imposter models for these new clusters and discriminatively optimizing taking advantage of similarity measures reduces the number of comparisons to be performed. In addition, the procedure also enhances the target models. The idea was also extended to the condition scenario, *i.e.,* the target speaker is now a target condition (which may enclose: noise type or Signal to Noise Ratio) with its corresponding competing conditions. An interesting partitioning-space approach was presented in [8]. This study clusters the space based on *a)* the vocal tract lengths or *b)* by using the maximum likelihood linear regression (MLLR) to build a model for each partition. They showed that by including a multiple background models the results improved in 13.65% and 15.43% for the Vocal Tract Length Normalization (VTLN) and the MLLR respectively.

In next sections, we detail the algorithm. Section 2 describes the methodology followed highlighting the training stage and the scoring procedure. Section 3 shows the experimental setup and gives the results obtained. Finally, section 4 gives the conclusion of this work and opens discussion for future research.

## 2 Methodology

In this section, we briefly explain the steps of the *ensemble modeling* approach to perform the adaptation. We follow the research in [9, 10]. As shown in Figure 1 we first build an ensemble of the data space; i.e., the space is divided into regions from which models are computed with the traditional techniques. The idea of this first partition provides a fair coverage of the data space. Afterwards, the method adapts the models to our new data, establishing a membership between every new sample and a certain region (note that the new data includes positive and negative samples). Next, we perform a discriminative optimization of the model parameters using the new data. For the test stage, the system computes the scores taking into account each partition in the same way the UBM would be used. After gathering all the scores and performing the fusion a final decision is obtained.

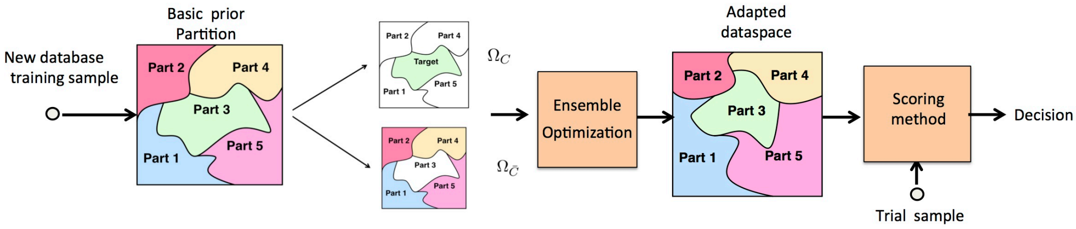


**Fig. 1.** System architecture

### 2.1 First stage training: partitioning background space

The first step of our approach is to build a primary set of partitions (basic prior partition with the same data used to compose the UBM), called *ensemble*, using a known database. Let us define the partitions $\Omega_1, \cdots, \Omega_P$ from the signal space. We explored two partition branches: supervised and unsupervised.

**Supervised partitioning:** The system has *prior* knowledge of the factors by which we wish to partition the space data. Below we consider two mechanisms: partition by *noise* and partition by *speaker*. Partitions may similarly be obtained by other factors such as channel variations.

*Environment-based partitions:* We partition the signal space according to the SNR. We divide the range of all possible SNR values into $P$ intervals. Each interval represents a partition of the signal space. Let $\mathrm{SNR}^C_{min}$ and $\mathrm{SNR}^C_{max}$ represent the minimum and maximum SNRs associated with partition $\Omega_C$. A signal $X$ with *signal-to-noise-ratio*, $\mathrm{SNR}_X$, is assigned to a partition $C$ such that $\mathrm{SNR}^C_{min} < \mathrm{SNR}_X \leq \mathrm{SNR}^C_{max}$. Note that partitions may also be formed based on noise type, or other known characteristics of the noise. In this research, however, we have only considered SNR.

*Speaker partitions:* We first compute a universal background model (UBM) from unpartitioned data. We then use an agglomerative clustering procedure to cluster speakers. Initially, each speaker forms his or her own cluster. From all the current clusters, we select two clusters with the smallest distance and merge them (the smallest distance may be computed using criteria as log likelihood, cross-entropy, euclidean distance between the means, among others.) The procedure continues until there is no cluster left. Agglomerative clustering iteratively merges the closest clusters until the desired number of clusters (and consequently, partitions) is obtained.

**Unsupervised Partitions:** When *a priori* knowledge about the training recordings is unavailable, partitions may be formed by clustering them using unsupervised methods. In this research we employed k-means as a clustering tool. The designer first defines an appropriate number of clusters, $K$. We relate the clusters according to a specific factor. For example, if we focus on noisy signals at different SNR (from 0 to 20) a suitable value of K is five, so that it accounts for five intervals. Once again, the factor by which partitions are formed can be controlled by using an appropriate distance function. Generic clustering based on Euclidean distances or likelihoods may be used to cluster the data by a dominant factor.

For each cluster $C$, the UBM is adapted via MAP adaptation to learn a model $\Lambda_C$. Then, two are the input to the adaptation: the models $\Lambda_C$ and the speakers models $\Lambda_S$. The latter are obtained using MAP from the UBM. These models will be adapted in a posterior stage to the new database or scenario.

### 2.2 Second stage training: Adaptation using the ensemble model

To perform the optimization we define the new positive and negative samples from the targets (enrollment database) and imposters (speakers extracted that are not in the enrollment database) . This procedure provides a natural way to adapt the current models ($\Lambda_C$ and $\Lambda_S$) to the new data.

We require each background model to be highly *specific* to the partition they represent and the new data. Let $\mathbf{X}_S$ represent the samples from speaker $S \in \mathcal{S}$ and $\mathbf{X}_{\bar{S}}$ the recordings from all imposters for $S$, *i.e.* $\mathbf{X} = \mathbf{X}_S \cup \mathbf{X}_{\bar{S}}$. The system establishes a rough membership (initialization) between every new sample to the most likely cluster. $\mathbf{X}_S$ belongs to a specific region $\Omega_{p,S}$ and $\mathbf{X}_{\bar{S}}$ belongs to $\Omega_{p,\bar{S}}$.

Each region is trained following the *discriminative* training procedure. Let $\Lambda_C$ represent the model for a partition $\Omega_C$. Let $\chi_C$ represent all (training) recordings assigned to $\Omega_C$ (including $\mathbf{X}$ of the new dataset). For any partition $\Omega_C$, let

$$\Omega_{\bar{C}} = \bigcup_{\bar{C}' \neq C} \Omega_{\bar{C}'} \tag{1}$$

represent the *complement* of $\Omega_C$, *i.e.* the union of all partitions that are not $\Omega_C$.

To perform the discriminative optimization, based on the minimization of the empirical error, the system adopts a three element formulation including: a set of discriminative functions, a missverification measure and a loss function:

1. Let $g(\chi; \Lambda_C) = \log P(\chi; \Lambda_C)$ be the discriminant function represent the log-likelihood of any recording $\chi$ computed with the distribution for partition $\Omega_C$.
2. Define $d(\chi, \Lambda_C)$, a misclassification measure for how likely it is that a data $\chi \in \chi_C$ from $\Omega_C$ will be misclassified as belonging to $\Omega_{\bar{C}}$ as

$$d(\chi, \Lambda_C) = -g(\chi; \Lambda_C) + G(\chi; \Lambda_{\bar{C}}), \tag{2}$$

$G(\chi; \Lambda_{\bar{C}})$ represents the combined score obtained from a partitions in $\Omega_{\bar{C}}$,

$$G(\chi, \Lambda_{\bar{C}}) = \log \left\{ \frac{1}{|\Omega_{\bar{C}}|} \sum_{C' : \Omega_{C'} \in \Omega_{\bar{C}}} \exp\left[\eta g(\chi, \Lambda_{C'})\right] \right\}^{\frac{1}{\eta}}, \tag{3}$$

where $|\Omega_{\bar{C}}|$ is the number of partitions included in $\Omega_{\bar{C}}$, and $\eta$ is a positive parameter.

3. Define a new objective function for discriminative training of $\Lambda_C$. This function takes the following form,

$$\ell(\Lambda_C) = \frac{1}{|\chi_C|} \sum_{X \in \chi_C} \frac{1}{1 + \exp\left[-\gamma\left(d(\chi, \Lambda_C) + \theta\right)\right]}, \tag{4}$$

where $|\chi_C|$ represents the number of recordings in $\chi_C$, and $\gamma$ and $\theta$ are control parameters.

Finally, the objective function in Equation 4 can be optimized by applying the following generalized probabilistic descent (GPD) update rule for $\Lambda_C$:

$$\Lambda_C^{t+1} = \Lambda_C^t - \epsilon \nabla \ell(\Lambda_C)|_{\Lambda_C^t}. \tag{5}$$

All of these models are GMMs, *i.e.*, $\Lambda_C = \{w_k^C, \mu_k^C, \Sigma_k^C\}$, where $w_k^C$, $\mu_k^C$ and $\Sigma_k^C$ are the mixture weight, mean and covariance matrices of the $k$-th Gaussian of the GMM for $\Lambda_C$. To optimize for these individual parameters, the system computes $\frac{\partial \ell(\Lambda_C)}{\partial w_k^C}$, $\frac{\partial \ell(\Lambda_C)}{\partial \mu_k^C}$ and $\frac{\partial \ell(\Lambda_C)}{\partial \Sigma_k^C}$. The solution is then plugged in for $\nabla \ell(\Lambda_C)$ in Equation (5). Details of the final solution of these gradient are shown in the Appendix.

Once the Ensemble method is performed and the clusters refined according to the new scenario, the system can be used in known methods such as Minimum Verification Error (MVE) [11] and FA (Factor Analysis) approaches. We obtained refined target and imposter models.

### 2.3 Minimum Verification Error (MVE) approach

The background models $\Lambda_C$ are obtained for all partitions. Next, the system trains *partition-specific* target-speaker models, $\Lambda_S^C$ by adapting the background model $\Lambda_C$ and using a similar discriminative approach to train $\Lambda_S^C$. Note that we consider the usual MVE approach [11] (the binary case), with just two classes. In the test stage, the system obtains a score for each trial sample with respect to every partition.

### 2.4 Factor Analysis approach

In the same way, the approach can be embedded in the JFA algorithm. Every partition model is now a set treated separately. The system trains the loading matrices $V$, $U$ and $D$ for each cluster and it computes the speaker factors for each $\Omega_C$. In the test stage, the scores for a trial are obtained against each partition.

### 2.5 Scoring method

In the test stage, the system performs a scoring method as in [9].

Given the pairs,

$$\{\Lambda_{C_1}, \Lambda_S^{C_1}\}, \{\Lambda_{C_2}, \Lambda_S^{C_2}\}, \cdots \{\Lambda_{C_P}, \Lambda_S^{C_P}\},$$

the set of background models for all $P$ partitions and their corresponding partition-specific target speaker models for any claimed speaker $S$, we can compute the score $\theta^S(X)$ to be employed in the likelihood ratio test for any recording $X$ in one of several ways. Let

$$\theta_C^S(X) = \log P(X|\Lambda_S^C) - \log P(X|\Lambda_C) \tag{6}$$

be the likelihood ratio computed in the log domain from the models for partition $\Omega_C$.

The final score, $\theta^S(X)$, for the target speaker with respect to the $P$ models is called *fusion*. For a trial $j$, target $a$, and $P$ different scores, the linear fusion is given by,

$$f_j = \alpha_0 + \alpha_1 a_{1,j} + \alpha_2 a_{2,j} + ... + \alpha_P a_{P,j}. \tag{7}$$

To obtain the $\alpha$ weights, we employ the logistic regression fusion as stated in [12, 13]. Target speaker trial scores and imposter trial scores form two matrices: $A$ is a score matrix of $P \times K$, where $P$ are the different classifiers (partitions) and $K$ are the number of target trials, and $B$ is a score matrix of $P \times L$, where $L$ are the number of imposter trials. Then, the objective logistic regression function based on the *cost* is

$$C = \frac{P_T}{K} \sum_{j=1}^{K} \log(1 + e^{-f_j - \text{logit} P_T}) + \frac{1 - P_T}{L} \sum_{j=1}^{L} \log(1 + e^{g_j + \text{logit} P_T}), \tag{8}$$

where $P_T$ is a prior target probability usually set to 0.5. The target and imposter scores are given by,

$$f_j = \alpha_0 + \sum_{i=1}^{N} \alpha_i a_{i,j} \quad g_j = \alpha_0 + \sum_{i=1}^{N} \alpha_i b_{i,j}. \tag{9}$$

## 3 Experimental Setup

This section describes the databases used and the proposed architecture.

**Baseline system:** The target speakers were initially computed using MAP (all data available for the 200 users was used to compute the models). The JFA experiments employed 50 eigenchannels for the latent variable $U$ and 100 for the eigenvoices $V$. All the approaches that require an *ensemble* discriminative optimization used the usual selection of cohort samples: 500 phrases from each NIST 2004, 2005 2006, switchboard 1, switchboard 2 and 500 phrases from NIST

2010 microphone (3000 phrases were used as imposter samples). For *ensemble modeling*, we limited to five partitions of the UBM data space obtained from NIST databases NIST 2004, 2005 2006, 2008 and 2010(microphone recordings)[3]. The partitioning was performed blindly and supervised. The score outputs for the different scoring strategies were fused.

### 3.1 Adaptation to Ahumada/Gaudi databases:

Ahumada and Gaudi are databases in Spanish [14]. Both databases were recorded in Castilian Spanish under controlled conditions and include 200 male users for Ahumada and 200 female users for Gaudi. The *ensemble modeling* adaptation using these databases represent our first attempt to extend our knowledge beyond languages. We followed the description in [15, 14], but we selected 100 users from each gender, from which imposters and target trials belong to matched and/or mismatched conditions. All the files were downsampled (if needed to 8 kHz). The feature vectors include 49 components: 16 static Cepstral, 1 log Energy, 16 delta Ceptral coefficients, and 16 double delta Ceptral coefficients. Using the Baseline system as a starting point (target models and partition models), we performed the *ensemble modeling* method to produce enhanced scores.

**Adaptation to noise conditions:** We used babble noise, extracted from the Aurora 2 database at different SNRS: 0,5,10,15 and 20 dB to be added to both databases. These new sets will be employed as new scenarios In a first stage, the *ensemble* training used five partitions to refine the models. The next stage adapts the target and the imposter models for each of the clusters.

### 3.2 Results

Tables 1 and 2 describe the performance for the verification task using different approaches. Separate results for male and female were computed. Maximum A posteriori Adaptation (MAP) from a general UBM from another database caused the EERs and minDCF to be higher than usual. Moreover, when applying *ensemble modeling* adaptation the results outperformed the previous baselines. Tables 1 and 2 show the same tendency for the noise conditions. The discriminative optimization improved the results better than the FA approaches in general. Once again, the best result was obtained using the *ensemble* method.

Figure 2 describes the relative improvements for Ahumada (male) database highlighting the unsupervised partition (consider it to be a more real scenario). *Ensemble modeling* resulted in greater improvements for the noise conditions.

Figure 3 shows the relative improvements for the clean and noise conditions. The results for the Gaudi (female) database are slightly better than the ones presented for Ahumada, but they follow the same tendency. *Ensemble modeling* approach provides a gain for both; moreover, we observe that the relative

[3] The reason to use microphone recordings is that the baseline models contain different channels.

| | *Clean* | | *Noise* | |
|---|---|---|---|---|
| | *EER* | *minDCF* | *EER* | *minDCF* |
| *Baseline MAP* | 17.35 | 9.6 | 22.52 | 15.3 |
| *MVE* | 15.24 | 8.9 | 18.64 | 11.0 |
| *JFA* | 13.82 | 7.7 | 20.32 | 13.1 |
| *Ensemble MVE U* | 10.32 | 6.8 | 12.18 | 7.2 |
| *Ensemble MVE S* | 8.63 | 6.3 | 11.6 | 6.9 |
| *Ensemble JFA U* | 11.25 | 7.1 | 15.24 | 8.6 |
| *Ensemble JFA S* | 10.50 | 6.9 | 13.12 | 7.7 |

**Table 1.** Final results (EER and minDCF) for the Ahumada database.

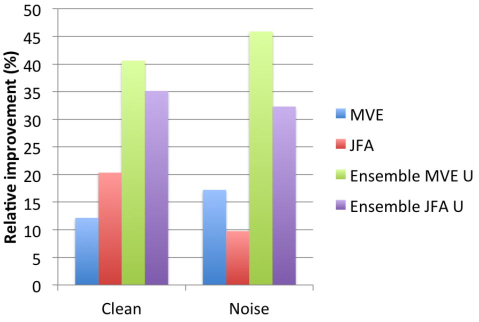


**Fig. 2.** Relative improvement summary for the Ahumada database.

improvement is higher for the noisy condition. Note also that the greater improvements are presented by the combination of MVE and *ensemble* approach. The discriminative optimization enhances both methodologies: JFA and MVE.

## 4 Conclusion

Adapting the model parameters from a baseline database to unknown data is a challenging issue. In this paper, we showed an *ensemble modeling* approach that naturally adapts for these new scenarios: new database or noise conditions. Moreover, Gaudi and Ahumada databases add a new problem since the adaptation has to deal with a new language. By partitioning the space in a first stage of the training, the new specific background models have shown to perform better adaptation than an isolated UBM. The results showed that the method improves the baseline, but more importantly, the adaptation of the new models gain specificity. For future research, the *ensemble* technique will decide the number of clusters in an autonomous way. Moreover, channel types may be examined in detail to have a complete system that can explored the data space at different levels.

| | Clean | | Noise | |
|---|---|---|---|---|
| | EER | minDCF | EER | minDCF |
| Baseline MAP | 16.5 | 9.3 | 21.26 | 14.7 |
| MVE | 14.4 | 7.9 | 17.45 | 10.6 |
| JFA | 12.4 | 6.6 | 19.1 | 12.4 |
| *Ensemble MVE U* | 9.2 | 6.0 | 11.08 | 6.6 |
| *Ensemble MVE S* | 8.9 | 5.8 | 10.88 | 6.3 |
| *Ensemble JFA U* | 11.05 | 6.8 | 13.65 | 7.7 |
| *Ensemble JFA S* | 10.94 | 6.5 | 13.20 | 7.2 |

**Table 2.** Final results (EER and minDCF) for the Gaudi database.

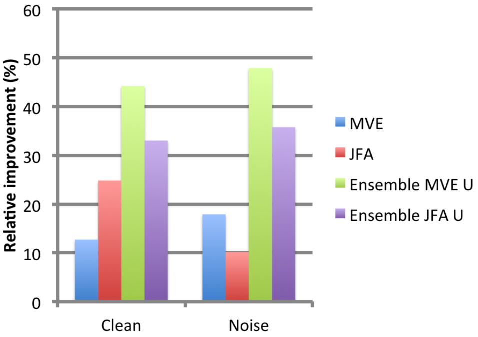


**Fig. 3.** Relative improvement summary for Gaudi database.

## 5 Appendix

The solution for the model parameters considers the following:

$$\mu = \sigma\,\tilde{\mu}, \qquad \tilde{\sigma} = \log(\sigma), \qquad w = \frac{\exp(\tilde{w})}{\sum_{k=1}^{K}\exp(\tilde{w})}.$$

For $\mu$, a specific $k$, dimension $d$ and to avoid bias, let

$$\nabla_\phi \ell(\chi, \tilde{\mu}_k) = \frac{1}{|\chi_C|}\sum_{X\in\chi_C} \gamma\ell(1-\ell)\frac{w_k\,\mathcal{N}(\chi|\mu_k,\sigma_k)}{\sum_{k'} w_{k'}\,\mathcal{N}(\chi|\mu_{k'},\sigma_{k'})}\left(\frac{\chi}{\sigma_k} - \tilde{\mu}_k\right). \tag{10}$$

For $\sigma_k$, let

$$\nabla_\phi \ell(\chi, \tilde{\sigma}_k) = \frac{1}{|\chi_C|}\sum_{X\in\chi_C} \gamma\ell(1-\ell)\frac{w_k\,\mathcal{N}(\chi|\mu_k,\sigma_k)}{\sum_{k'} w_{k'}\,\mathcal{N}(\chi|\mu_{k'},\sigma_{k'})}\left\{\left(\frac{\chi-\mu_k}{\tilde{\sigma}_k}\right)^2 - 1\right\}. \tag{11}$$

Finally, for the set of weights $w_k$,

$$\nabla_\phi \ell(\chi, \tilde{w}_k) = \frac{1}{|\chi_C|}\sum_{X\in\chi_C} \gamma\ell(1-\ell)\frac{\mathcal{N}(\chi|\mu_k,\sigma_k)}{\sum_{k'} w_{k'}\,\mathcal{N}(\chi|\mu_{k'},\sigma_{k'})}. \tag{12}$$

## References

1. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.
2. P. Kenny, P. Oueleet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, pp. 980–988, 2008.
3. L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," in *Proceedings of SRE11 Analysis Workshop*, 2011.
4. Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on.* IEEE, 2012, pp. 4253–4256.
5. T. Hasan and J. Hansen, "Integrated feature normalization and enhancement for robust speaker recognition using acoustic factor analysis," in *Proc. Interspeech*, 2012, pp. 1568–1571.
6. Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 6788–6791.
7. D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for ivector-based noisy speaker recognition," sumitted to *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, IEEE, 2014.
8. A. Sarkar and S. Umesh, "Multiple background models for speaker verification using the concept of vocal tract length and MLLR super-vector," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 351–364, 2012.
9. L. P. García-Perera, B. Raj, and J. A. Nolazco-Flores, "Ensemble approach in speaker verification." in *INTERSPEECH*, 2013, pp. 2455–2459.
10. Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE transactions on Speech*, vol. 43, pp. 781–785, August 1994.
11. B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.
12. N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.
13. N. Brummer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch: University of Stellenbosch, 2010.
14. J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "Ahumada: A large speech corpus in spanish for speaker characterization and identification," *Speech communication*, vol. 31, no. 2, pp. 255–264, 2000.
15. J. Ortega-Garcia, S. Cruz-Llanas, and J. Gonzalez-Rodriguez, "Facing severe channel variability in forensic speaker verification conditions." in *EUROSPEECH*, 1999, pp. 783–786.

# Using glottal parameters for detecting vocal pathology

Carlos Lázaro Carrascosa[1] and Pedro Gómez Vilda[2]

[1]Universidad Rey Juan Carlos, Móstoles, España
carlos.lazaro@urjc.es
[2]Universidad Politécnica de Madrid, Madrid, España
pedro@fi.upm.es

**Abstract.** This paper summarizes a comprehensive statistical study of a set of glottal parameters with the ultimate aim of measuring their capacity to discriminate and detect vocal pathology. The study is concentrated in the analysis of relevance of a set of parameters obtained from the analysis of phonated speech, specifically an open vowel as /a/. The speech signal was inversely filtered to obtain the glottal source, which on its turn was used to generate a set of 51 parameters, describing its biometrical and biomechanical properties. The study of relevance is based on correlation, factorial analysis, parametrical and non-parametrical hypothesis tests, with the aim of assessing the pathologic/normophonic condition of the speaker. The validation of the results is based on discriminant analysis. The conclusions allow establishing the most relevant parameters for pathological voice detection. These parameters get high classification rates.

**Keywords:** Glottal Parameters, Vocal Disease, Diagnostic Support, Discriminant Analysis

## 1 Introduction

Voice pathologies have become recently in a social problem that has reached a certain concern. Pollution in cities, smoking habits, air conditioning, etc. contributes to it. This problem is more relevant for professionals who use their voice frequently: speakers, singers, teachers, actors, telemarketeers, etc. Therefore techniques that are capable of drawing conclusions from a sample of recorded voice are of particular interest for the diagnosis as opposed to other invasive ones, involving exploration by laryngoscopes, fiber scopes or video endoscopes, which are techniques much less comfortable for patients. Voice quality analysis has come a long way in a relatively short period of time. In regard to the diagnosis of diseases, we have gone in the last fifteen years from working primarily with parameters extracted from the voice signal (both in time and frequency domains) and with scales as GRBAS [1] drawn from subjective assessments by experts, to produce more accurate evaluations with estimates derived from the glottal source. The importance of using the glottal source resides broadly in that this signal is linked to the state of the speaker's laryngeal structure. Unlike the voice signal (phonated speech) the glottal source, if conveniently

adfa, p. 1, 2011.

reconstructed using adaptive lattices, may be less influenced by the vocal tract [2]. As it is well known the vocal tract is related to the articulation of the spoken message and its influence complicates the process of voice pathology detection, unlike when using the reconstructed glottal source, where vocal tract influence has been almost completely removed.

The estimates of the glottal source have been obtained through inverse filtering techniques developed by our research group, based on the original work of Alku [3, 4]. We have also deepened into the nature of the glottal signal, dissecting it and relating it to the biomechanical parameters of the vocal folds, obtaining several estimates of items such as mass, loss or elasticity of cover and body of the vocal fold, among others. From the components of the glottal source also arise the so-called biometric parameters, related to the shape of the signal, which are themselves a biometric signature of the individual. We will also work with temporal parameters related to the different stages that are observed in the glottal signal during a cycle of phonation. Finally, we will take into consideration classical perturbation and energy parameters [2, 5].

In short, we have now a considerable amount of glottal parameters in multidimensional statistical basis, designed to be able to discriminate people with pathologic or dysphonic voices from those who do not show pathology.

On the same line, authors like Kuo [6] have worked with glottal parameters and Discriminant Analysis, with the aim of separating healthy voices and voices with presence of nodules, achieving success rates of 95%. Oliveira Rosa [7], meanwhile, has investigated about the discriminatory power of some single parameters, concluding that *jitter* is the best of them, with a rate above 54%.

This work addresses several issues: first, we will study the distribution of the parameters. We will take special care in the analysis of the difference between distributions from healthy subjects and the distributions from pathological subjects. To reach this goal we will use different statistical techniques such as: generation of descriptive items and diagrams, tests for normality and hypothesis testing, both parametric and nonparametric. These latter techniques consider the difference between the groups of healthy people and groups of people with an illness related to voice.

In addition, we are interested in finding statistical relationships between parameters. There are various reasons behind that: eliminate possible redundancies in the model, reduce the dimensionality of the problem and establish a criterion of relative importance in the parameters. The latter reason will be done in terms of discriminatory power for the criterion pathological/healthy. To this end, statistical techniques such as Bivariate Linear Correlation and Factor Analysis based on Principal Components [8,9] will be applied. Finally, we will use the well-known technique of Discriminant Analysis [8,10] classification applied to different combinations of parameters and factors to determine which of these combinations offers more promising success rates.

To perform the experiments we have used a balanced and robust database, consisting of two hundred speakers, one hundred of them males and one hundred females. We have also used a well-balanced proportion where subjects with vocal pathology as well as subjects who don´t have a vocal pathology are equally represented.

The different statistical analyses performed will allow us to determine what parameters contribute in a more decisive way in the detection of vocal pathology. The issue of parameter selection will be left for future studies. Therefore, some of the analyses will even allow us to present a ranking of the parameters based on their importance for the detection of vocal pathology. On the other hand, we will also conclude that it is sometimes desirable to perform a dimensionality reduction in order to improve the detection rates. Finally, detection rates themselves are perhaps the most important conclusion of the work, so this article will be mainly focused on them.

All the analyses presented in this work have been performed for each of the two genders in agreement with previous studies [11] showing that male and female genders should be treated independently, due to the observed functional differences between them.

## 2 Glottal Parameters

This work focuses primarily on the study of parameters extracted from the signal known as glottal source. Admittedly parameters obtained directly from the speech signal as perturbations in the fundamental frequency or amplitude, or as signal-to-noise ratios or relationships established between the first harmonics and formants have been effective in their ability for the detection of pathology. However the presence of the vocal tract in the analyzed signal provokes some contamination due to acoustic-phonetic influences that vocal tract produces itself [2]. To avoid this effect, our research group has managed to estimate the signal known as glottal source model eliminating the transfer function associated with the vocal tract through the use of inverse filtering techniques. The glottal source is a direct consequence of the biomechanical behavior of the vocal folds, and can decompose into two different signals: the average acoustic wave, associated with the movement of the body of the vocal fold and mucosal wave, associated with the movement of the cover of the vocal fold [5].

A set of new parameters were extracted from the glottal signals previously mentioned. These parameters can be classified according to their nature into the following groups [2]:

- Disturbance and energy parameters: the first relate to the differences of different characteristic values of the glottal source with respect to the average values collected in their own environment; second collect values related to the mucosa wave or acoustic wave energy. We can see the list in Table 1.

**Table 1.** Perturbation and energy parameters used in the study. Parameters 1-7 are estimated from the glottal source. 8 is estimated from the mucosal wave (MWC) and the average acoustic wave (AAW). 9-14 are estimated from the mucosal wave. The set numbering is internal.

| Nº | Disturbance and energy parameters |
|---|---|
| 1 | Pitch |
| 2 | Jitter |
| 3 | Amplitude Shimmer |
| 4 | Slenderness Shimmer |
| 5 | Area Shimmer |

| Nº | Disturbance and energy parameters |
|---|---|
| 6 | Ratio Diff. Closure Spike Amp. Neighbours Avg. Value |
| 7 | Ratio Diff. Closure Spike Slenderness Amp. Nbr. Avg. Value |
| 8 | Ratio Energy MWC/AAW |
| 9 | Ratio 2nd. Harm / Fundamental MWC |
| 10 | Ratio 2nd. Harm. / 1st. Harm MWC |
| 11 | 1st. Energy Bin MWC |
| 12 | 2nd. Energy Bin MWC |
| 13 | 3rd. Energy Bin MWC |
| 14 | 4th. Energy Bin MWC |

- Spectral or biometric parameters: related to the characteristic shape associated to the envelope of the glottal source power spectral density. We can see the list in Table 2.

**Table 2.** Biometric parameters used in the study. All parameters are estimated from the envelope of the power spectral density (PSD) of the glottal source. The set numbering is internal, and takes into account other parameters that are working on this work.

| Nº | Biometrical Parameter | Nº | Biometrical Parameter |
|---|---|---|---|
| 15 | PSD Origin Amplitude Relative | 25 | PSD Initial Min. Position Relative |
| 16 | PSD Initial Min. Amplitude Relative | 26 | PSD 1st Max. Position ABS |
| 17 | PSD 1st Max. Amplitude ABS | 27 | PSD 1st Min. Position Relative |
| 18 | PSD 1st Min. Amplitude Relative | 28 | PSD 2nd Max. Position Relative |
| 19 | PSD 2nd Max. Amplitude Relative | 29 | PSD 3rd Max. Position Relative |
| 20 | PSD 3rd Max. Amplitude Relative | 30 | PSD 2nd Min. Position Relative |
| 21 | PSD 2nd Min. Amplitude Relative | 31 | PSD 4th Max. Position Relative |
| 22 | PSD 4th Max. Amplitude Relative | 32 | PSD End Val. Position Relative |
| 23 | PSD End Value Amplitude Relative | 33 | Slenderness of the 1st V notch |
| 24 | PSD Origin Position Relative | 34 | Slenderness of the 2nd V notch |

- Biomechanical parameters: related to mechanical properties directly associated with the vocal folds: mass loss and elasticities and their unbalances. Referred to both the body and the cover of the vocal fold. We can see the list in Table 3.

**Table 3.** Biomechanical parameters used in the study. 35-40 are estimated from the envelope of the power spectral density of the average acoustic wave. 41-46 are estimated from the envelope of the power spectral density mucosal wave. The set numbering is internal, and takes into account other parameters that are working on this study.

| Nº | Biomechanical parameter | Nº | Biomechanical parameter |
|---|---|---|---|
| 35 | Body Mass | 41 | Cover Mass |
| 36 | Body Losses | 42 | Cover Losses |
| 37 | Body Stiffness | 43 | Cover Stiffness |
| 38 | Body Mass Unbalance | 44 | Cover Mass Unbalance |
| 39 | Body Losses Unbalance | 45 | Cover Losses Unbalance |
| 40 | Body Stiffness Unbalance | 46 | Cover Stiffness Unbalance |

- Temporal parameters: related to singular points extracted from the time domain representation of a cycle of the glottal source phonation. We can see the list in Table 4.

**Table 4.** Temporal parameters used in the study. All parameters are estimated from the mucosal wave and its derivative. The set numbering is internal, and takes into account other parameters that are working on this study.

| Nº | Temporal Parameter |
|---|---|
| 47 | Open Phase Time |
| 48 | Closed Phase Time |
| 49 | Recovery Phase Time |
| 50 | 1st Max. Position Mucosal Wave Correlate |
| 51 | 2nd Max. Position Mucosal Wave Correlate |

## 3 Materials and methods

The protocol used in the experimentation is based on the following steps:

1. Descriptive analysis of the parameters, paying special attention to the separation between the two main groups in our study: people with speech pathology and people free from speech pathology.
2. Tests of normality of the parameters, taking into account the distinction made in the previous section.
3. Mean difference parametric test (Student's t) between the two groups mentioned above. The extent complementary effect size analysis was also included to assess the relevance of the differences found. Levene's test for homogeneity of variances.
4. Nonparametric contrasts between the two groups mentioned. In particular, tests of Mann-Whitney, Moses, Kolmogorov-Smirnov and Wald-Wolfowitz. These tests, as well as the ones mentioned in the previous point were performed on all parameters, in order to facilitate the comparison of results.
5. Bivariate linear correlation studies between the different parameters, and also between each of them and the criterion of belonging to the group of healthy subjects.
6. Factor Analysis based on Principal Components applied on the full set of parameters, and also on significant subsets thereof.
7. Discriminant Analysis on the full set of parameters and on significant subsets thereof. We will also apply this analysis to the factors obtained in the previous section.

The last of the steps above is the core of the study, on which it is mainly based, for two main reasons: it offers detection rates, one of the main objectives we pursued. Besides, Discriminant Analysis is related with the previous techniques performed, since the results given by those are frequently used to determine the inputs of the Discriminant Analysis. For example, we have carried a test out using the parameters whose effect size rank has achieved the highest values. We have also performed a test using the parameters whose correlation index is highest, and so on.[1]

The database used for the study is a corpus of 200 subjects (100 male and 100 female), created by our research group, which evenly distributes the presence and absence of vocal pathology. Thus, the database can be described as: 50 non-pathological female subjects, 50 pathological female subjects, 50 non-pathological male subjects and 50 male pathological subjects. This data will be split in two sets for Discriminant Analysis: train and test. We have tried several sizes for the sets, in different experiments, but the more frequently proportion used is ninety per cent for training - ten per cent for testing.

The majority of studies analyzes data collected from male and from female independently. In general, the need for male and female samples to reach general conclusions is well known [11]; also the research in this field recommends taking into consideration the different physiological characteristics of both genders. The ages of the subjects range from 19 to 56, with a mean of 30.02 years and a standard deviation of 9.94 years. Normal condition has been determined by electroglottography, video-endoscopy and evaluation GRABS. Also, previously, some criteria that patients must meet to fulfill such a condition were established:

- self-report not having any laryngeal pathology;
- have a voice according to gender, age and cultural group speaker, plus a suitable pitch, timbre, volume and flexibility of diction;
- be a non-smoker;
- no history of any surgery related to any laryngeal pathology.
- no history of endotracheal intubation in the last year.

Concerning pathological cases, the sample set contains about the same number of mild (functional) and moderate pathology grades, comprising polyps, nodules and Reinke’s edemae. The recording protocol comprises three different utterances of the vowel /ah/ with duration longer than three seconds, taken with a headset microphone placed 10 cm from the speaker's mouth. The signals were sampled at 44100 Hz, lately undersampled to 22050 Hz. The resolution used was 16 bits. The recordings were performed using the external sound board Creative Studio. Segments of 0.2 seconds long were extracted from the central part (the more stable) for analysis and parameterization. These segments included a variable number of cycles of phonation: about twenty in male voices, with a fundamental frequency of 100 Hz and about forty in female voices, with a fundamental frequency of 200 Hz [5].

[1] The full details of the work exceed the scope of this article. They will be addressed in future works.

## 4 Results

This section is dedicated to show the results from the most important study: Discriminant Analysis. This study aims to measure the classification ability from the different combinations of parameters. It also measures factors used to consider the fact of having or not having a vocal pathology.

These combinations were chosen based on three main reasons: first, the use of parameters and factors from the parameters with no prior assumption; second, the nature of the parameters and factors, i.e., the family from which they come from: perturbation/energy, biometric, biomechanical and temporal; third, the statistical results of previous studies made. Remarkably, we performed a double study in the latter case, based first in considering the best parameters of each test (and the factors extracted from these parameters), and secondly to consider all parameters (and related factors) except the worst from each test. Only in the study related to boxplots it was not possible to make this distinction, because only two categories for qualitative parameters are set in the mentioned study.

We will highlight one study that has been carried out using all parameters except the one labeled with code 024 (mucosal wave power spectral density origin position, relative value). The reason to exclude this parameter was found in the study related to partial correlation. This study advises the exclusion of the parameter. The list of the studies that have been conducted can be seen in Table 5, which also reports the average detection rate achieved in each case (this overall rate has been calculated as the arithmetic mean between the rate for selected cases, the rate for unselected cases and the rate for a process of cross-validation). The latter rate is also presented in Table 5, as a measure of additional test. Finally, we would like to emphasize that the whole study is divided into two parts: male and female data.

**Table 5.** Tests and detection rates obtained for the bases of male and female data

| TEST | Female data | | Male data | |
|---|---|---|---|---|
| | Avg. Value (%) | Cross Valid. (%) | Avg. Value (%) | Cross Valid. (%) |
| All the parameters | 90.8 | 86.4 | 88.2 | 82.1 |
| All the factors | 95.9 | 95.6 | 87.4 | 86.5 |
| PERTURBATION/ENERGY Factors | 84.4 | 84.8 | 82.5 | 82.1 |
| BIOMETRICS Factors | 87.2 | 87.0 | 73.5 | 73.1 |
| BIOMECHANICS Factors | 86.0 | 85.2 | 82.2 | 82.2 |
| TEMPORAL Factors | 73.7 | 73.1 | 65.5 | 65.2 |
| PERTURBATION/ENERGY Parameters | 85.8 | 84.9 | 81.6 | 80.3 |
| BIOMETRICS Parameters | 89.5 | 86.6 | 85.5 | 79.4 |

| TEST | Female data | | Male data | |
|---|---|---|---|---|
| | Avg. Value (%) | Cross Valid. (%) | Avg. Value (%) | Cross Valid. (%) |
| BIOMECHANICS Parameters | 91.1 | 89.8 | 86.1 | 83.8 |
| TEMPORAL Parameters | 77.2 | 75.9 | 69.8 | 68.3 |
| NO PERTURBATION/ENERGY Parameters | 90.0 | 86.0 | 87.1 | 82.4 |
| NO BIOMETRICS Parameters | 93.2 | 89.9 | 87.2 | 84.3 |
| NO BIOMECHANICS Parameters | 91.4 | 87.8 | 83.1 | 75.4 |
| NO TEMPORAL Parameters | 90.3 | 87.0 | 90.0 | 85.8 |
| NO PERTURBATION/ENERGY Factors | 90.1 | 88.9 | 81.4 | 80.7 |
| NO BIOMETRICS Factors | 93.8 | 92.9 | 86.7 | 86.6 |
| NO BIOMECHANICS Factors | 94.1 | 93.5 | 86.7 | 85.4 |
| NO TEMPORAL Factors | 93.1 | 92.2 | 85.3 | 83.5 |
| ALL Parameters EXCEPT 24 | 91.0 | 88.1 | 88.1 | 82.8 |
| ALL Factors EXCEPT 24 | 95.9 | 95.6 | 87.3 | 86.0 |
| Best Parameters BOXPLOT | 92.9 | 91.6 | 87.4 | 87.3 |
| Best Factors BOXPLOT | 92.1 | 91.6 | 85.3 | 85.0 |
| Best Parameters DISTRIBUTION DIFFERENCE | 91.6 | 88.6 | 87.3 | 86.2 |
| Best Factors DISTRIBUTION DIFF. | 92.9 | 92.9 | 84.7 | 83.5 |
| Best Parameters EFFECT SIZE | 92.4 | 91.4 | 83.7 | 81.3 |
| Best Factors EFFECT SIZE | 91.1 | 90.5 | 82.0 | 80.8 |
| Best Parameters HEALTH CORRELATION | 88.3 | 87.1 | 84.7 | 84.1 |
| Best Factors HEALTH CORR. | 87.7 | 87.5 | 85.1 | 85.1 |
| Worst Param. Excluded DISTRIBUTION DIFF. | 91.1 | 87.1 | 86.4 | 83.1 |
| Worst Factors Excluded DISTRIBUTION DIFF. | 95.5 | 94.5 | 88.0 | 87.7 |
| Worst Parameters excluded EFFECT SIZE | 92.2 | 87.5 | 88.8 | 83.4 |
| Worst Factors Excluded EFFECT SIZE | 94.8 | 94.0 | 86.8 | 87.5 |
| Worst Parameters Excluded HEALTH CORR. | 91.5 | 88.1 | 86.4 | 82.0 |
| Worst Factors Excluded HEALTH CORR. | 94.4 | 93.2 | 87.2 | 86.9 |

The results of the different Discriminant Analysis studies will allow us to conclude that for the female database, the best combinations obtained are:

1. Using the factors extracted from all the glottal parameters except labeling number 24[2]: mucosal wave power spectral density origin position, relative value (overall measure: 95.9%).
2. Using the factors extracted from all the glottal parameters (overall measure: 95.9%).
3. Using the factors extracted from all considered parameters except the worst from the study that takes into account the difference in the distributions: slenderness shimmer and ratio mucosal wave correlate / average acoustic wave energy, labeled with the numbers 4 and 8, respectively (overall measure: 95.5%).

Regarding the male database, the best combinations obtained are:

1. Using all the glottal parameters except the so-called temporals (overall measure: 90.0%).
2. Using all parameters except the worst seen from the study that takes into account the effect size: amplitude of the first maximum, slenderness shimmer, slenderness of the first groove, vocal fold body mass, vocal fold body elasticity, ratio of the difference between the closure spike amplitude of neighbor cycles with respect to their average value and ratio mucosal wave correlate / average acoustic wave energy, labeled with the numbers 17, 4, 33, 35, 37, 6 and 8, respectively (overall measure: 88.8%).
3. Using all the glottal parameters (overall measure: 88.2%).

The results show some interesting issues:

- We obtain better rates with female data than with male data, in all cases.
- Cross validation rates are quite high, they are comparable to the average values.
- Temporal parameters seem to have less discriminatory power than those of the other categories.
- Factors work better with female data; parameters work better with male data.
- The best results are achieved using a large number of parameters (or factors). In fact, the use of all parameters (or factors) reaches one of the best results in both databases. The statistical experiments performed can adjust these results, sometimes improving the rates obtained by eliminating the least favorable parameters for each study.
- Not contradicting the previous point, the categories of parameters considered are able to obtain high classification rates by themselves.

## 5 Conclusions

We have carried out a comprehensive statistical study of glottal parameters using a database composed of individuals of both genders. The study has mainly focused on

[2] Recall that the motivation to eliminate this parameter is found in studies of partial correlation.

finding out the most suitable parameters to discriminate pathological and non-pathological voices, using the Discriminant Analysis technique.

Thus, the above mentioned technique applied to different combinations of the parameters and factors allows us to conclude that the use of these parameters reached high rates of success in regard to the detection of pathologies of the vocal apparatus.

Specifically, for the female database, the combination that provides a higher rate of success is the one that considers extracted factors from all parameters used in the study except the so-called "mucosal wave power spectral density origin position, relative value". On the other hand, the best combination of male data is achieved when all parameters are used except the so-called temporal.

## References

1. Hirano, M. et al., 1988. Acoustic analysis of pathological voice. Some results of clinical application. Acta Otolaryngologica 105 (5-6), pp. 432-438.
2. Gómez, P., et al., 2009. Glottal Source biometrical signature for voice pathology detection. Speech Communication, 51. pp. 759-781.
3. Alku, P., 1992. An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulseform. ICASSP'92, pp. II/29-32.
4. Alku, P., 2003. Parameterisation methods of the glottal flow estimated by inverse filtering. In: Proc. VOQUAL'03, pp. 81–87.
5. Gómez, P., et al., 2005. PCA of perturbation parameters in voice pathology detection. Actas de INTERSPEECH-2005, pp. 645-648.
6. Kuo, J. et al., 1999. Discriminating speakers with vocal nodules using aerodynamic and acoustic features. Proc. ICASSP'99, Vol. 1, pp. 77–80.
7. De Oliveira Rosa, M., Pereira, J.C., Grellet, M., 2000. Adaptive estimation of residue signal for voice pathology diagnosis. IEEE Trans. Biomed. Eng. 47 (1), pp. 96-104
8. Härdle, W., & Simar, L. (2007). Applied multivariate statistical analysis (Vol. 22007). Berlin: Springer.
9. Ho, R., 2006. Handbook of univariate and multivariate data analysis and interpretation with SPSS. CRC Press.
10. Klecka, W. R. (Ed), 1980. Discriminant analysis (No. 19). Sage.
11. M. Teresa Ruiz y Lois M. Verbrugge, 1997. A two way view of gender bias in medicine. Journal of Epidemiology and Community Health. Vol. 51, pp.106-109.

# Speaker Clustering for Variability Subspace Estimation

Alicia Lozano-Diez, Ivan Gomez-Piris, Javier Franco-Pedroso, Javier Gonzalez-Dominguez, and Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group
Universidad Autonoma de Madrid (UAM). Spain
{alicia.lozano@uam.es,iva.gomez@estudiante.uam.es,javier.franco@uam.es,
javier.gonzalez@uam.es,joaquin.gonzalez@uam.es}

**Abstract.** This work analyses the use of clustering as an unsupervised approach to exploit unlabelled speech in speaker verification. Particularly, we apply Agglomerative Hierarchical Clustering (AHC) over a large set of unlabelled i-vectors to discover the underlying speaker identities they belong to. Then, we use those new learned identities as the ground truth for a subsequent supervised stage based on Linear Discriminant Analysis. Experiments are conducted on data belonging to NIST SRE 2012 following the protocol defined in the related NIST i-vector Machine Learning Challenge 2014. Results show how the proposed unsupervised stage properly exploits unlabelled speech data without a significant loss of performance ($\sim 7\%$ in terms of EER) with respect to a baseline system where correct labels are provided.

**Keywords:** Clustering, Speaker Recognition, AHC, LDA.

## 1 Introduction

Most of the approaches of state-of-the-art speaker recognition systems rely on speaker labelled data. However, the labelling process implies a big amount of resources as human assistance, specially when large databases are needed to achieve high-accurate performance.

We analyse in this work the use of clustering as an unsupervised stage to group speaker identities from unlabelled data. Those new learned labels are used as a ground-truth for a subsequent supervised scheme.

This issue has been recently approached in the NIST i-vector Machine Learning Challenge [14, 7], where participants were asked to develop speaker recognition systems from unlabelled i-vectors [4].

Motivated by this, we present an analysis of the influence of a clustering stage in order to label i-vectors to train a speaker recognition system based on cosine scoring over channel-compensated i-vectors by using Linear Discriminant Analysis (LDA) technique. Moreover, this could be extend to use its probabilistic variant, Probabilistic Linear Discriminant Analysis (PLDA) [11], widely used to deal with speaker recognition tasks.

As reference, a system from manually labelled i-vectors has been considered and compared with the performance of the same system but from automatically labelled data.

In order to evaluate the performance of our clustering algorithm, we use a subset belonging to NIST SRE'12 [13, 7] to perform the experiments.

The rest of this paper is organized as follows. The clustering algorithm used is introduced in Section 2 and the i-vector extraction process is briefly described in Section 3. Then, the speaker recognition systems and the performance measures employed are presented in Sections 4 and 5, respectively. Finally, in Sections 6 and 7, results and conclusions of this work are presented.

## 2 Agglomerative Hierarchical Clustering

The Agglomerative Hierarchical Clustering (AHC) algorithm is a well-known unsupervised algorithm used in many machine learning tasks. We briefly describe its basis in this section.

As an agglomerative approach, each observation is initialised as a cluster or group. Then, those clusters are iteratively grouped in pairs according to a defined distance (or similarity) measure, so that the two closest groups create a new cluster that groups both of them.

Several distance measures can be chosen to compute the distance between two data points $(x_i, x_j)$. Most common ones are cosine distance and euclidean distance, defined as follows:

$$d_{cosine} = 1 - \frac{x_i x_j^t}{\sqrt{(x_i x_i^t)(x_j x_j^t)}} \tag{1}$$

$$d_{euclidean} = \sqrt{(x_i - x_j)(x_i - x_j)^t} \tag{2}$$

Another key aspect of AHC is the method used to group two clusters ($r$ and $s$). This method is known as *linkage* method. We chose the Unweighted Pair Group Method With Averaging (UPGMA, Eq. 3) and its weighted version (WPGMA, Eq. 4) due to the fact that both methods provide robustness against *outliers*. These two distances are defined as:

$$d_{UPGMA}(r, s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} d(x_{ri}, x_{sj}) \tag{3}$$

$$d_{WPGMA}(r, s) = \frac{d(p, s) + d(q, s)}{2} \tag{4}$$

where $n_r$ and $n_s$ are the number of objects in clusters $r$ and $s$, respectively, and $x_{ri}$ is the *ith* data point in cluster $r$. Finally, it is assumed that cluster $r$ was created by combining clusters $p$ and $q$, so that the definition of the WPGMA distance is recursive.

As an *stopping criterion* to get final cluster, we considered the following two. Firstly, the maximum number of final clusters was provided considering that the number of speakers is known. However, if we want to perform a complete unsupervised clustering stage, this number should be considered unknown too. Then, the maximum distance among clusters to group was used in the second set of experiments.

The main issue of this last criterion is how to fix the value of that maximum distance to reach good clustering performance. In this work, that parameter was set empirically according to nearest neighbours distances in the development dataset.

## 3 I-vectors Extraction

The i-vectors extraction process used in this work is as follows.

First, a Wiener filter (ICSI [2]) is used to reduce noise of audio files, followed by a stage of Voice Activity Detection (VAD) consisting in a combination of a simple VAD based on energy and a VAD SoX tool [1]. Then, a Hamming window of 20 ms with 10ms of overlap is applied. Resulting frames of speech are parameterized (19 MFCC + c0 + delta) and channel-compensated (Cepstral Mean Normalization [5], RASTA [8] and Feature Warping are applied [15]).

Then, a gender independent Universal Background Model (UBM) of 2048 Gaussian components is trained using a subset of the development data from NIST Speaker Recognition Evaluations (from SRE'04 to SRE'10).

Finally, a Total Variability scheme was employed. This scheme shares the same principles as Joint Factor Analysis (JFA) systems [10, 16], 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* (represented by a low-rank T matrix), which jointly includes speaker and session variability is computed instead of computing two separate subspaces as in JFA (matrices U and V). The T matrix (600 dimensions) obtained is gender independent, and it was trained with the same data used to train the UBM.

Furthermore, *whitening* and *length normalization* [6] were applied to the resulting i-vectors in order to deal with non-Gaussian behaviour of i-vectors distribution and palliate duration variability of utterances.

## 4 Speaker Recognition System

In this work, a Linear Discriminant Analysis (LDA) modelling has been used in order to compensate the speaker inter-variability. Thereby, a LDA matrix of 400 dimensions was trained with the development i-vectors and their corresponding speaker identity labels (manually or automatically obtained). Then, the development, model and test i-vectors are projected to this new space and a simple cosine score over the resulting transformed data has been performed.

It should be pointed out that as the SRE'12 task considered multi-training, i.e. more than one i-vector to obtain the speaker model, each speaker model has been obtained by averaging all the i-vectors belonging to the same speaker.

## 5 Performance Measures

Two stages of the speaker recognition system are separately evaluated: the clustering algorithm and the speaker verification task.

### 5.1 Clustering Performance

Two measures have been selected to measure how well the i-vectors are grouped according to speaker identities, both of them described in [12].

- *Cluster impurity*, a measure of how many speakers have been grouped within the same cluster.
- *Class impurity*, which shows whether the i-vectors of the same speaker have been assigned to just one cluster or more.

Both are complementary, so the best working point is considered as the one where both of them are minimum over the dataset.

### 5.2 Speaker Recognition Systems Performance

In order to measure the performance achieved by the whole speaker recognition system when facing the speaker verification task, three metrics have been considered: the Equal Error Rate (EER); the minimum of the decision cost function (DCF) defined as The NIST i-vector Machine Learning Challenge 2014 [14]; and, finally, the minimum of the log-likelihood ratio cost (minCllr) [3].

## 6 Experiments and Results

### 6.1 Experimental Framework

We perform experiments using a subset of the NIST SRE'12 database. In particular, three separate subsets have been considered:

- *Development set:* Used to train the LDA matrix. It consists of 38766 utterances of approximately 10, 30 and 150 seconds of duration from 1775 speakers. These utterances have been automatically labelled by using AHC algorithm to performed the experiments. For some experiments, i-vectors computed from utterances of less than 30 seconds of duration have been discarded. This was motivated by the fact that i-vectors coming from short utterances are less reliable [9] and could be affecting the performance of the clustering algorithm or the LDA training stage. When these utterances are discarded, 27401 i-vectors remain for the development of the system and the number of speakers remains the same. For the whitening of the data in the speaker recognition system, the whole development set has been used.

It should be taken into account that this dataset is a subset of the development set provided by NIST in SRE'04, SRE'05, SRE'06, SRE'08 and SRE'10, and half of the speakers given as training data in SRE'12. All the speakers considered satisfy the constraint of having at least 8 utterances.

– *Train set or model set:* Used to create the speaker models by averaging the i-vectors corresponding to each target speaker. It has 14693 utterances belonging to 959 different speakers, which correspond to the other half of speakers from the training data of SRE'12 that have not been used in the development set.
– *Test set:* It consists of 16218 utterances of 300 seconds of duration used to evaluate the performance of the speaker recognition system. These data come from the test dataset of NIST SRE'12.

All the speech considered in this work corresponds to telephone speech, and therefore, there is not channel mismatch among development, train and test data.

In this work, we applied clustering over i-vectors. The cosine distance considers only the angle between the two i-vectors and, then, provides independence of the i-vector magnitude, which is believed to be related to the channel information [4]. Thus, removing the magnitude improves the robustness of the i-vector system, so that cosine distance has been used in the experimental part of this work. The euclidean distance (Eq. 2) was also used performing a previous stage of length normalization over the i-vectors, but the performance was similar to the case of cosine distance, so the last one was selected.

Regarding the *linkage* method, the performance of initial experiments was slightly better with the UPGMA method, so that most of the experiments of this work are based on this type of linkage.

### 6.2 Baseline and Reference Systems

Considering the problem of no having development labels and that we just use the training and the test datasets to perform the speaker recognition task, the best we can do is to compute the cosine score between the model and the test i-vectors. This system will be considered the baseline that we want to outperform by using the development dataset. The results with this kind of scoring are shown in Table 1.

As reference system to compare with, we consider a simple LDA-based speaker recognition system, where the LDA matrix has been trained with real (manually labelled) speaker identity labels. The results of this system considering different development subsets are shown in Table 1. These results represent a lower bound, i.e., the best we can do with an ideal clustering.

| | **EER(%)** | **minDCF** | **minCllr** |
|---|---|---|---|
| **Baseline** | 7.13 | 0.510 | 0.233 |
| **Ref. 1** | 5.61 | 0.451 | 0.187 |
| **Ref. 2** | **5.32** | **0.436** | **0.180** |

Table 1: Baseline system based on cosine scoring and reference systems (Ref. 1 and Ref. 2) based on LDA-cosine scoring. Ref. 1 system was trained with the whole development set while Ref. 2 system was train with utterances of more than 30 seconds of duration.

### 6.3 Experiments over Automatically Labelled i-vectors

We have considered a first preliminary set of experiments where we suppose that the number of speakers is known. This means that the stopping criterion used in the AHC algorithm is the maximum number of clusters to create (equal to 1775 in our case). These experiments have been performed in order to select the parameters of the clustering algorithm, which are the distance measure and the type of *linkage* method. Results of the LDA system can be seen in Table 2. The whole development dataset has been employed. In the case of the euclidean distance experiments, i-vectors have been previously length normalized.

The best results were obtained when using cosine distance and UPGMA as linkage method or euclidean distance and WPGMA. Since the results are pretty similar for these two configurations and the typical distance measure in the i-vector space is the cosine distance, we select this for the further experiments in this paper.

In the second set of experiments, the number of speakers is considered unknown. Then, the stopping criterion chosen is the maximum distance to continue grouping clusters. The value of this parameter (known as *cutoff*) was heuristically obtained, although the range of values was limited by computing distances to nearest neighbors in the development set.

Different experiments were performed according to this. The specific configuration of each experiment is shown in Table 3, as well as the resulting impurity values of the clustering obtained. Results can be seen in Table 4. The first row (EXP 1) corresponds to the results obtained by using the whole development dataset with the value of *cutoff* that gave us the best performance. In the second

| **AHC params** | | **LDA-based System** | | |
|---|---|---|---|---|
| **Distance** | **Linkage** | **EER(%)** | **minDCF** | **minCllr** |
| **Cosine** | **UPGMA** | 6.93 | **0.472** | **0.220** |
| | **WPGMA** | 6.95 | 0.504 | 0.226 |
| **Euclidean** | **UPGMA** | 7.13 | 0.480 | 0.231 |
| | **WPGMA** | **6.26** | 0.500 | **0.220** |

Table 2: Performance of LDA-based system considering that the number of speakers is known when performing the clustering stage.

| | **Clust.Data** | **Pre-Whitening** | **Cutoff** | **#Clusters** | **Dev.Data** | **Impurities** |
|---|---|---|---|---|---|---|
| **EXP 1** | All | No | 0.76 | 3396 | All | 0.29 |
| **EXP 2** | All | Yes | 0.91 | 2669 | All | 0.21 |
| **EXP 3** | All | Yes | 0.91 | 1573 | Clusters≥ 8 | 0.12 / 0.19 |
| **EXP 4** | >30s | Yes | 0.88 | 2120 | >30s | **0.12** |

Table 3: Configuration and resulting impurities of each experiment when the number of speakers is considered unknown. It should be pointed out that class and cluster impurities are the same in the chosen points for experiments 1, 2 and 3.

| | **EER(%)** | **minDCF** | **minCllr** |
|---|---|---|---|
| **Baseline** | 7.13 | 0.510 | 0.233 |
| **EXP 1** | 7.09 | 0.494 | 0.226 |
| **EXP 2** | **5.91** | 0.467 | **0.198** |
| **EXP 3** | 5.92 | **0.455** | **0.198** |
| **EXP 4** | **5.91** | 0.463 | **0.198** |
| **Ref. 2** | **5.32** | **0.436** | **0.180** |

Table 4: Performance of LDA-based system considering that the number of speakers is unknown when performing the clustering stage.

experiment (EXP 2), a whitening operator was applied to i-vectors previously to the clustering stage, which has shown to be advantageous for the algorithm. In the third one (EXP 3), clusters with less than 8 i-vectors were discarded in order to remove what could be noisy data. And finally, the last row (EXP 4) of the table shows the results obtained when data of less than 30 seconds were discarded and development data were whitened before the clustering stage.

Additionally, impurity curves were computed, and the value of *cutoff* that minimized both cluster and class impurities was the same that the one that gave us the best performance after several experiments with different values of that parameter.

The mentioned curves are presented in Figure 1. According to them, both applying the whitening operator and discarding short i-vectors resulted in improvements of the performance of AHC algorithm. This leads to improvements in the performance of the speaker recognition systems, as we can see in Table 4.

## 7 Conclusions

In this work, we have presented the use of a well-known hierarchical clustering (AHC) approach to group i-vectors according to speaker identity. Once they have been labelled, a speaker recognition model has been trained and evaluated. The results have been compared to a reference system that has been trained using the actual speaker labels provided by NIST.

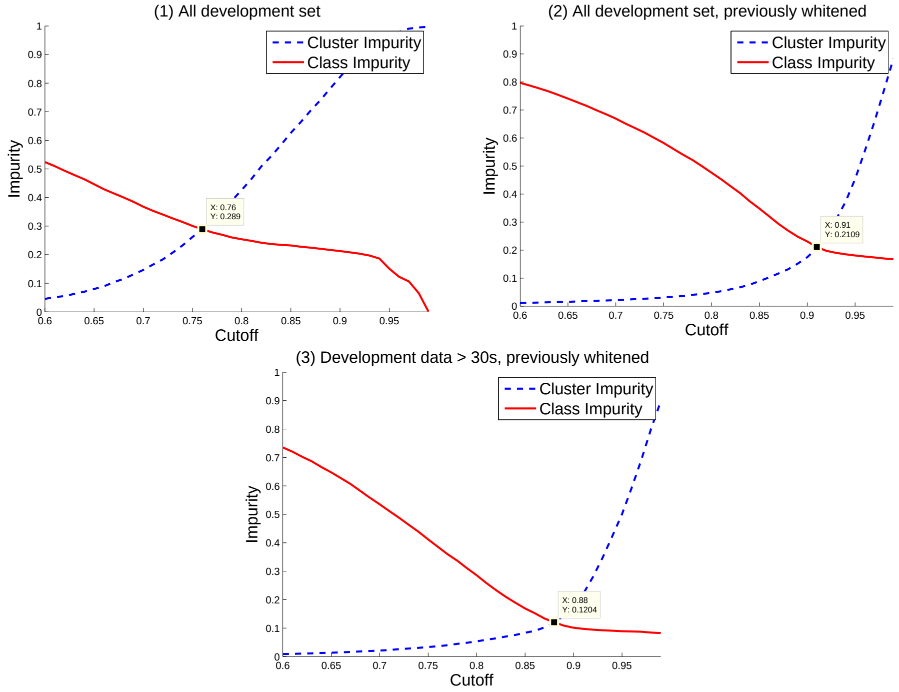


Fig. 1: Cluster and class impurities for different values of *cutoff* parameter of AHC algorithm. Each graphic corresponds to a different development set and different pre-processing of data according to the developed experiments: (1) all development set without any pre-processing stage; (2) all development set, previously whitened; (3) just data of more than 30 seconds, previously whitened.

It should be highlighted that the performance of the system trained with estimated labels is pretty similar to the one obtained with the real labels. This can be considered as an important advantage since it allows taking advantage of unlabelled data to train speaker recognition systems with little loss in terms of performance. It also means that the system is robust to incorrect labelled data. Moreover, it has been observed that this approach presents robustness against the number of clusters created as well and it does not seem a critical factor when referring to speaker recognition system performance.

Furthermore, the performance yield by LDA-based systems trained over estimated labels is higher than the simple cosine scoring (baseline system), which means that using development data is beneficial to the speaker recognition task even though the speaker clusters are not completely accurate.

As future work, we propose to extend this to more complex speaker recognition systems, such as Probabilistic Linear Discriminant Analysis (PLDA), and also, to perform experiments using other techniques such as Support Vector Machines (SVM).

**Acknowledgments.** This work has been developed within the project *CMCV2: Caracterización, Modelado y Compensación de Variabilidad en la Señal de Voz* (TEC2012-37585-C02-01), funded by *Ministerio de Economía y Competitividad*, Spain.

## References

1. http://sox.sourceforge.net/
2. Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivadas, S.: Qualcomm-icsi-ogi features for asr. In: Proc. ICSLP. pp. 4–7 (2002)
3. Brmmer, N., du Preez, J.: Application-independent evaluation of speaker detection. Computer Speech & Language 20(23), 230 – 275 (2006), http://www.sciencedirect.com/science/article/pii/S0885230805000483, odyssey 2004: The speaker and Language Recognition Workshop Odyssey-04 Odyssey 2004: The speaker and Language Recognition Workshop
4. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech & Language Processing 19(4), 788–798 (2011)
5. Furui, S.: Cepstral analysis technique for automatic speaker verification. Acoustics, Speech and Signal Processing, IEEE Transactions on 29(2), 254–272 (Apr 1981)
6. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of i-vector length normalization in speaker recognition systems. In: INTERSPEECH. pp. 249–252 (2011)
7. Gonzalez-Rodriguez, J.: Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014). Loquens 1(1) (2014), http://loquens.revistas.csic.es/index.php/loquens/article/view/9/21
8. Hermansky, H., Morgan, N.: Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on 2(4), 578–589 (Oct 1994)
9. Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M.: i-vector based speaker recognition on short utterances. In: INTERSPEECH. pp. 2341–2344 (2011)
10. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. Speech and Audio Processing, IEEE Transactions on 13(3), 345–354 (May 2005)
11. Kenny, P., Stafylakis, T., Ouellet, P., Alam, M., Dumouchel, P.: Plda for speaker verification with utterances of arbitrary duration. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 7649–7653 (May 2013)
12. van Leeuwen, D.A.: Speaker linking in large data sets. In: Odyssey. p. 35 (2010)
13. NIST: The nist year 2012 speaker recognition evaluation plan (2012), http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf
14. NIST: The 2013-2014 speaker recognition i-vector machine learning challenge (2013), http://www.nist.gov/itl/iad/mig/upload/sre-ivectorchallenge_2013-11-18_r0.pdf
15. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: 2001: A Speaker Odyssey - The Speaker Recognition Workshop. pp. 213–218. International Speech Communication Association (ISCA), Crete, Greece (2001), http://eprints.qut.edu.au/10408/

16. Vogt, R.J., Sridharan, S.: Explicit modelling of session variability for speaker verification. Computer Speech & Language 22(1), 17–38 (January 2008), `http://eprints.qut.edu.au/12631/`

# Analysis of Complexity and Modulation Spectra parameterizations to characterize Voice Roughness

Laureano Moro-Velazquez*, Jorge Andrés Gómez García, Juan Ignacio Godino-Llorente,

*Centro de Tecnología Biomédica*, Universidad Politécnica de Madrid

**Abstract.** Disordered voices are frequently assessed by speech pathologists using acoustic perceptual evaluations. This might lead to problems due to the subjective nature of the process and due to the influence of external factors which compromise the quality of the assessment. In order to increase the reliability of the evaluations the design of new indicator parameters obtained from voice signal processing is desirable. With that in mind, this paper presents an automatic evaluation system which emulates perceptual assessments of the roughness level in human voice. Two parameterization methods are used: complexity, which has already been used successfully in previous works, and modulation spectra. For the latter, a new group of parameters has been proposed as Low Modulation Ratio (LMR), Contrast (MSW) and Homogeneity (MSH). The tested methodology also employs PCA and LDA to reduce the dimensionality of the feature space, and GMM classifiers for evaluating the ability of the proposed features on distinguishing the different roughness levels. An efficiency of 82% and a Cohen's Kappa Index of 0.73 is obtained using the modulation spectra parameters, while the complexity parameters performed 73% and 0.58 respectively. The obtained results indicate the usefulness of the proposed modulation spectra features for the automatic evaluation of voice roughness which can derive in new parameters to be useful for clinicians.

**Keywords**: GRBAS, Complexity, Modulation Spectra, Kappa Index, GMM, voice pathology, Roughness.

## 1 Introduction

Voice pathology assessment aims at diagnosing and evaluating the condition of patients with vocal pathologies, in order to find an appropriate treatment for their disorders. On this context, speech pathologist often employ perceptual analysis of patient's phonation to indicate the perceived level of perturbation of the voice. In these cases, specialists listen to the voice of the patient producing a sustained vowel or reading a particular passage and rate it conforming to a

* Corresponding e-mail address: laureano.moro@upm.es

specific procedure. Most of the times, a numeric value is assigned according to the dysfunction level, where one of the most used rating scale is $GRBAS$ [1]. This scale is divided into five traits which evaluate different speech quality characteristics: Grade ($G$), Roughness ($R$), Breathiness ($B$), Aesthenia ($A$) and Strain ($S$).

Each characteristic ranges from 0 to 3, where 0 indicates no affection, 1 slightly affected, 2 moderately affected and 3 severe affected voice regarding to the corresponding trait.

The main issue affecting the perceptual analysis of voice quality is the inherent subjectivity of the process, where external factors might compromise the quality of the assessment, such as the assessor's mood, its background training, fatigue, stress or cultural issues, among many others [2, 3].

With this in mind, acoustic analysis of voice signal techniques might be considered for reducing the uncertainty of perceptual evaluations. The acoustic analysis is widely used as a tool for monitoring the patient's evolution after the application of a treatment, mainly due to the simplicity of the process, as well as its low cost and non-invasiveness. Therefore its use in quality assessment of voice might be beneficial to clinicians, giving them tools to perform evaluations in a more objective and reproducible manner.

By using classification systems which emulate a perceptual evaluation it might be possible to identify new acoustic features or parameterizations which could be used by clinicians as a basis to perform a more objective assessment. The present paper describes an automatic $Roughness(R)$ evaluation system, based in complexity and Modulation Spectra (MS) features.

Complexity measurements have been used in several studies to determine the presence of a pathology related to the phonatory system [4–6] whereas in [7] GRBAS traits are classified using complexity among other parameterizations. Besides, MS has been used in [8, 9] to detect pathological voices. In [10, 11] it was utilized in pathology automatic classification and in [8] to obtain objective parameters to quantify voice quality. In these studies MS is revealed as a source of parameters to characterize pathological voices. On the other hand, works as [12] use acoustic parameters for automatic classification of Breathiness, obtaining a 77% of efficiency whereas [13] uses MFCC in a GRBAS classification system obtaining 65% efficiency. On [14–16], Linear Frequency Spectrum Coefficients (LFSC) are used to classify different traits in order to test the influence of frequency range in GRBAS perceptual and automatic assessments.

In this paper new MS features are proposed. Moreover, dimensionality reduction techniques and Gaussian mixture models (GMM) are employed for taking decisions on the level $(0-3)$ of $R$ trait using the proposed parameters as input.

The paper is organized as follows: Section 2 presents the theoretical background of complexity and modulation spectra features. Section 3 presents the experimental setup and describes the database used in this study. Section 4 presents the obtained results. Finally, section 5 presents the discussion, conclusions and future work.

## 2 Theoretical Background

### 2.1 Complexity measures

For extracting complexity measures, it is first necessary to represent the time series in a $m$-dimensional space, called *phase* or *state space*. In this manner, the dynamical evolution of the system, all its states and its evolution are described. The procedure usually employed is called *embedding*. Through *embedding* it is possible to calculate an *attractor* which is used to obtain the complexity measurements [17].

Some popular features are the Correlation Dimension (CD), the Largest Lyapunov Exponent (LLE) and also some Regularity measurements such as the Approximate Entropy (ApEn), the Sample Entropy (SampEn) and the Fuzzy Entropy (FuzzyEn).

***Correlation dimension*** CD is the autosimilarity of an embedded time series. It is estimated as presented in [17].

***Largest Lyapunov Exponent*** LLE is a measure of the divergence of nearby orbits in phase space, thus measuring the sensitivity to initial conditions of embedded systems. It is estimated as in [17].

***Regularity measurements*** ApEn was proposed in [18], and tries to measure the regularity of a system. Since ApEn is biased due to a phenomena called self-matching, the *Sample Entropy* is proposed in [19]. The *Fuzzy entropy* is a further improvement which changes the measurement function used in ApEn and SampEn by a Fuzzy membership function [20]. All ApEn, SampEn and FuzzyEn rely on the choosing of the tolerance parameter $r$, which is usually calculated as $r = \alpha\,\mathrm{std}(\cdot)$, where $\alpha$ is varied from within a delimited range and $\mathrm{std}(\cdot)$ is the standard deviation of the time series.

### 2.2 Modulation Spectra

On this study new MS parameters are proposed to characterize the voice signal. MS provides information about the energy of modulation frequencies that can be found in the carriers of a signal. It is a bidimensional representation where abscissa usually represents modulation frequency and ordinate axis, acoustic frequency. This kind of representation allows observing different voice features simultaneously such as the harmonic nature of the signal and the frequency modulation of fundamental frequency and harmonics. To obtain MS, signal passes through a short-Time Fourier Transform (sTFT) filter bank whose output is used to detect amplitude and envelope. This output is finally analyzed using FFT [21]. To calculate MS, Modulation Toolbox library ver 2.1 has been employed [22].

After obtaining MS it is needed to extract some parameters representative enough to be used in the classification stage. The MS is parameterized using centroids [23] (MSC) and a set of 5 new features: Low Modulation Ratio (LMR) in modulus and Contrast (MSW) and Homogeneity (MSH) in modulus and phase.

***Centroids*** MSC are obtained along the modulation frequency bands. MS is reduced to an entire number of bands usually ranging from 6 to 26. Once the reduced MS is computed, centroids are calculated and normalized taking into account the energy at the voice pitch in acoustic frequency.

***Low Modulation Ratio*** LMR is the ratio between the energy in the first modulation band at pitch frequency, $E_0$, and global energy in all modulation bands covering at least from 0 to 25 Hz at pitch frequency (acoustic band), $E_{25}$, as it is shown in equation 1

$$LMR = 10 \cdot \log(\frac{E_{25}}{E_0}) \tag{1}$$

***Contrast and Homogeneity*** Representing MS as two dimensional images it is observed that pathological voices usually seem to have more complex distributions. Images related to normal voices are frequently more homogenous and have less contrast, as can be seen in Fig. 1

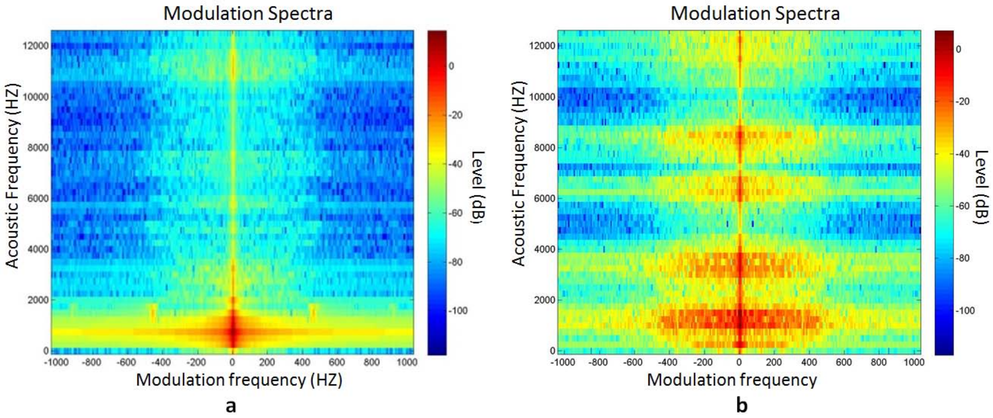


**Fig. 1.** MS modulus of a normal voice (a) and pathological voice of a patient with gastric reflux, edema of larynx and hyperfunction (b).

Homogeneity is computed using the Bhanu method described by equation 2 as stated in [24].

$$MSH = \sum_{I_m}\sum_{I_a}[f(I_m, I_a) - \overline{f}(I_m, I_a)]^2, \tag{2}$$

being $MSH$ the MS Homogeneity value, $f(I_m, I_a)$ the modulation spectra (modulus or phase) at point $(I_m, I_a)$, and $\overline{f}(I_m, I_a)$ the average value in a $3 \times 3$ window centeredred at the same point, representing $I_m$ the frequency modulation bands and $I_a$ the acoustic frequency bands.

Contrast is computed using a variation of the Weber-Fechner contrast relation method described by equation 3 as stated in [24].

$$MSW(I_m, I_a) = \sum_{I'_m}\sum_{I'_a} C_{I'_m, I'_a} \tag{3}$$

where

$$C_{I_m, I_a} = \frac{|f(I_m, I_a) - f(I'_m, I'_a)|}{|f(I_m, I_a) + f(I'_m, I'_a)|} \tag{4}$$

being $f(I_m, I_a)$ MS value (modulus or phase) at point $(I_m, I_a)$ and $(I'_m, I'_a)$ vertical and horizontal adjacent points to $(I_m, I_a)$. The global MSW is considered as the sum of all points in $MSW(I_m, I_a)$.

Regarding MSH and MSW, modulus and phase parameters are used on this study.

## 3 Experimental Setup

### 3.1 Database

The original database used for this study contained 226 recordings of maintained vowel /a:/ and the 'Rainbow passage' from the Massachusetts Eye & Ear Infirmary (MEEI) Voice Disorders Database, distributed by Kay Elemetrics [25]. Sample frequency has been adjusted to 25 kHz and quantization to 16 bits when necessary. Duration of the files used for parameterization (only sustained vowel) ranges from 1 to 3 seconds. Level of R trait has been estimated three times by two voice therapists. One of them evaluated the whole database once and the other performed the assessment in two different sessions. Only the 85 files with total agreement among the three assessments were chosen with the aim of using a database with highly consistent labels. This reduced set includes 34 male voices with age ranging from 26 to 58 years with an average of 38 and 51 female with age ranging from 22 to 52 years with an average of 35. Class distribution is shown in figure 2.

### 3.2 Methodology

The methodology employed in this paper is shown in Fig. 3, while each one of its stages is explained next. Firstly, each signal is framed and windowed using Hamming windows overlapped 50%. The window lengths are varied in the range of

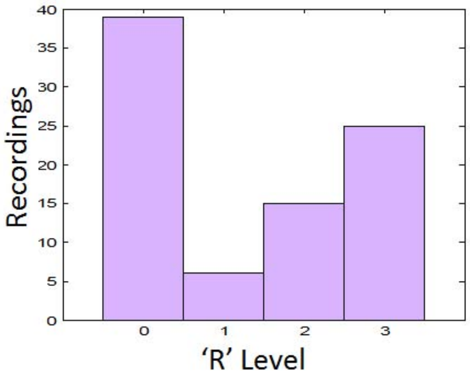


**Fig. 2.** Class histogram for trait 'R'

40-200 ms in 20 ms steps. Then, in the characterization stage, MS and complexity features are employed. The feature vector extracted from the MS amplitude is composed of the following: MSC, LMR, MSW and MSH. Additionaly, MSW and MSH are computed from the phase. The number of centroids for the MSC feature is varied in the range of $[6, 26]$ with a step size of 2. The complexity set of features is composed by CD, LLE, ApEn, SampEn, and FuzzyEn. The $\alpha$ parameters used for ApEn, SampEn, and FuzzyEn is varied in the following range: $[0.10, 0.35]$ with a step size of 0.05. Following the characterization, a 6-fold cross-validation [26] was used for evaluating results, were two experiments are defined: one without a dimensionality reduction technique, which then feeds the classification stage, and another one using various dimensionality reduction techniques previous classification. In the dimensionality reduction stage PCA [27] and LDA [28] techniques are used, varying the amount of desired features reduction from 25 to 54 %. Regarding PCA and LDA techniques, only the training data set is used to obtain the models which are employed to reshape all the data: training and test data sets. This process is repeated in every iteration of GMM training-test process carried out for validation. The reduction of dimensions is applied for both MS and complexity features separately. Finally in the classification stage, a GMM whose parameters were varied 8 to 128. The assessment of the classifier was performed by means of efficiency and Cohen's Kappa Index [29].

## 4 Results

Best results can be observed in Table 1. All tests were performed using the described reduced database with and without PCA and LDA techniques. The training set (5 folds from a total of 6) was used to train the GMM models which were validated with the remaining test fold following a 6 fold cross-validation technique.

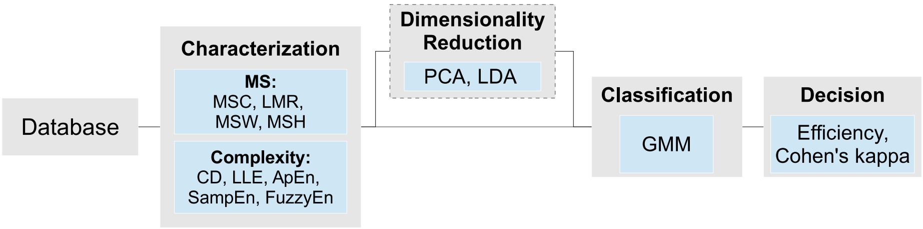


**Fig. 3.** Outline of the 'R' automatic detector presented in the paper

**Table 1.** Best Results. Efficiency ± standard deviation and Kappa Index

| Parameters | Efficiency | Kappa Index |
|---|---|---|
| Complexity | 71 ± 7 % | 0.53 |
| Complexity+PCA | 67 ± 8 % | 0.46 |
| Complexity+LDA | 73 ± 8 % | 0.58 |
| MS | 61 ± 8 % | 0.35 |
| MS+PCA | 73 ± 15 % | 0.56 |
| MS+LDA | **82 ± 7 %** | **0.73** |

Best results were obtained using MS in 180 ms frames, 8 centroids, 54 % data reduction through LDA and 14 GMM. Regarding Complexity parameters, best results are obtained with $\alpha = 0.25$, 25 % LDA dimensionality reduction and 16 GMM. All results are expressed in terms of efficiency and Cohen's Kappa Index, the latter expressing the grade of agreement between the labels assigned by the GMM classifier and the perceptual assessment done by therapists.

On Tables 2 and 3 confusion matrices are shown respectively for MS and complexity features. These matrices are the sum of the confusion matrices obtained in each of the six test folds.

**Table 2.** MS Parameters Confusion Matrix. TR are targets and PR predictions

| | PR 0 | PR 1 | PR 2 | PR 3 |
|---|---|---|---|---|
| TR 0 | **38** | 0 | 1 | 0 |
| TR 1 | 3 | **1** | 2 | 0 |
| TR 2 | 1 | 0 | **10** | 4 |
| TR 3 | 2 | 0 | 2 | **21** |

## 5 Discussion and Conclusions

On this study an analysis of two different parameterizations applied to human voice to characterize the level of *Roughness*$(R)$ has been performed. Dimension-

**Table 3.** Complexity Parameters Confusion Matrix. TR are targets and PR predictions

| | PR 0 | PR 1 | PR 2 | PR 3 |
|---|---|---|---|---|
| TR 0 | **34** | 0 | 0 | 5 |
| TR 1 | 2 | **1** | 2 | 1 |
| TR 2 | 2 | 1 | **4** | 8 |
| TR 3 | 2 | 0 | 0 | **23** |

ality reduction methods as LDA and PCA and GMM classification techniques have been used to analyze the capability of both types of parameterizations to characterize voice roughness. Best results are obtained with the proposed new MS parameters and LDA, producing 82 % of efficiency and 0.72 Cohen's Kappa Index. As it can be inferred from Altman interpretation of Cohen's index [30], shown in Table 4, agreement is considered as good. Moreover, most of errors are placed in adjacent classes as it can be deduced from confusion matrices in Tables 2 and 3.

**Table 4.** Altman interpretation of Cohen's index

| Kappa Index | Agreement |
|---|---|
| $\leq 0.20$ | Poor |
| 0.21 - 0.40 | Fair |
| 0.41 - 0.60 | Medium |
| 0.61 - 0.80 | Good |
| 0.81 - 1.00 | Excelent |

As a starting point, most of the previously exposed tests were performed with the extended database (226 files) using the three available label groups separatedly: one of them generated by one therapist and the other two created by the other therapist in two different sessions. In these cases, on spite of having a higher number of files and a more class-balanced database, results rarely outranged 62 % of efficiency. The details of these tests have not been included in this work for the sake of simplicity and conciseness. This demonstrates that consistency of the database labeling is a key point for future work. New studies should utilize only consistent labels obtained with several therapists in different sessions.

It is difficult to compare these results with other studies such as [12–16] due to, as it is stated in [31], there is not a standard database and in this particular case labeling is different for each work, although results in most of them are under 80% efficiency. The definition of a standard database with a consistent and known labeling would lead to comparable results.

As a conclusion, it might be said that results suggests that the proposed MS parameters could be used as an objective basis to help clinicians to assess

Roughness according the GRBAS scale reducing uncertainty. The use of MS seems to provide better results than complexity. It would be advisable to study the creation of a new parameter from the combination of the proposed ones, being suitable for therapists and physicians. But to obtain highly relevant and representative results, new tests with a larger database should be performed which will allow the use of a a slightly different and more consistent methodology.

## 6 Acknowledgements

The authors of this paper have developed their work under the grant of the project *TEC2012-38630-C04-01* from the Ministry of Economy and Competitiveness of Spain and *Ayudas para la realización del doctorado (RR01/2011)* from Universidad Politécnica de Madrid, Spain.

## References

1. M. Hirano, *Clinical examination of voice.* Springer Verlag, 1981.
2. I. V. Bele, "Reliability in perceptual analysis of voice quality." *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 555–73, Dec. 2005.
3. W. F. L. De Bodt, M S. *et al.*, "Test-retest study of the grbas scale: influence of experience and professional background on perceptual rating of voice quality." *Journal of voice : official journal of the Voice Foundation*, vol. 11, no. 1, pp. 74–80, 1997.
4. Y. Zhang, J. J. Jiang, L. Biazzo *et al.*, "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis." *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 519–28, Dec. 2005.
5. J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement." *Journal of voice : official journal of the Voice Foundation*, vol. 20, no. 1, pp. 2–17, Mar. 2006.
6. G.-L. J. I. S.-L. N. Arias-Londono, J. D. *et al.*, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.
7. J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón *et al.*, "Automatic GRBAS Assessment Using Complexity Measures and a Multiclass GMM-Based Detector," *Seventh International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2011.
8. M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions On Audio Speech And Language Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.
9. J. D. Arias-Londoño, J. I. Godino-Llorente *et al.*, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices." vol. 36, no. 2, pp. 60–9, Jul. 2011.
10. T. F. Q. Nicolas Malyska, "Automatic dysphonia recognition using biologically inspired amplitude-modulation features," in *Proc. ICASSP*, vol. 1. IEEE, 2005, pp. 873–876.
11. M. Markaki and Y. Stylianou, "Modulation Spectral Features for Objective Voice Quality Assessment: The Breathiness Case," *Sixth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications.*, 2009.

12. A. Stráník, R. Cmejla, and J. Vokřal, "Acoustic Parameters for Classification of Breathiness in Continuous Speech According to the GRBAS Scale." *Journal of voice : official journal of the Voice Foundation*, vol. 28, no. 5, Sep. 2014.
13. N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz *et al.*, "Automatic assessment of voice quality according to the GRBAS scale," *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 2478–2481, 2006.
14. G. Pouchoulin, C. Fredouille, J. Bonastre *et al.*, "Dysphonic Voices and the 0-3000Hz Frequency Band," *Interspeech 2008. ISCA*, pp. 2214–2217, 2008.
15. G. Pouchoulin, C. Fredouille, J.-F. Bonastre *et al.*, "Characterization of the pathological voices (dysphonia) in the frequency space," *Proceedings of International Congress of Phonetic Sciences (ICPhS)*, pp. 1993–1996, 2007.
16. G. Pouchoulin, C. Fredouille, J. Bonastre, A. Ghio *et al.*, "Frequency Study for the Characterization of the Dysphonic Voices," *Interspeech 2007. ISCA*, pp. 1198–1201, 2007.
17. H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. Cambridge University Press, 1 2004.
18. S. M. Pincus, "Approximate entropy as a measure of system complexity." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 6, pp. 2297–301, Mar. 1991.
19. J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy." *American journal of physiology. Heart and circulatory physiology*, vol. 278, no. 6, pp. H2039–49, Jun. 2000.
20. W. Chen, Zhuang *et al.*, "Measuring complexity using fuzzyen, apen, and sampen," *Medical Engineering & Physics 31 (2009) 61-68*, 2009.
21. S. Schimmel, L. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," *EEE International Conference in Acoustics, Speech and Signal Processing, 2007. ICASSP*, vol. 4, 2007.
22. L. Atlas, P. Clark, and S. Schimmel, "Modulation Toolbox Version 2.1 for MATLAB," 2010. [Online]. Available: http://isdl.ee.washington.edu/projects/modulationtoolbox/
23. B. Gajic and K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 600–608, Mar. 2006.
24. R. Peters and R. Strickland, "Image complexity metrics for automatic target recognizers," *Automatic Target Recognizer System and Technology Conference*, 1990.
25. "Voice Disorders Database," Lincoln Park, NJ., 1994.
26. B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, 1983.
27. L. Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, 2002.
28. R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, pp. 13–16, 1992.
29. J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. XX, no. 1, pp. 37–46, 1960.
30. D. G. Altman, *Practical statistics for medical research.* CRC Press, 1990.
31. N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz *et al.*, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006.

# A Grouping Method based on Energy Clustering for Reverberant Speech

Zhao Li and Thorsten Herfet

Telecommunications Lab, Saarland University, Saarbrücken, Germany
{li, herfet}@nt.uni-saarland.de

**Abstract.** The motive of this paper is to group speech units in the time frequency (T-F) domain based on a general monaural cue, whose applications include speaker localization, speech enhancement, speech separation, etc. Based on the observation that signal energy from each sound source tends to form a cluster in the T-F domain resulting from a bank of gammatone filters, in this paper we use the energy distribution as a monaural cue to group the T-F units similar to the classic watershed algorithm but under the control of cluster shape and curve fitting. Experimental results show that the proposed energy clustering has high grouping accuracy (over 93%) and excellent robustness against reverberation. It is also shown that the energy clustering improves a purely localization cue-based separation system by 15-30 percents in the term of hit minus false alarm rate.

**Keywords:** Energy clustering, Sequence grouping, Speech Separation

## 1 Introduction

Grouping speech in the time-frequency (T-F) domain and classifying these groups into separated streams have many potential applications including speaker localization, speech enhancement, speech separation, audio information retrieval, hearing prostheses systems, etc. When multiple microphones are available, the most ubiquitous approach is beamforming, which attempts to improve the signal-to-noise ratio (SNR) of a source using directional information [1]. Other approaches perform a time-frequency decomposition of the mixture signals and use interaural time difference (ITD) and interaural level difference (ILD) in each T-F unit to estimate an output signal that originates from a particular direction [2]. These systems use localization information as a primary cue to group or to classify speech units, and show rapid performance degradation as reverberation is added to the recordings.

Darwin has suggested that it is unlikely that the auditory system uses spatial information as a primary means of attending to individual sources, as individual localization cues are highly unreliable in reverberant environments [3]. He argues that much of the psychophysical evidence supports an alternative framework in which sound source localization may be a product of source separation, rather than a primary cue used to achieve it. In this account, in [4], monaural cues (e.g. harmonicity, onsets/offsets, amplitude and frequency modulation) are used first to form auditory objects (groups of T-F units). These objects are then localized to create auditory space which is used as one major cue for classifying objects into separated auditory streams.

Following the above idea, in this paper, we investigate the distribution of the signal energy of T-F units and use it as a monaural cue to group the T-F units, called energy clustering. Based on the observation that, in T-F domain resulting from a bank of gammatone filters, the signal energy from each sound source tends to form a cluster, the T-F units are grouped based on their energy level, similar to the classic watershed algorithm [5], but under the control of cluster shape and curve fitting. Energy clustering covers some of the T-F units which have unreliable or false auditory cues, hence has robustness against reverberations.

Based on our survey, there is only one related system in literature which was proposed by Kollmeier and Koch [6]. They describe a speech enhancement algorithm which works in the domain of the modulation spectrum, i.e. a two-dimensional representation of AM frequency vs. centre frequency. Energy from each sound source tends to form a cluster in the modulation spectrum, allowing sources with different modulation characteristics to be separated from one another. Our energy clustering is in the T-F domain and is used to group and separate T-F units for speech separation.

The rest of the paper is organized as follows. Section 2 shows the cue of energy distribution. The proposed grouping method is presented in Section 3. In Section 4, we evaluate the proposed grouping method and a separation system which is based on it via numerical experiments. Section 5 concludes this paper and shows the direction of future work.

## 2 Energy Distribution in the T-F Domain

As discussed above, the localization cues are unreliable in reverberant environments. T-F unit grouping, which refers to the process of joining T-F units together that are thought to primarily contain energy from a single source, can be performed in order to pool data across time and/or frequency of a larger region, allowing for more reliable classification decisions. Auditory segmentation [4, 8] takes into consideration contextual information beyond individual T-F units. However, at the segment level, the localization cue is still responsible mainly for all T-F grouping due to short time range. Monaural cues are often used to group the T-F units or to form simultaneous streams [4] which refer to a collection of T-F units over a continuous time interval that are thought to be dominated by the same source. However, some monaural cues, e.g. pitch cues, are only reliable for voiced speech [11].

The motive of this paper is to find a more general cue to group the T-F units with robustness against reverberations. We make use of the ideal binary mask (IBM) [7] as a mechanism for grouping T-F units generated by the same source. The IBM is defined in terms of premixed target and interference where, for a T-F unit, if the target energy in this T-F unit is greater than the interfering energy, it will be labelled as 1 and otherwise it will be labelled as 0.

It is found that the energy distribution of the audio mixtures in the T-F domain looks like mountains and almost each peak belongs to one sound source only, as shown in Figure 1.a. Specifically, the signal energy from each sound source tends to form peaks in both frequency and time domain (Figure 1.b and 1.c). Gaussian model is often used to model speech spectrum [12, 13]. However, due to a heavy tailed distribution, only the part of spectrum that is close to the average point satisfy Gaussian curve [14]. Super-model is a better approximation of the speech spectrum in the frequency domain [15]. Meanwhile, due to energy vibration and echoes, the distribution of speech energy in the time domain does not like single regular peaks rather than a supervision of various irregular peaks, as shown in Figure 1.c. As a result, most of the literature of speech enhancement and speech separation consider only the distribution of signal amplitude in the frequency domain (speech spectrum) [12-15].

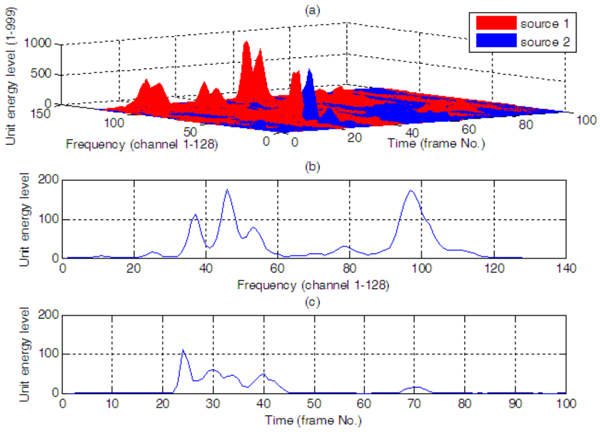


**Fig. 1.** The *distribution of signal energy in the T-F domain resulting from the gammatone filters in Section 3.1 (a. a mixture of two audio sources in the T-F domain; b. a single audio source in the frequency domain; c. a single audio source in the time domain)*

In this paper we explore the distribution of speech energy in the T-F domain and use the energy distribution as a cue to group the T-F units, called energy clustering. As observed, the energy of speech is like water and it floods from each energy spot (the peaks of signal energy). Only the fact that the signal energy tends to form clusters in the T-F domain is used in our grouping method, even though the energy distribution in the frequency or time domain is not regular curve.

## 3 Energy Clustering

In this Section, we first present the used auditory periphery and then describe the procedure of energy clustering.

### 3.1 Auditory Periphery

Human cochlear filtering can be modeled by a bank of bandpass filters. The filterbank employed here consist of 128 fourth-order gammatone filters [16]. The impulse response of the i-th filter has the following form:

$$g_i(t) = \begin{cases} t^3 exp(-2\pi b_i t) cos(2\pi f_i t + \phi_i), if\ t \geq 0 \\ 0, \quad otherwise \end{cases} \tag{1}$$

where $b_i$ is the decay rate of the impulse response, related to the bandwidth of the filter, $f_i$ is the centre frequency of the filter, and $\phi_i$ is the phase (here $\phi_i$ is set to zero). The Equivalent Rectangular Bandwidth (ERB) scale is a psychoacoustic measure of auditory filter bandwidth. The centre frequencies $f_i$ are equally distributed on the ERB scale between 80Hz and 5kHz, and specifically for each filter we set the bandwidth according to the following equations [17]:

$$ERB(f_i) = 24.7(4.37 f_i / 1000 + 1) \tag{2}$$

$$b_i = 1.019 ERB(f_i) \tag{3}$$

In order to simulate the middle-ear transfer function, the gains of the gammatone filters are adjusted according to the data provided by Moore et al. [18]. We include this middle-ear processing for the purpose of physiological plausibility. In the final step of the peripheral model, the output of each gammatone filter is half-wave rectified in order to simulate firing rates of the auditory nerve. Saturation effects are modelled by taking the square root of the rectified signal. Each bandpass filtered signal is divided into 20 ms time frames with a 10 ms overlap to create a cochleagram.

### 3.2 The Procedure of Energy Clustering

As shown in Figure 2, the idea of the watershed algorithm has been introduced in 1979 by S. Beucher and C. Lantuéjoul in [5]. It consists in placing a water source in each regional minimum, to flood the relief from sources, and build barriers when different sources are meeting. The resulting set of barriers constitutes a watershed by flooding. We consider the energy of speech as water. It floods from each energy spot (the peaks of signal energy). The procedure of the proposed energy clustering is quite like the watershed algorithm but with several control constraints which will be discussed below.

Given the T-F units generated by a bank of gammatone filters, we first assign each of them an energy level (0-999) uniformly based on their signal energy with an energy threshold which covers more than 95% or 99% of the total signal energy of all T-F units, where the scale range 1000 is chosen for a proper granularity. Please note that here different scales can be used to enlarge low energy levels where the unit energy are densely located which is left as one of the related future work. Besides the energy level, another control for the energy clustering is cluster shape control where the expected cluster shape is rectangle or ellipse. This is to avoid long and narrow tails in energy clusters.

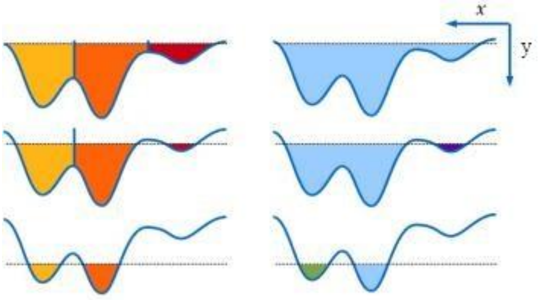


**Fig. 2.** The *classic watershed algorithm by flooding*

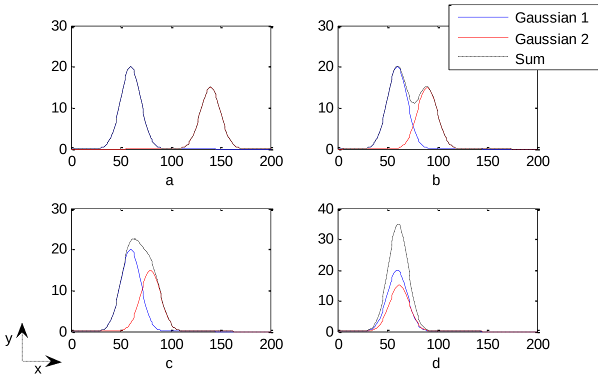


**Fig. 3.** Mixture *of two Gaussian curves*

Considering the mixture of two Gaussian curves, there are mainly four cases as shown in Figure 3. From the figure we can figure out straightforwardly that the classic watershed algorithm can catch the contour of these two peaks for the cases a and b as they do not wrap each other. For the case d where the two peaks have almost the same location, it is difficult to distinguish the original two peaks. Luckily, experiment results show that this case is quite rare (less than 1% for the mixture of two speeches). Here we just ignore it and leave it for the further separation with other auditory cues, e.g. ITD, ILD, pitch etc. For the case c, Gaussian curve fitting cannot be used to recover the original two peaks because that the curves of the signal energy in the frequency domain are not exactly Gaussian function as discussed in the previous section. Here, we approximate the separating boundary to be the point located at the distance of one third of the half peak range from the peak centre in the wider side. Experiments show that the deviation (error) of the proposed curve fitting method is at maximum three channels (overall 128 channels), which is acceptable. Moreover, this deviation can be reduced by the control methods based on auditory cues and cluster shape. This method is referred to as curve fitting control in the rest of this paper.

Another issue is to identify the above cases a-d. Given a peak we first check whether it is symmetric. When the deviation of symmetry –the width difference between the left and right sides - is over a threshold (e.g. 3 channels in our experiments with 128 channels in all), the peak is considered as case c. Otherwise, the peak is considered as from a single source and belongs to case a, b or d, where the classic watershed algorithm is used to catch the contour of peaks as discussed above.

The procedure of the energy clustering is as follows:

[a] Consider the residual (e.g. not yet grouped) T-F unit with the highest energy level, compare its energy level with the ones of its four neighbours and determine:

[a.i] group this unit with one of its neighbour if this neighbour has the highest energy level among all neighbours, this energy level of this neighbour is higher than the one of the current unit and the current unit satisfy the constraint of the curve fitting control and shape control.

[a.ii] assign a new group number to it, otherwise.

[b] Repeat step [a] until all the units with nonzero energy level have been handled.

In this algorithm, we use only the energy cue of T-F units, which is a more general cue than other auditory cues like ITD, ILD, pitch, etc. Please also note that in the process of energy clustering, the localization cues and pitch cues in the average manner can be also used for the control of the energy clustering process, which will be shown in the next section.

## 4 Performance evaluation

We evaluate the performance of the proposed energy clustering-based grouping and a separation system based on it by using a corpus which contains 20 utterances (each lasts 3s) spoken by three men and two women. All utterances are down sampled from 44.1kHz to 20kHz. The test set consists of 30 utterances (each last 3s) mixed by two speakers randomly at -5, 0 and 5dB. Please note that the test cases of -5 and 5dB are in the same mixtures where the stronger one is noted as 5dB while the other one is noted as -5dB. The two sources are located at azimuth ±30 ° respectively. The software Roomsim [10] is used to simulate the convolutive mixture in different reverberant rooms: anechoic, percent50 and acoustic tile suspend (ATS). And the original sources are used to compute the unit accuracy of the separated speeches.

We first evaluate the performance of the proposed grouping method and then show how well it can improve the speech separation system which is based on location cues only. We only consider the audio mixtures of two sources for simplicity. Separation of the audio mixtures of more than two sources will be evaluated in the future work.

### 4.1 The Performance of Energy Clustering

In this section, we evaluate the performance of energy clustering method based on the energy cue of T-F units only. We first show the distribution of group size got from the energy clustering method for the same mixture of two speeches (1s long) in different reverberant rooms, shown in Figure 4. The results show that about 150-200 energy clusters are generated by the proposed energy clustering for a one second long mixture of two speeches. We can see that the maximum group size goes up to 280 units and the average group size is about 50 units which is much larger than the segment size that is around 9 units normally [8]. We also can see that the average group size is almost the same in different reverberant rooms. This result confirms that the proposed energy clustering has excellent robustness against reverberation.

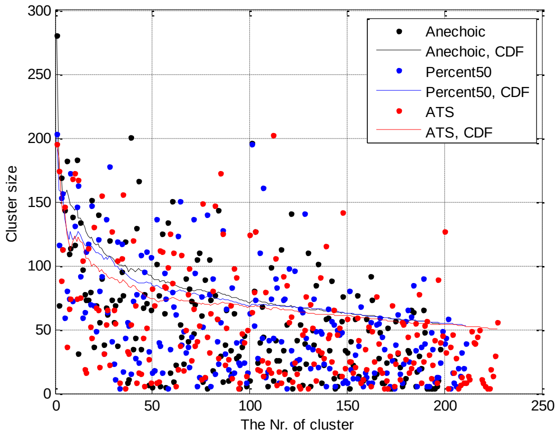


**Fig. 4.** The *distribution of group size*

We evaluate the performance of grouping further based on the grouping accuracy, which is defined as the unit assignment accuracy when all the groups are correctly classified. Table 1 shows the results. We can see that the grouping accuracy is as high as 93% and 96% for energy saved rate 99% and 95% respectively. This means that high unit accuracy can be obtained by speech separation systems based on the energy groups, which is confirmed in the following section. We can also see that the reverberation has almost no influence on the grouping accuracy, which confirms that the proposed energy clustering-based grouping has excellent robustness against reverberation. For the evaluation of speech separation in the next section, we only consider the case of energy saved rate 99% which is a very high rate of energy recovery.

**Table 1.** Grouping *accuracy*

| Energy saved | Anechoic | Percent50 | ATS |
|---|---|---|---|
| **95%** | 0.9656 | 0.9661 | 0.9633 |
| **99%** | 0.9320 | 0.9343 | 0.9355 |

### 4.2 The Performance of the Separation System

The grouping method based on energy clustering can be used to improve the performance of speech separation. A separation system is evaluated where the groups got from energy clustering are classified into separated speech based on the average location cue of each group which is computed with the Skeleton method [2] and precedence effect weighting [9] [19], shown as follows.

$$A(\phi) = \sum_i \sum_j w_{i,j} S(i,j,\phi) \quad if\ w_{i,j} > Thres_{PE} \tag{4}$$

$$S(i,j,\phi) = Q(i,j,\phi) * \exp\left(-\phi^2 / 2\sigma_i^2\right) \tag{5}$$

where $w_{i,j}$ is a real-valued weight which measures the energy ratio between unit $u_{i,j}$ and $u_{i,j-1}$ and has a specified threshold ($Thres_{PE}$ e.g. 0.8). A *Skeleton* function $S(i,j,\phi)$ is formed by superimposing Gaussian functions at azimuths corresponding to local maxima (denoted as $Q(i,j,\phi)$ ) in the corresponding cross-correlation function, where $i, j$ and $\phi$ are channel index, time (frame) index and azimuth in degree respectively.

Moreover, the localization cues (ITD and ILD) and autocorrelation (ACC) cues are also used for the further control of the energy clustering process. To avoid the unreliable cue of a single unit, we perform the control of auditory cues in the term of average cue for a group or a serial of units. The basic idea is that when the average cue of a serial of units is closer to other source than to the source of the current group, this serial of units are cut off from the current group and are assigned to a new group number. The separation system is denoted as “Energy Clustering-SS”.

We also compare the performance of our separation system with a purely localization cue-based approach. This is an improved purely localization cue-based system, noted as PureLoc-IMP, which computes the IBM using segmentation, ITD/ILD, precedence effect-based methods, onset/offset, etc. In fact, the proposed separation system above is just PureLoc-IMP plus energy clustering-based grouping. This is to evaluate how the proposed energy clustering can improve a purely localization cue-based speech separation system.

To measure the separation performance, we compute the HIT rate, which is the percentage of the target-dominant units in the IBM correctly classified, and the FA rate, which is the percentage of the interference-dominant units in the IBM wrongly classified. It has been shown that HIT–FA is highly correlated to human speech intelligibility [11]. We also compute the unit classification accuracy, which is the percentage of correctly classified units.

We first compare the unit classification accuracy of these two methods: PureLoc-IMP and Energy Clustering-SS, shown in Table 2. With enhancements, we can see that the proposed Energy Clustering-SS achieves 10-20% improvements over the PureLoc-IMP method. Moreover, the unit accuracy of the PureLoc-IMP method drops heavily in reverberant environments. However, the unit accuracy of the Energy Clustering-SS merely drops in reverberant environments. This performance confirms again that the proposed energy clustering method has the capability to handle reverberation.

We compare these two methods further in the metrics of HIT, FA and HIT-FA, as shown in Table 2. The Energy Clustering-SS achieves a high HIT-FA rate (74-92%) which is about 15-30% improvements over the PureLoc-IMP. Please also note that the HIT-FA of the PureLoc-IMP method degrades rapidly in reverberant environments. Meanwhile, as expected, the HIT-FA of the Energy Clustering-SS merely drops in reverberant environments. Moreover, based on our survey, the HIT-FA performance (74-92%) of the proposed system is quite high even compared roughly with the results obtained by other systems from literature, e.g. 30-70% in [11].

**Table 2.** T-F *unit accuracy (the first four rows are for Energy Clustering-SS while the rest four rows are for PureLoc-IMP; H-F=HIT-FA, Acc.=accuracy)*

| | **anechoic** | | | **Percent50** | | | **ATS** | | |
|---|---|---|---|---|---|---|---|---|---|
| | -5dB | 0dB | 5dB | -5dB | 0dB | 5dB | -5dB | 0dB | 5dB |
| **HIT** | 0.896 | 0.930 | 0.963 | 0.907 | 0.914 | 0.935 | 0.920 | 0.900 | 0.895 |
| **FA** | 0.090 | 0.073 | 0.042 | 0.142 | 0.088 | 0.041 | 0.178 | 0.103 | 0.045 |
| **H-F** | **0.806** | **0.857** | **0.921** | **0.765** | **0.826** | **0.894** | **0.742** | **0.797** | **0.850** |
| **Acc.** | 0.945 | 0.932 | 0.945 | 0.928 | 0.918 | 0.928 | 0.907 | 0.902 | 0.907 |
| **HIT** | 0.890 | 0.860 | 0.857 | 0.823 | 0.783 | 0.765 | 0.883 | 0.746 | 0.628 |
| **FA** | 0.247 | 0.142 | 0.062 | 0.339 | 0.195 | 0.116 | 0.418 | 0.206 | 0.072 |
| **H-F** | **0.643** | **0.718** | **0.795** | **0.484** | **0.588** | **0.649** | **0.465** | **0.540** | **0.556** |
| **Acc.** | 0.863 | 0.851 | 0.863 | 0.762 | 0.765 | 0.762 | 0.716 | 0.733 | 0.716 |

## 5 Conclusions

In this work, we use the energy distribution as a monaural cue to group the T-F units in the T-F domain similar to the watershed algorithm but under the control of curve fitting and cluster shape constraints. Experimental results show that the proposed energy clustering has high grouping accuracy (over 93%) and excellent performance against reverberation. A separation system is applied which classifies the energy groups into separated auditory streams based on the average localization cue of each group computed using Skeleton and precedence effect weighting. It is shown that the energy clustering improves the purely localization cue-based separation system by 15-30 percents in the term of HIT-FA.

The future work includes refining the grouping method based on energy clustering, evaluating the performance for the mixtures of more than two sources, combining the energy clustering with monaural cues and developing a monaural speech separation system.

## 6 References

1. M. Brandstein and D. Ward, Eds., "Microphone Arrays: Signal Processing Techniques and Applications", Springer, 2001.
2. Roman, N., D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," J. Acoust. Soc. Amer., vol. 114, pp. 2236-2252, 2003.
3. Darwin, C. J., "Spatial Hearing and Perceiving Sources," in Auditory perception of sound sources, Eds Yost, W. A., Fay R. R. and Popper, A. N., Springer-Verlag, 2008.
4. J. Woodruff and D.L. Wang, "On the role of localization cues in binaural segregation of reverberant speech", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009, pp. 2205-2208.
5. Serge Beucher and Christian Lantuéj, "Use of Watersheds in Contour Detection", International Workshop on image processing: Real-time edge and motion detection, 1979.
6. B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," in J. Acoust. Soc. Am., 1994, vol. 95, pp. 1593–1602.
7. D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in Speech separation by humans and machines, P. Divenyi, Ed., pp. 181–197, 2005.
8. S. Schulz and T. Herfet, "Binaural Source Separation in non-ideal Reverberant Environments", Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France, September 10-15, 2007.
9. Li, Z.; Herfet, Th.; Grochulla, M.; Thormählen, T.: "Multiple Active Speaker Localization based on Audio-visual Fusion in two Stages", 2012 IEEE International Conference on Multisensor Fusion and Information Integration (MFI 2012), Germany, 2012.
10. Campbell, D.R., The ROOMSIM, http://media.paisley.ac.uk/~campbell/Roomsim/.
11. K. Han and D.L.Wang., "A classification based approach to speech segregation", Journal of the Acoustical Society of America, vol. 132, pp. 3475-3483, 2012.
12. P. J. Wolfe & S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," Spectral Issue EURASIP JASP Digital Audio Multim. Commun., No.10, 2003, pp. 1043-1051.
13. I. Cohen & B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal Process., Vol. 81, No. 11, 2001, pp. 2403-2418.
14. J. Porter & S. Boll, "Optimal estimators for spectral restoration of noisy speech," in Proc. ICASSP'84, 1984, pp. 18A.2.1-18A.2.4.
15. T. Lotter & P. Vary, "Noise reduction by maximum a posteriori spectral amplitude estimation with supper-gaussian speech modeling," Proc. 8th Internat. Workshop on Acoustic Echo and Noise Control, 2003, pp. 83-86.
16. Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P.. "An efficient auditory filterbank based on the gammatone function," Applied Psychology Unit (APU), Report 2341 (Cambridge, UK), 1988.
17. Glasberg, B. R., and Moore, B. C. J., "Derivation of auditory filter shapes from notched-noise data", Hear. Res. 44, 1990.
18. Moore, B. C. J., Glasberg, B. R., and Baer, T. "A model for prediction of thresholds, loudness, and partial loudness," J. Audio Eng. Soc. 45, 224–240, 1997.
19. J. Woodruff and D. L. Wang, "Integrating monaural and binaural analysis for localizing multiple reverberant sound sources," in Proc. ICASSP, Mar. 2010, pp. 2706-2709.

# Residual VQ-quantization for speech frame loss concealment

Domingo López-Oller, Angel M. Gomez, José L. Pérez-Córdoba *

Departamento de Teoría de la Señal, Telemática y Comunicaciones
University of Granada, Spain
{domingolopez,amgg,jlpc}@ugr.es

**Abstract.** In this paper, we propose a mitigation technique based on an Minimum Mean-Square Error (MMSE) estimation from the last correctly received frames. This technique provides replacements to reconstruct not only the LPC coefficients but also the residual signal.A critical requirement in the proposed estimator is the quantization of the speech parameters, both LPC and excitation signal. Hence, in this paper we present several VQ methods based on the well known Linde-Buzo-Gray (LBG) algorithm extended with different modifications. The performance of our proposals are evaluated over the AMR codec in terms of speech quality using the PESQ algorithm. This novel mitigation technique achieves a noticeable improvement over the legacy codec under adverse channel conditions with no increase of bitrate, with a low computational cost and without any delay in the decoding process.

**Keywords:** Speech reconstruction, frame erasure, packet loss concealment, code-excited linear prediction (CELP), AMR, LPC-residual concealment, MMSE

## 1 Introduction

In recent years, the widespread deployment of packet-based networks has facilitated the development of new services and applications which suppose a lower cost and accessible alternative to the traditional telephony (i.e. VoIP service). However, those packet-based networks were originally designed for non-real time data communications suffering from network congestion and packet delays. Due to this, the speech transmissions under this kind of networks often leads to high rates of packet losses and/or consecutive packet losses (bursts) that must be concealed by the speech codecs. Moreover, the problems of packet losses are aggravated by the fact that modern speech codecs are based on the Code Excited Linear Prediction (CELP) paradigm [1]. This paradigm provides a high quality synthesis at a remarkably low bitrate which makes it suitable for bandwidth limited channels. One of the most widely used codecs for speech coding is the

* This work has been supported by an FPI grant from the Spanish Ministry of Education and by the MICINN TEC2013-46690-P project.

Adaptive Multi-Rate (AMR) [2] as it operates at different bitrates. However, CELP-based codecs are more vulnerable to frame erasures due to the extensive use of the predictive filters, particularly the long-term prediction filter (LTP), responsible of the error propagation [3, 4]. This is obviously a major drawback of these codecs when they operate over packet-switched networks as a single lost frame can degrade the quality of many subsequent frames albeit these were correctly received. Nevertheless, there are already proposed a number of techniques to minimize or even avoid the propagation error [5–7] but in this paper we are interested in the concealment of loss itself.

In a previous work [8], we applied an error mitigation technique to the iLBC [9] codec in order to restore lost frames in a transmission. In that work, we got the concealment for the LPC coefficients and also for the residual signal from a replacement super vector obtained with a source-based model of the speech. Under that approach, a critical aspect was the vector quantization (VQ) of the residual signal (excitation). In this paper, we propose several modifications to this VQ method, originally based on the well known Linde-Buzo-Gray (LBG) algorithm [10], in order to further improve the achieved results. In this paper, the proposals are evaluated with the AMR codec instead of iLBC codec in order to test our approach with a codec with error propagation.

The remainder of this paper is organized as follows. In Section 2, we describe our mitigation technique. In Section 3, we explain how to obtain the replacement super vectors for each parameter (LPC coefficients and the residual signal) and the VQ process involved. In Section 4 we describe the experimental framework and the achieved results with AMR codec. Finally the conclusions are summarized in Section 5.

## 2 Source-based mitigation of lost frames

When a packet loss occurs, most of the Packet Loss Concealment (PLC) algorithms try to minimize the degradation on the perceptual quality by extrapolating and gradually muting (in the case of consecutive lost frames) the speech signal. However, the information contained in the source can be exploited to get better estimations of the lost information. The problem is how to model this information in an efficient way. In [8], we only considered the last correctly received frame to estimate the lost information by means of a Minimum Mean-Square Error (MMSE) approach. In this paper we will consider also when there are more than one correctly received frame ($\mathcal{O}$ frames) in order to obtain better estimates.

### 2.1 First order data-source model

In the case of the 1st-order data-source model, we will only consider the last correctly received frame. Thus, assuming that the speech frame parameters (i.e. LPC coefficients and residuals) are quantized we can compute the conditional probability of a certain symbol $j$ (vector quantization center $\mathbf{c}^j \in \mathcal{C}$) at instant

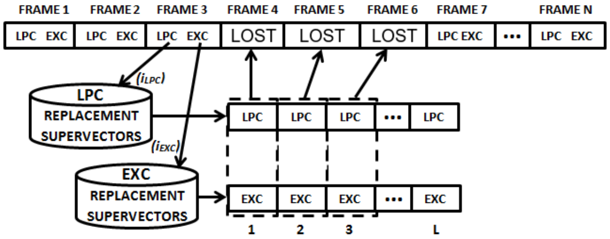


**Fig. 1.** Diagram of the replacement scheme for the LPC and residuals (EXC) parameters in a burst.

$t+l$ granted that symbol $i$ has been previously received at instant $t$ ($P(i_{(t+l)} = j|i_t = i)$). In such a way, we can compute an MMSE estimation of the lost parameters as [11]:

$$\hat{\mathbf{c}}_l(i) = \sum_{j=0}^{C-1} \mathbf{c}^{(j)} P(i_{t+l} = j|i_t = i), \qquad (1 \leq l \leq L), \tag{1}$$

where $\hat{\mathbf{c}}_l(i)$ is the replacement computed for the $l$-th frame in the burst, of maximum size $L$, provided index $(i)$ was the last symbol received before it and $C$ is the number of vectors in codebook $\mathcal{C}$.

As we can see in Figure 1, when we have bursts of consecutive losses, the indexes which represent the LPC filter and the residual signal ($i_{LPC}$ and $i_{EXC}$) quantization are extracted from the last correctly received frame and then the corresponding super vectors ($V_{LPC}$ and $V_{EXC}$) are selected. These super vectors will provide an approximation for each loss in the burst up to a maximum length $L$. Each replacement super vector is defined as $V(i) = (\hat{\mathbf{c}_1}(i), \hat{\mathbf{c}_2}(i), ..., \hat{\mathbf{c}_L}(i))$ which can be pre-computed from a training database causing no computational burden at the decoding stage.

### 2.2 Data-source model of second and further orders

The previous scheme can be extended to an estimation based on the two or more nearest correct frames. Thus, the estimations can be obtained in a similar way as (1) but extending the symbol transition probabilities to $P(i_{t+l} = k|i_t = i_1, ..., i_{t-\mathcal{O}} = i_{\mathcal{O}})$, where $\mathcal{O}$ is the desired order of data-source model.

We could obtain the super vectors in the same way as we did before. However, it must be noted that the procedure to obtain $P(i_{t+l} = k|i_t = i_1, ..., i_{t-\mathcal{O}} = i_{\mathcal{O}})$ present two important issues. On the one hand, a data-source model of arbitrary order $\mathcal{O}$ can not be affordable due to the amount of required memory, which increases exponentially along with the order ($C^{\mathcal{O}}$). On the other hand, a certain combination of $\mathcal{O}$ indexes could appear less than a certain number $\mu$ of times in the training database leading to bad estimates.

In order to solve both problems, it is possible to reduce the memory requirements by keeping those combinations with length $\mathcal{O}$ which appear more than

$\mu$ times in the database. Thus, the required memory will not only decrease but also we avoid possible bad estimates. Nevertheless, this solution does not ensure a replacement vector for each combination of $\mathcal{O}$ indexes so that it is necessary to use lower order estimates in order to provide a replacement for non-trained combinations. Further details of this approach can be found on [11].

In any case, the success of the resultant replacement super vectors will depend on the quantized parameters (LPC coefficients and the residual signal) so in the next section we will focus on their representation and codebook generation.

## 3 Frame reconstruction using LPC and residual signal estimations

The success of our error mitigation technique depends on how good is our codebook to provide efficient quantized signals. In the next subsections we describe the representation and codebook generation for LPC coefficients and residual signal.

### 3.1 LPC representation and codebook generation

In this paper, we focus on the residual representation since little or nothing has been done for this signal in the bibliography. Thus, we keep the same representation and codebook generation for the LPC as we did in [8]. Hence, we have extracted the LPC coefficients for each frame and we consider its representation as Line Spectral Frequencies (LSF). The LSF coefficients are more practical for coding and estimation because they exhibit the properties of ordering and distortion independence [12]

In order to calculate the conditional probability of (1) and further orders ($\mathcal{O}-$ *order*) over the LPC coefficients, an LSF codebook is necessary. This codebook is obtained by applying the LBG algorithm over the training database. Then, estimates $\hat{\mathbf{c}}_l(i_1, i_2, ..., i_{\mathcal{O}})$ can be easily obtained by averaging the LSF coefficients found at time instant $t+l$ for each time in which the combination $(i_1, i_2, ..., i_{\mathcal{O}})$ is observed at instant $t$ until $t-\mathcal{O}$ over the entire speech database. In such a way, the conditional probability of (1) is not necessary (but implicitly used) and quantization errors can be alleviated. Once this is completed, LSF super vectors are reverted to LPC in order to provide replacements for these parameters at positions from $l=1$ to $L$.

### 3.2 Residual signal representation and codebook generation

The representation of the residual signal is achieved in different ways by the different speech codecs. However, in this work we are interested in obtain a good approximation of the entire residual signal $\hat{e}_c(n)$, using an unique codebook, that minimizes the synthesis error $\epsilon$ with respect to the original speech signal. This synthesis error is commonly defined as:

$$\epsilon = \sum_{n=0}^{N-1} \left(h(n) * \hat{e}_c(n) - s(n)\right)^2 , \tag{2}$$

where $N$ is the length of the frame, $h(n)$ is the LP impulse response and $s(n)$ is the target signal obtained from the original signal $s(n)$ after removing the zero-input component.

In this paper, the residual signal approximation ($\hat{e}_c(n)$) will be given by a codebook obtained with a VQ method. Nevertheless, although LBG is a well known method for vector quantization, this method presents several problems to obtain suitable codebooks for the residual signal:

- The selected centroid from a cell $\mathcal{B}$ could lose important information such as the pitch due to the averaging applied during the computation of the centroid.
- Two signals with similar shape but different amplitude could be assigned to different centroids due to the euclidean distance. The residual signals assigned to a cell $\mathcal{B}$ could be significantly different across themselves and we can obtain a bad representative centroid by averaging.

For these reasons, the LBG algorithm is modified in order to determine which signals are assigned to each cell and how to choose the corresponding centroid. In the next subsections we will describe these modifications.

**LBG algorithm with kmedoids center criteria** In order to solve both problems, firstly we propose a new method for splitting the training database into clusters by considering the shape of the signal only and not its amplitude. Hence, prior to the computation of the euclidean distance between the residual signal and the center in each cell $\mathcal{B}$, these signal are rescaled into the $[-1, 1]$ range from their minimum and maximum amplitude (respectively). In such a way the residual signal is assigned to the cell whose components show a similar normalized waveform.

Secondly, we consider the kmedoids algorithm [13] such that the cell representative (medoid) will be one of the components of the cell $\mathcal{B}$. This medoid is selected as the closest observed residual to the mean of the cell. Some implementations differ on the proposed distance function but in this case, the euclidean distance is selected. We will refer to the proposal whose replacement vectors ($V_{EXC}$) are obtained through this method as LBG with kmedoids (LBGK).

**Synthesis LBG algorithm.** Although the previous algorithm will provide better codebooks than the standard LBG, we must take into consideration that our objective is find that approximation $\hat{e}(n)$, which minimize the synthesis error (2). As it is well known, the euclidean distance is not enough to ensure that the selected centroid will minimizes the error after the LP filtering ($h(n)$) [8]. Hence, the centroid will be calculated as the one which minimizes the synthesis error with all the elements of the cell $\mathcal{B}$. In this case, we need to modify the kmeans optimal cell and center criteria in order to obtain the cells and their centroids in accordance with (2).

During the clustering process of the LBG algorithm, we will consider a synthesis distance defined from the synthesis error in (2) instead of the euclidean one. Thus, given a residual $\mathbf{e}_b$ corresponding to the $b$-th speech frame in the training database, this is assigned to a centroid $\mathbf{c}^{(i)}$ iff $\epsilon\left(b, \mathbf{c}^{(i)}\right) < \epsilon\left(b, \mathbf{c}^{(j)}\right) \forall i \neq j$, being $\epsilon(b, \mathbf{c})$ defined as,

$$\epsilon(b, \mathbf{c}) = \sum_{n=0}^{N-1} (h_b(n) * e_b(n) - h_b(n) * c(n))^2, \tag{3}$$

where $h_b(n)$ and $e_b(n)$ are the impulse response and the residual of the $b$-frame respectively.

Once the $C$ regions are filled, it is necessary to find the optimal centroid. Thus, given a set of frames, $\mathcal{B}_i$, all of them corresponding to the cell $i$, its optimal center is obtained as,

$$\mathbf{c}_{new}^{(i)} = \underset{\mathbf{c}}{\operatorname{argmin}} \left( \sum_{b \in \mathcal{B}_i} \epsilon(b, \mathbf{c}) \right) \tag{4}$$

Nevertheless, obtaining the optimum center through the previous equation can be troublesome due to the convolution operation in (3). To solve this, we can consider performing the minimization in the spectral domain by applying the DFT transform to the involved signals. In order to linearize the implicit circular convolution, zero-padding is applied to signals $h_b$ and $e_b$, which are extended to $K = 2N - 1$ samples (being $K$ the size of the DFT). Thus, we obtain the following synthesis distance,

$$\varepsilon(b, \mathbf{C}) = \sum_{k=0}^{K-1} (H_b(k)E_b(k) - H_b(k)C(k))^2 \tag{5}$$

where $\mathbf{C} = (C(0), ..., C(K-1))$ is the DFT of the zero-padded extension of $\mathbf{c}$. It must be noted that both distances, $\varepsilon$ and $\epsilon$, are not identical and lead to different meanings for the residual optimization (see explanation in [8]). Hence, by using distance $\varepsilon$, we can rewrite (4) as follows,

$$\mathbf{C}_{new}^{(i)} = \underset{\mathbf{C}}{\operatorname{argmin}} \sum_{k=0}^{K-1} \sum_{b \in \mathcal{B}_i} (H_b(k)E_b(k) - H_b(k)C(k))^2 \tag{6}$$

where each DFT bin can be independently optimized as quadratic distances always return positive values. In this way, each DFT component of the centroid $\mathbf{C}_{new}^{(i)}$ is given by:

$$C_{new}^{(i)}(k) = \underset{C}{\operatorname{argmin}} \sum_{b \in \mathcal{B}_i} (H_b(k)E_b(k) - H_b(k)C)^2 \qquad 0 < k < K-1 \tag{7}$$

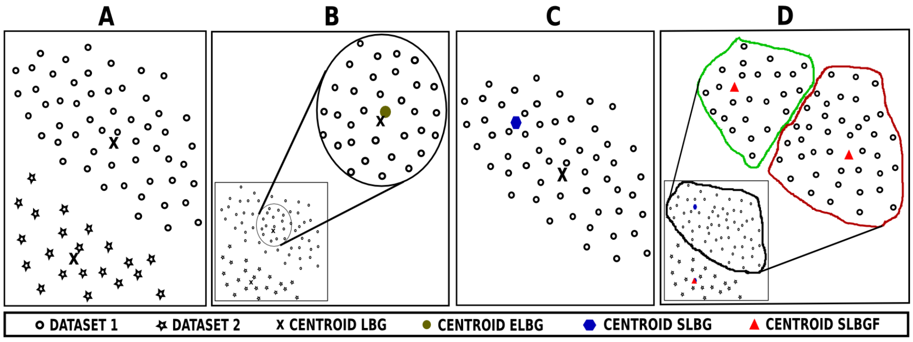


**Fig. 2.** Different methods to obtain the centroid for each algorithm. a)LBG algorithm b)LBGK algorithm c)SLBG algorithm d)SLBGF algorithm

By means of a least square error procedure we can compute each component as:

$$C^{(i)}_{new}(k) = \frac{\sum_{b \in \mathcal{B}_i} H^*_b(k) H_b(k) E_b(k)}{\sum_{b \in \mathcal{B}_i} H^*_b(k) H_b(k)}. \tag{8}$$

The centroid $\mathbf{c}^{(i)}_{new}$ can be retrieved as the IDFT transform of $\mathbf{C}^{(i)}_{new}$. We will refer as Synthesis LBG (SLBG) to this proposal.

**Synthesis LBG algorithms with Fixed or Dynamic splitting.** Although the previous method guarantee that the final residual signal $\mathbf{c}^{(i)}_{new}$ represents the cell $\mathcal{B}$ in terms of minimum distance of (2), this codebook can be improved. This improvement is based on the fact that the number of elements assigned to each cluster $\mathcal{B}$, in each iteration, is not the same. Thus, the most populated cell can be split in order to obtain a better training.

In this way, we split the cell $\mathcal{B}_i$, iff $||\mathcal{B}_i||_0 > ||\mathcal{B}_j||_0, \forall i \neq j \in \mathcal{C}$, and then we propose two training databases:

- Split $\mathcal{B}_i$ into two new cells keeping the others cells without changes. We will call Synthesis LBG with Fixed splitting (SLBGF) to this proposal.
- Split the cell $\mathcal{B}_i$ in two new cells but applying again convergence over all the other cells with new partitioning. We will call Synthesis LBG with Dynamic splitting (SLBGD) to this proposal.

Before a cell splitting, we considered a parameter $\alpha$ in order to decide when a cell $\mathcal{B}$ should be split or not. The value of $\alpha$ will be critical because if we want to obtain good estimates in the MMSE estimation, it is necessary to have a minimum number of elements ($\alpha$) in each cell of the training database.

The figure 2 shows an example of the partitioning and the resultant centers of the different proposals. As we can see, each method will select a centroid for each cluster and will determinate the codebooks for the residual signal. Once a VQ codebook is trained, the replacement super vectors $V_{EXC}$ are computed for each of its centers. In contrast to LSF vectors, residuals found after every instance of a given center $i$ are not averaged. Instead of that, all these residuals are grouped into a new set $\hat{\mathcal{M}}$ and then we apply the corresponding algorithm to find the center (kmeans, kmedoids or the synthesis distance algorithm).

## 4 Experimental results

We have considered an objective speech-quality test applied over the speech codec AMR (mode 12.2 Kbps) in order to evaluate the performance of the different proposals. The objective quality test performed is the ITU Perceptual Evaluation of Speech Quality standard (PESQ) [14].

The frame erasures are simulated by a Gilbert channel, commonly used in the evaluation of PLC techniques over loss-prone channels, with average burst lengths (ABL) from 1 to 12 frames and with packet loss ratios (PLR) from 10% to 50% [8]. Although some channels conditions are known to be non realistic loss scenarios, in this paper we are interested in testing conditions with long bursts. Thus, in order to provide good statistics with a number of long burst, high packet loss ratios must be assumed.

For the PESQ test we have considered a subset of the TIMIT database [15], downsampled at 8 kHz and composed of a total of 1328 sentences (928 for training and 450 for test) uttered by a balanced number of male and female speakers. Since PESQ test recommends utterance lengths between 8 and 20 seconds, those sentences uttered by a same speaker were joined to obtain longer utterances of approximately 7 s. The scores obtained for every test sentence are weighted by their relative length in the overall score.

The replacement super vectors obtained with the different proposals will allow us to regenerate the lost frames. Nevertheless, these vectors will imply considerable memory requirements. The LPC replacement super vector is a matrix which size is $C \times L \times p$, where $p$ is the number of LPC coefficients (typically 10) and $C$ is the number of combinations of $\mathcal{O}$ indexes for the data-source model. Those indexes are provided by codebooks of 1024 centers. The residual replacement super vector is also a matrix which size is $C \times L \times N$, where $N$ is the number of samples of the residual signal. We have considered $L = 20$ which seem enough for most of bursts. If the burst length is longer than $L$, the last replacement in the selected super vector will be repeated. As can be noted, it is a considerable amount of required memory but it is still affordable for currently available devices.

As shown in the previous section, the different tested proposals differ in the VQ quantization method (LBG, LBGK, SLBG, SLBGF, SLBGD) for the residual signal. In the case of SLBGF and SLBGD we have consider $\alpha = 100$ as the minimum number of elements in a cell required to split it. The results of our proposals are presented in Table 1. This table also includes the results obtained by the AMR codec, with its own PLC algorithm, as a reference (Baseline). As can be observed, better results are obtained as ABL increases, since our technique performs better for long bursts.

| **(PLR - ABL)** | **10-1** | **10-4** | **10-12** | **20-1** | **20-4** | **20-12** | **40-1** | **40-4** | **40-12** |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 2.819 | 2.820 | 2.925 | 2.273 | 2.223 | 2.273 | 1.348 | 1.446 | 1.450 |
| 1st-order MMSE | | | | | | | | | |
| LBG | 2.836 | 2.865 | 2.924 | 2.295 | 2.329 | 2.369 | 1.513 | 1.653 | 1.729 |
| LBGK | 2.871 | 2.917 | 2.933 | 2.366 | 2.415 | 2.477 | 1.575 | 1.716 | 1.835 |
| SLBG | 2.945 | 2.974 | 2.982 | 2.438 | 2.474 | 2.553 | 1.686 | 1.792 | 1.897 |
| SLBGF | 2.987 | 3.014 | 3.052 | 2.513 | 2.538 | 2.600 | 1.767 | 1.879 | 2.033 |
| SLBGD | 3.014 | 3.037 | 3.086 | 2.530 | 2.554 | 2.638 | 1.772 | 1.899 | 2.064 |
| 2nd-order MMSE | | | | | | | | | |
| LBG | 2.852 | 2.926 | 3.014 | 2.369 | 2.405 | 2.466 | 1.553 | 1.705 | 1.798 |
| LBGK | 2.902 | 2.994 | 3.074 | 2.423 | 2.458 | 2.527 | 1.622 | 1.790 | 1.929 |
| SLBG | 2.996 | 3.044 | 3.107 | 2.495 | 2.566 | 2.638 | 1.717 | 1.882 | 2.018 |
| SLBGF | 3.031 | 3.078 | 3.139 | 2.595 | 2.595 | 2.654 | 1.764 | 1.913 | 2.068 |
| SLBGD | 3.068 | 3.106 | 3.153 | 2.558 | 2.617 | 2.676 | 1.799 | 1.941 | 2.096 |

**Table 1.** Average PESQ scores obtained with legacy codec AMR (Baseline) and our proposals under different channel conditions

In order to consider what is the best proposal, we can observe that the algorithms which apply the synthesis distance (SLBG, SLBGF, SLBGD) obtain better results and the best results are achieved by the SLBGD algorithm. However, its high computational cost and the relatively small improvement in performance compared with SLBGF, the use of SLBGF algorithm is more suitable.

Moreover, table 1 also shows the results obtained through a 2nd-order model. In this model, we have set a threshold $\mu = 100$ as the minimum of combinations of two indexes observed in the training database. As can be observed, significative improvements are obtained in each quantization method, as source-model is further exploited. Nevertheless, as before, best results are obtained by the SLBGD and SLBGF approaches. Finally, it must be noted that our proposals do not incur in neither bitrate increase nor delay.

## 5 CONCLUSIONS

In this paper we have proposed a mitigation technique in which LPC coefficients and the residual signal are both concealed when long bursts appear. In order to obtain the replacements, a model of the speech source is exploited by an MMSE estimation. To this end, we have computed separated codebooks for LPC and residuals parameters. In particular, we present several algorithms which modify the LBG algorithm and minimize the synthesis error in case of the residual signal. The best residual codebook is obtained with the SLBGD algorithm but due to its complexity, the use of SLBGF is more suitable. We also tested our proposals through a 2nd-order model in which further improvements are obtained. Although in this work codebooks with the entire residual signal are considered, a better quantization can be achieved by splitting this signal into several subframes. This issue will be addressed in future work.

The objective quality tests have shown the suitability of our technique in adverse channel conditions. Moreover, the proposed technique does not increase neither the bitrate nor delay in the decoder while computational complexity is kept at minimum as the replacement vectors are obtained off-line. In addition, we have proposed an error mitigation technique which is not restricted to a particular codec.

## References

1. M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *IEEE ICASSP*, vol. 10, pp. 937–940, 1985.
2. 3GPP TS 26.090, "Mandatory speech codec speech processing functions; adaptive multi-rate (AMR) speech codec," 1999.
3. M. Serizawa and H. Ito, "A packet loss recovery method using packet arrived behind the playout time for CELP decoding," *IEEE ICASSP*, vol. 1, pp. 169–172, 2002.
4. M. Chibani, R. Lefebvre, and P. Gournay, "Fast recovery for a CELP-like speech codec after a frame erasure," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2485–2495, 2007.
5. A. Gomez, J. Carmona, A. Peinado, and 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. 18, pp. 1258–1268, Aug. 2010.
6. A. Gomez, J. Carmona, J. González, and V. Sánchez, "One-pulse FEC coding for robust CELP-coded speech transmission over erasure channels," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 894–904, 2011.
7. H. Ehara and K. Yoshida, "Decoder initializing technique for improving frame-erasure resilience of a celp speech codec," *IEEE Trans. Multimedia*, vol. 10, pp. 549–553, Apr. 2008.
8. D. Lopez-Oller, A. Gomez, and J. L. Pérez-Córdoba, "Source-based error mitigation for speech transmissions over erasure channels," *EUSIPCO*, Sept. 2014.
9. IETF RFC 3951, "Internet Low Bit Rate Codec (iLBC)," 2004.
10. Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Commun.*, vol. 28, no. 1, pp. 84–95, 1980.
11. A. Gomez, A. Peinado, V. Sánchez, and A. Rubio, "A source model mitigation technique for distributed speech recognition over lossy packet channels," *EUROSPEECH*, 2003.
12. F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, p. 35, 1975.
13. L. Kaufman and P. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," *Willey*, 1990.
14. ITU-T Recomendation P.862, "Perceptual evaluation of speech quality (PESQ)," 2001.
15. J.S. Garofolo et al., "The Structure and Format of the DARPA TIMIT CD-ROM Prototype," 1990.

# The implementation of medial stop clusters in European Portuguese:

## An articulatory account and some perceptual data

Conceição Cunha

Institute of Phonetics and Speech Processing, LMU Munich, Germany
cunha@phonetik.uni-muenchen.de

**Abstract.** The main aim of this paper was to investigate the temporal and spatial organisation of medial stop consonant clusters and its influence on perception. More specifically, timing and coordination of two following stops were investigated and related to its perception by native listeners. The hypothesis to be tested here is that if consonants are strongly overlapped, C2 may mask C1, leading to the perception of a singleton consonant and a discrepancy between production and perception.

In order to test this, physiological movement data were recorded using a 3D electromagnetic articulograph from five first language speakers of EP. For the measurements, temporal overlap and spatial magnitude has been analysed with RM-ANOVA in R. 37 native speakers of EP participated in the forced choice perception experiment with an identification task.

The results showed that stop clusters presented two consonant targets, but some of them were so overlapped, that listeners recover indeed only C2 in perception.

**Keywords:** EMA, consonant cluster, double articulation consonants, perception, overlap, compensation for coarticulation.

## 1 Introduction

In spoken Portuguese medial stop clusters present currently a great variation, in which both stops can be realized, the first element of the sequence can be deleted ([ɐtual] vs. [ɐktual], *actual*, ‘actual’) or both elements can be realized with an intermediate vocalic element ([kɐptaɾ] vs. [kɐpɨtaɾ], captar, ‘to get’, Mateus & d’Andrade 2000, Veloso 2007 for European Portuguese, EP). Evidence from Brazilian Portuguese (BP), in which intermediate vowel epenthesis is even more frequent than in EP (Bisol 1999), attests vowel epenthesis after bilabial, but its occurrence after velar consonant is unattested (*[fakitu] *facto*, ‘fact’) or at least seldom (?[kompakitadu] *compactado*, ‘compact’).

In a diachronic point of view, sequences of two stops were relatively unstable in the words that took part of the oral evolution from Latin to Romance: the first stop became vocalized or deleted in Portuguese (nŏcte > /nojte/, ’night’ and lūcta > /luta/

‘fight’), palatalized in Spanish giving rise to an affricate (nŏcte > /notʃe/ and lūcta > /lutʃa/), vocalized or a geminate in French (nŏcte > /nuit/ and lūcta > /lutte/) and a geminate in Italian (nŏcte > /notte/ and lūcta > /lotta/. The remaining stop sequences were directly imported from Classical Latin later on and they did not undergo the evolution processes from the spoken transmission mentioned before, Boyd-Bowman 1954: 41-43, Williams 1968: 86, 94-95).

From a speech motor perspective, stops are realised with a strong closure that can mask or cover adjacent segments maximally. Therefore, sequences of stops are not expected to be parallelly transmitted or to overlap without masking each other (Mattingly 1981). However, the place of articulation showed some influence on the recoverability of adjacent segments. Stop clusters in Georgian (Chitoran et al. 2002) and Moroccan Arabic (Gafos et al. 2010) showed consistently a place-order effect in production, in which both stops in front-to-back clusters (eg. /pk/) were timed closer together and showed more overlap than stops in the other way round (eg. /kp/). The pointed explanation was that sequencing of places of articulation in front-to-back order is perceptually more easily recoverable than in back-to-front clusters (Chitoran et al. 2002).

These results are in accordance with previous results showing shorter intervals between $C^1$ and $C^2$ closures in /tk/ than in /kt/. The production of the first cluster involve only tongue dorsum raising von /t/ to /k/, while repositioning of the tongue body may be needed from velar to alveolar position in /kt/ and the increased flexibility of the tongue tip or blade for /t/ may lead to ore coarticulation of this articulator, compared to the tongue dorsum for /k/ (Hardcastle and Roach, 1979). Large amounts of temporal overlap cause accommodation of the spatial distribution of the gestures leading to the bending of two sequence gestures in a new intermediate configuration (Browman and Golstein, 1989, Recasens et al. 1993).

In order to test, which of the precedent prediction better describe the organisation of stop clusters in Portuguese, we use physiological EMA-data to analyse the temporal and spatial organisation of bilabial-alveolar and velar alveolar clusters. In this analysis the synchronic variation will be related to the place order hypothesis in terms of temporal coordination and timing differences in the model of articulatory phonology (Browman & Goldstein, 1986, 1988, ff.). A second issue was to relate the temporal consonant coordination in production with an identification task in perception in order to test, if coordination is matched with the identification of consonant clusters, singleton consonants or intermediate vocalic elements. For a baseline with consonant overlap in production (e.g. Gafos 2002), the correspondence to the perception of a consonant cluster (CC) is expected if production and perception are matched. However, it is possible that changes on the consonantal overlap may not be compensated for and they lead to different perceptions. In this case we predict two scenarios: a) $C^2$ would be timed to occur later relative to $C^1$ in a /$C^1C^2$/ cluster and this would give rise to the perception of an intervening vocalic element between the consonants ($C^VC$) b). If $C^2$ would was timed closed after $C^1$, the latter may be strongly covered by $C^2$ and not perceptible by listeners. They may perceive the sequence as a single consonant, which should be $C^2$ or as a double articulation consonant. The last effect is expected to be even stronger in back-to-front clusters according to the predictions of the place-order hypothesis, in which velar C1 are even more difficult to recover than bilabial ones (Chitoran et al. 2002).

Regarding the Portuguese varieties, the frequent vowel reduction (Mateus & d'Andrade 2000) and the high vowel deletion (), suggest a trend to strong coarticulation in EP. The resulting prediction is stronger consonantal overlap in EP than in BP and, if perception follows overlap, listeners should also be more inclined to (mis)perceive more singleton consonants in EP than in BP. The variation of lexical stress permits to test, whether unstressed tokens show more coarticulation and overlap than stressed as found before (Byrd et al. 1996).

The last aim of this study was to investigate whether the listener's variety plays a role in the perception of the stop clusters. The prediction here is that clusters should be more readily perceived as singleton consonants by BP than EP listeners, if perception matches production: that is, if BP requires a wider phasing between consonants of a cluster in production, then BP listeners might also require a relatively wide phasing between consonants for them to hear a cluster rather than a singleton stop in perception. In particular, BP listeners should hear the majority of the tightly timed clusters in European Portuguese as singleton consonants.

## 2 Temporal and spatial organisation

### 2.1 Methods

Physiological movement data were recorded using a 3D electromagnetic articulograph from five first language speakers of EP and four speakers of BP. The EP participants were from the Porto region (three male, two female and aged between 24 – 29 years) and the BP speakers from the city of São Paulo (three male, one female and aged between 23 – 29 years). At the time of the recordings, the speakers were exchange students or visitors in Munich and had lived there less than ten months. They were all naïve regarding the purpose of the experiment and did not report any speech, reading or hearing disorders. The sensors were fixed mid-sagittally on the lips (upper and lower lip), jaw and three on the tongue (tip, mid, back). Reference sensors were attached on the maxilla, nose and on the left and right mastoid bones.

**Stimuli and data processing**

The speakers repeated every target word embedded in a carrier sentence 8 times. The stimuli consisted of the lexical words presented on table 1, containing velar or bilabial stop in $C^1$ position and alveolar stop in $C^2$ in medial position. The following vowel was held constant for each stimulus set.

**Table 1.** Target words for both experiments presented.

| **cluster** | captar<br>captador<br>compactado<br>compactador |
|---|---|

| | |
|---|---|
| **CeC** | tapetar<br>tapetador<br>quetado<br>quetadado |
| **CuC** | computar<br>computador<br>cutano<br>cutanoso |
| **C** | pada<br>padada<br>cádiz<br>cadastro |

The audio signal was automatically segmented on the basis of the acoustic signal using the Munich Automatic Segmentation System (MAUS: Schiel, 2004), the target cluster were corrected manually. The physiological trajectories was labelled with the EMU System (Harrington 2010) with a semi-automatic algorithm that identify the peak velocities of the closing (PVEL1) and opening (PVEL2) gesture and the maximal constriction (MAX) from the respective signal. MAX corresponds to the point of minimal velocity between the onset and offset of the target. The remaining values for gesture onset (START) and offset (END) as well as the achievement (TARGET) and release (RELEASE) of the constriction were interpolated from the velocities profiles and corresponded to the 20% threshold of the difference between the closest velocities peaks (see Bombien, 2011:71-80 for a detailed description).

For the physiological annotation of the bilabial segments, the lip aperture (la) was calculated as the Euclidean distance between upper and lower lip sensors. The other two consonants were defined based on the vertical movement of tongue tip (tt) for /t/ and tongue back (tb) for /k/.

### Measurements

OVERLAP of the velar or bilabial $C^1$ and the alveolar $C^2$ was defined as the difference between the points in time from $C^2$Start to $C^1$Target divided by the duration of the $C^1$Plateau (the plateau was defined as the duration between $C^1$Release and $C^1$Target) and subtracted by 1. By subtracting it from 1, higher values indicate more overlap (Gafos et al., 2010, Chitoran et al., 2002).

- An overlap value between 0-1 means that $C^2$ starts during the $C^1$Plateau (e.g. Gafos 2002, Gafos et al. 2010, Chitoran et al. 2002).
- A zero or a negative value means that $C^2$ starts simultaneously to or after the $C^1$Release, respectively, which means no overlap of both consonant gestures.
- A value greater than 1 means that $C^2$ starts before the $C^1$Plateau and both consonants overlap more.

For the spatial analysis we defined the magnitude of the tongue back and lip aperture for $C^1$ between the highest and lowest point in the movement of the closing gesture. The velar and bilabial stops in the cluster were compared with intervocalic singletons (VCV).

**Statistics**

The overlap variable were analysed in a Mixed Model (library lme4, Bates et al., 2011) in R with type (three levels: cluster, CeC vs. CuC), speaker variety (two levels: BP vs. EP), $C^1$ (two levels: bilabial vs. velar) and stress (two levels: stressed vs. unstressed) as fixed factors and speaker as a random factor. In the case of significant interactions between two or more factors, subsequent Tukey post-hoc-tests were run with the help of the multcomp library (Hothorn et al., 2008) in the same statistical environment.

## 2.2 Results

Fig. 1 displays the overlap measurements: The upper panels present the results for four BP and the lower panels for the five EP speakers with stressed tokens on the left and unstressed on the right. Lexical clusters plotted in black can be compared to the lexical CVC sequences with intermediate front vowel in dark grey and with back vowel in light grey. Positive values mean overlap of both consonants and negative values an intermediate lag between both consonants.

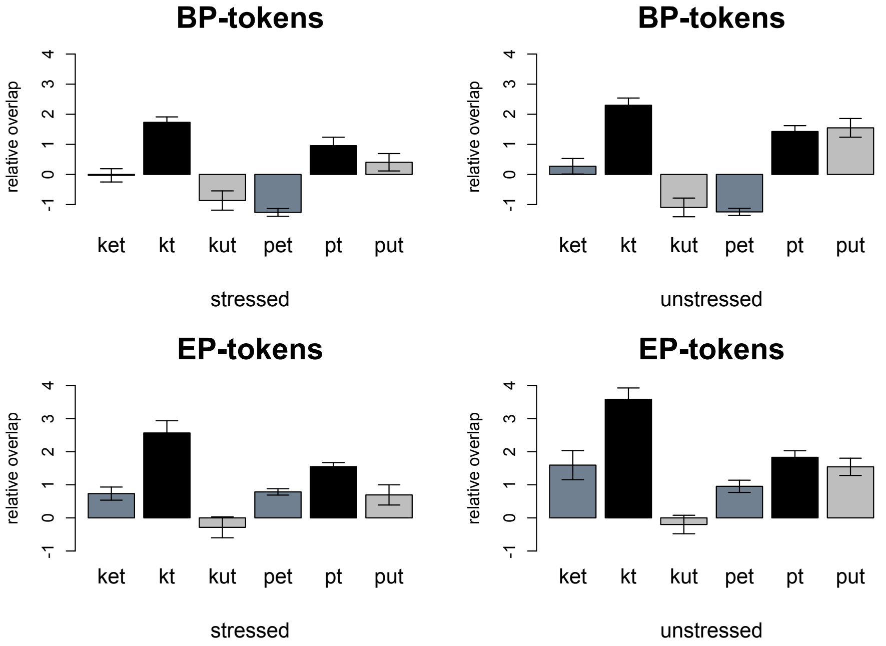


**Fig. 1.** Relative gesture overlap [1 − ($C^2$Onset- $C^1$Target)/($C^1$Release - $C^1$Target)] of the stop tokens. The graph shows means and standard deviations over four BP and five EP speakers. /Ct/, in which C=/p, k/, corresponds to lexical clusters, /CeC/ and /CuC/ corresponds to the CVC sequences with front and back vowels, respectively. For an easier comparison, types were plotted in the same colours and do not have any further meaning.

The most prominent main effect was of stimuli's type ($\chi^2[2] = 226.1$, $p < 0.001$) with lexical clusters showing most more overlap than CeC and CuC. However, it is evident that the positive values obtained in this data are much higher as predicted: The values are close to or greater than one, meaning that the second consonant starts before $C^1$ is achieved. Although similar trends are showed for both varieties, the speaker's variety had a significant main effect ($\chi^2[1] = 13.1$, $p < 0.01$) on the overlap values, in which tokens produced by EP participants showed more overlap than by BP, as well as place of articulation of $C^1$ ($\chi^2[2] = 62.7$, $p < 0.001$), with velar showing much more overlap than clusters starting with the bilabial $C^1$. Unstressed tokens showed more overlap than stressed as expected, nonetheless the effect was greater for the EP ($\chi^2[1] = 17.9$, $p < 0.001$) than for the BP tokens ($\chi^2[1] = 9.0$, $p < 0.01$). Tokens in EP showed more overlap than in BP, as well as velar position comparing to bilabial and unstressed comparing to stressed, so that the combination of the three factors corresponds to the most overlapped tokens (unstressed /kt/ in EP).

Magnitude of both consonants was not significantly smaller in cluster than in the singleton consonant. Consequently, we will concentrate on the relationship between temporal overlap and perception in the next sections.

## 3 Perception experiment

### 3.1 Methods

The stimuli for the perception experiment were recorded as part of a physiological study (cf. Section 3) with a multichannel DAT device from one native speaker of BP and one of EP. The target stimuli from the third repetition of both speakers were excised between their acoustic onset and offset and presented in randomized order in an online forced choice experiment. The participants listened to each stimulus separately and carried out an identification task in which they responded to each /p/-initial token with one of the following four choices <t>, <pt>, <pet>, <put> and to each /k/-initial token with one of the following four <t>, <kt>, <ket>, <kut>. Since the main research question in this study is the number of elements perceived, we analysed the responses for one (/t/) and three elements ($C^VC$) in separated mixed models.

37 native speakers of EP (11m/26f, originated from Porto and aged between 24-36 years) and 32 native speakers of BP (10m/22f, originated from Campinas and aged between 22-35 years) participated in the perception experiment. None of the subjects reported any hearing or reading problems.

Each listener group was subjected to all stimuli of the EP and BP speaker, respectively. The responses were analysed with two binomial generalized linear mixed model (GLMM) on the binary responses for the proportion of /t/ and $C^VC$ in the R statistical program with the listener name as a random factor and listener variety as one of the fixed factor. The listener groups will not be presented separately, because we did not found a significant influence of this variable (European vs. Brazilian listener group). The statistical models will be reported in detail with the presentation of the results.

### 3.2 Results

Fig. 2 present the perceptual answers of what were intended lexical clusters in production as produced by the BP speaker on the top and by the EP speaker on the bottom with the lexical clusters with bilabial $C^1$ in the upper bar and with velar $C^1$ in the lower bar. The both upper panels present the answers to the BP production tokens and the lower the EP token. Black corresponds to the perception of two following consonants ($/C^1C^2/$), grey to the perception of the same consonants with an intermediate vowel and light grey to one singleton consonant.

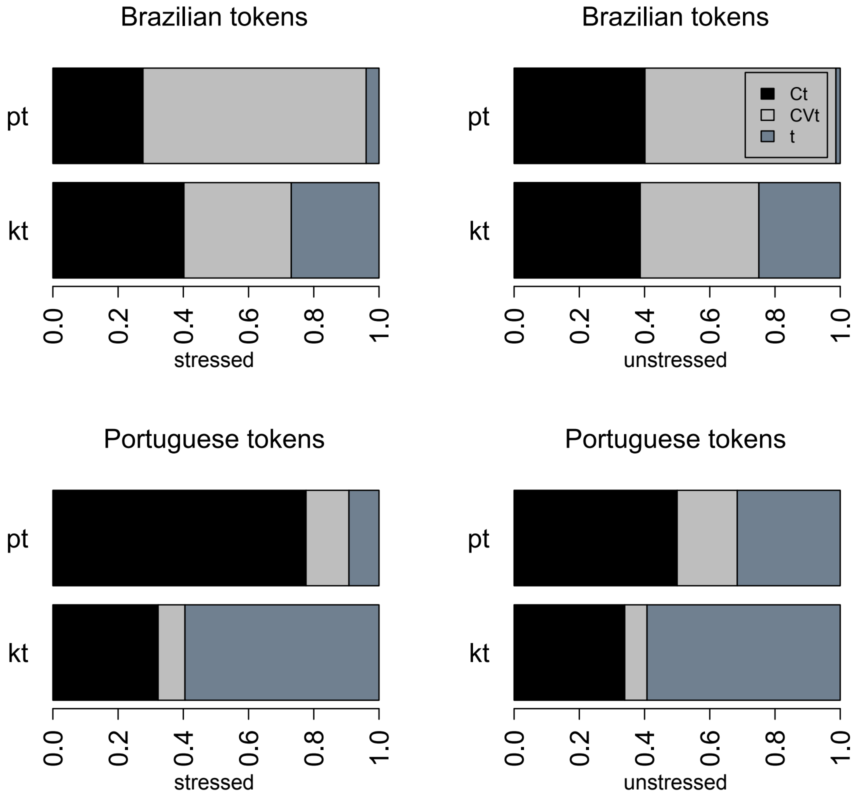


**Fig. 2.** Proportion of the responses to the BP tokens on the top and to the EP tokens in the bottom. The left panels show stressed and the right panels the unstressed tokens.

A quick look on Fig. 2 makes clear, that the perception of the lexical cluster showed much variation in the perception of intermediated /CVt/, cluster or singleton consonant. The mixed model on the percentage of sequence with intermediate vowel, irrespective of the vowel was run with speaker variety, $C^1$ and stress as fixed factors and listener as a random factor. It showed a main effect of speaker's variety ($\chi2[1]=7.3$, $p<0.001$) and $C^1$ ($\chi2[1]=7.5$, $p<0.001$). Stress showed a small trend to more identification of clusters in unstressed position, but this was not significant. Tukey post-hoc tests on the combination of both factors confirmed what became evident from Fig. 2: The variety of speaker had an influence on the percentage of /CVt/ for each cluster in both stress conditions, since the percentage of /CVt/ were significantly greater in the Brazilian than in the European tokens ($p<0.001$ for each combination). In the Brazilian tokens, the bilabial $C^1$ leaded to the perception of an intervocalic element ($p<0.001$). The same trend was true for the European perception, but it was significant just in the unstressed condition ($p<0.002$) for this variety. Stress and listener variety were further tested separately for each production set, but both factors had no significant influence on the perceptual responses.

The proportion of singleton /t/ analysed with a second mixed model showed a significant main effect of speaker ($\chi2[1]=8.9$, $p<0.001$) and C1 ($\chi2[1]=5.3$, $p<0.001$) and a significant interaction between the factors ($\chi2[4]=32.0$, $p<0.001$). In the perceptual responses the proportion of singleton /t/ was significantly greater in velar context for the EP than for the BP data ($p<0.001$ in both stress conditions). After bilabial consonant the varieties had an influence on unstressed /pt/ ($p<0.001$) with EP showing more singleton than BP, in stressed the same trend in visible but it was not significant.

Velar $C^1$ leaded to a greater amount of singletons than bilabial in both varieties ($p<0.001$ for both varieties in both stress conditions). The stress condition had an influence only in the perception of EP /pt/, in which unstressed tokens were perceived often as singleton than stressed ($p<0.001$). All other combinations showed no effect of stress.

## 4 Discussion

The strong overlap of the stops in consonant clusters found here was much greater than predicted. This data seem to contradict the idea that stops due to the strong closure do not overlap strongly in order to not mask the preceding stop. The implications of the strong overlap for perception were tested with a part of the stimuli. The perceptual results clear showed that the majority of the strong overlapped clusters in EP was perceived as singleton consonants. Therefore, the strong overlap cannot be compensate for in perception, since listeners just give the second element in perception.

The analysis of magnitude showed that both gestures still being realized, so that blending of the gestures could not be confirmed. However, the implication of this experiment was that even being produced both gestures in EP, they cannot be recover anymore even by native listeners. In this sense, production and perception do not match, since two gestures are produced and just one can be recovered. The came across due to unstable timing relationship, in which two gestures with a strong closure are so tiny timed that the first one may not be released and cannot be recovered in perception.

In BP both stops overlap less, but a similar trend was found. If velar-alveolar stop clusters cannot be recovered in perception, they may compromise in the language in a near future.

The strong overlap in EP can be related with the stronger coarticulation of the segments in EP and the frequent high vowel deletion (Silva 1998). The comparison between EP and BP make the relationship between production and perception of stop clusters even more evident, since smaller overlap lead to the perception of more elements.

The place-order effect seems to be contradicted in the production part, since sequences with velar $C^1$ presented more overlap than with bilabial. However, the perception part make clear implications of strong overlap in this context for perception with the lost of over 60 % of the cluster in EP.

## 5 Conclusion

This study clear showed an apparent contradiction between production and perception, with two elements being produced and just one of them perceived. However, the apparent contradiction could be explained with the clear relationship between overlap in production and the perception of stop clusters in two Portuguese varieties.

## 6 References

1. Bisol, L. 1999. „A sílaba e seus constituintes. In M. H. M. Neves“. In Gramática do Português Falado, 7:701–741. Campinas: Humanitas/ ed. Unicamp.
2. Bombien, L. 2011. „Segmental and prosodic aspects in the production of consonant clusters“. Ludwig-Maximilians-Universität München. Dissertation.
3. Boyd-Bowman, P. 1980. From Latin to Romance in Sound Charts. Washington, D.C.: Georgetown Univ. Press.
4. Browman, C. P, and L. Goldstein. 1986. Towards an articulatory phonology. *Phonology yearbook 3* (21): 9–252.
5. Browman, C. P, und L. Goldstein. 1988. „Some notes on syllable structure in articulatory phonology“. Phonetica 45 (2-4): 140–155.
6. Browman, C. P and L. Goldstein. 1992: “Targetless” schwa: an articulatory analysis. In: G.J. Docherty and D.R. Ladd. *Papers in Laboratory Phonology* II: Gesture, Segment, Prosody. Cambridge: Cambridge University Press, 26-67.
7. Browman, C., and L. Goldstein. 2000. Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlée* 5: 25–34.
8. Byrd, D. 1996. Influences on articulatory timing in consonant sequences. *Journal of Phonetics* 24: 209–244.
9. Chitoran, Ioana, L. Goldstein, and D. Byrd. 2002. Gestural Overlap and Recoverability: Articulatory Evidence from Georgian. In *Laboratory Phonology*, 419–448. 7. Berlin, New York: Mouton de Gruyter.
10. Cunha, Conceição. *Submitted*. Portuguese lexical clusters and CVC sequences in speech perception and production.
11. Gafos, Adamandis. 2002. A grammar of gestural coordination. Natural Language and Linguistic Theory 20: 269–337.

12. Gafos, A., P. Hoole, D. Roon, und C. Zeroual. 2010. „Variation in overlap and phonological grammar in Moroccan Arabic clusters". Laboratory Phonology X, Mouton de Gruyter, Berlin/New York.
13. Harrington, J. 2010. Phonetic Analysis of Speech Corpora. 1. Aufl. John Wiley & Sons.
14. Kühnert, Barbara, Philip. Hoole, Christine. Mooshammer. 2006. Gestural overlap and C-center in selected French consonant clusters. *Proc. 7th International Seminar on Speech Production*, UFMG Belo Horizonte, pp. 327-334.
15. Marin, Stefania., and Marianne. Pouplier. 2010. Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model. Motor Control 14 (3): 380–407.
16. Mateus, M. H., und E. d'Andrade. 2000. The Phonology of Portuguese. Oxford University Press.
17. Mattingly, Ignatius. G. 1981. „Phonetic Representation and Speech Synthesis by rule T. Myers, J. Laver & J. Anderson". The cognitive representation of speech: 415–42.
18. Recasens, D., Fontdevila, J., Pallarès, M.D. & Solanas, A. 1993. An electropalatographic study of stop consonant clusters, Speech Communication, 12, 335-355.
19. Silva, D. J. 1998. „Vowel Lenition in São Miguel Portuguese". Hispania 81: 166–178.
20. Veloso, J. 2007. Da influência do conhecimento ortográfico sobre o conhecimiento fonológico. München: Lincom.

# First steps towards Skipping NNLMs

A. Palacios-Corella[1], F. Zamora-Martínez[2], S. España-Boquera[1], and M.J. Castro-Bleda[1]

[1] Departamento de Sistemas Informáticos y Computación,
Universitat Politècnica de València, Valencia, Spain
[2] Departamento de Ciencias Físicas, Matemáticas y de la Computación,
Universidad CEU Cardenal Herrera, Alfara del Patriarca (Valencia), Spain

**Abstract.** Statistical language modeling greatly suffers from the effects of data sparsity which is tackled by means of smoothing techniques. Continuous space language models are able to interpolate unseen word histories but new problems and challenges arise, as a very high computational cost during evaluation of $N$-gram probabilities, due to the softmax normalization constant. Several approaches to study how to reduce this computational cost have been proposed in the literature. This work tries to improve the use of pre-computed softmax normalization constants tables by including the Skipping $N$-grams technique into Neural Network Language Models (NN LMs) and describes some experiments conducted on IAM-DB corpus to validate the viability of the proposed technique. The skipping for NN LMs works as regularization, but additionally helps to simplify the use of pre-computation of softmax normalization constants, as will be shown in the preliminary experiments of this paper.

## 1 Introduction

The estimation of a-priori probabilities of word sequences is one of the main purposes of language models (LMs). They play a key role in many natural language processing applications such as speech or text recognition, machine translation, part-of-speech tagging, or information retrieval. The most commonly used LMs are statistical $N$-gram models [6, 4], which only consider the $N-1$ previous words to estimate the LM probability for a sequence of words of length $|W|$:

$$p(w_1 \dots w_{|W|}) \approx \prod_{i=1}^{|W|} p(w_i | w_{i-n+1} \dots w_{i-1}) \tag{1}$$

The use of $N$-grams is usually restricted, in practice, to low orders, as is the case of trigrams. Although trigram LMs work well in practice, there are many improvements over this simple model, including higher-order $N$-grams, smoothing techniques, skipping models, clustering models or cache models [4].

The estimation of these models are based on counting occurrences of word histories in a training corpus. A more recent alternative to the classical "count-based" $N$-grams are those based on a continuous representation of the lexicon using connectionist approaches, as is the case of Neural Network Language Models (NN LMs) based on multilayer perceptrons [15, 18, 19].

NN LMs do not require the use of explicit smoothing techniques usually employed in count-based $N$-grams (e.g., backing-off), but important computational issues appear when using large vocabularies, majorly due to output softmax activation function. Short-list [15] and pre-computation of softmax normalization constants [18, 19] allow to reduce significantly this computational cost.

This paper describes the first steps of this research and its main contribution is to study whether the connectionist skipping $N$-gram LMs can help to improve NN LMs performance as a regularization method and as a new technique for smoothed NN LMs presented in [18].

## 2 Related language models

### 2.1 Skipping $N$-gram LMs

$N$-grams suffer from data sparsity making it necessary the use of smoothing techniques. Skipping $N$-grams [5, 14, 13, 10, 16, 4] can improve the generalization ability of standard smoothing techniques. The idea is that the exact context will have not probably been seen during training, but the chance of having seen a similar context (with gaps that are skipped over) increases as the order of the $N$-gram does.

Let us explain the idea with an example: suppose that the sentence *"The little boy likes pizza very much"* appears in the training data and we are trying to estimate a 5-gram. The training sentence has contributed to the estimation of $p(\text{pizza}|\text{the little boy likes})$. Unfortunately, this sentence cannot help much in the estimation of $p(\text{pizza}|\text{the little girl likes})$. The usual technique of backing-off would consist of using lower order N-grams. In this case, we would need to descend until $p(\text{pizza}|\text{likes})$. By skipping some words from the past history, the training sentence is useful to estimate $p(\text{pizza}|\text{the little — likes})$. For instance, the probability of $p(\text{pizza}|\text{the little boy eats})$ would benefit from $p(\text{pizza}|\text{the little boy —})$ whereas backing-off would require to descent until unigrams as far as the example sentence is concerned.

Skipping $N$-grams are not only based on skipping words from the past history but also on the combination of different ways of performing these skips. Each different way of skipping words can be considered a lower order LM in some way and the Skipping $N$-grams are a mixture of them. For example, the representation of several skipped trigrams (at most two context words are not skipped) may approximate a higher order $N$-gram using less resources, which explains why some authors have considered Skipping $N$-grams as a *poor man's* alternative to higher order $N$-grams. Nevertheless, our emphasis here is that they can also be useful to improve NN LMs probability computation.

### 2.2 NNLMs

NN LMs are able to learn a continuous representation of the lexicon [15, 18, 19]. A scheme of a NN LM is illustrated in Figure 1 where a multilayer perceptron

(MLP) is used to estimate $p(w_i|w_{i-n+1}\dots w_{i-1})$. There is an output neuron for each word $w_i$ in a vocabulary $\Omega'$, a subset of the most frequent words of the task vocabulary $\Omega$, allowing to increase computation of output layer.[3] The input of the NN LM is composed of the sequence $w_{i-n+1},\dots,w_{i-1}$. A local encoding scheme would be a natural representation of the input words, but it is not suitable for large vocabulary tasks due to the huge size of the resulting Neural Network (NN). To overcome this problem, a distributed representation for each word is learned by means of a projection layer. The mapping is learned by backpropagation in the same way as the other weights in the NN. After the projection layer, a hidden layer with non-linear activation function is used and an output layer with the softmax function will represent the $N$-gram LM probability distribution. The projection layer can be removed from the network after training, since it is much more efficient to replace it by a pre-computed table which stores the distributed encoding of each word. In order to alleviate problems with rare words, the input is restricted to words with frequency greater than a threshold $K$ in training data.

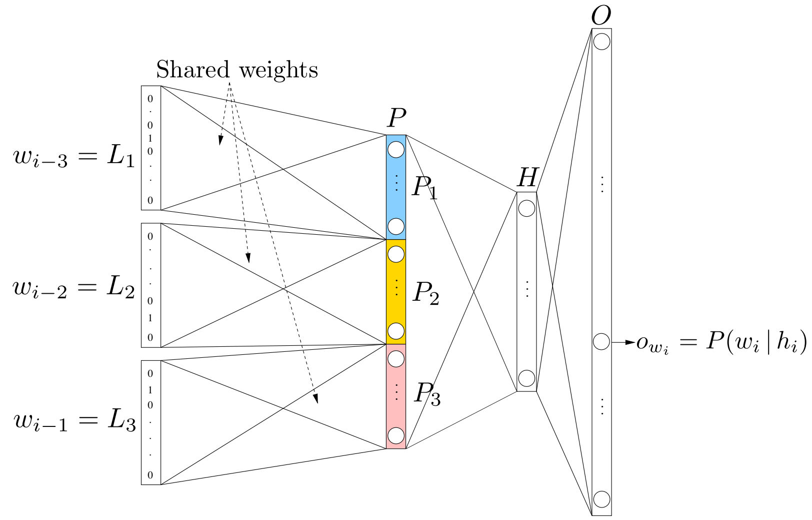


Fig. 1: The architecture of a 4-gram NN LM during training. In this example, the history of word $w_i$ is represented by $h_i = w_{i-1,}, w_{i-2}, w_{i-3}$.

Pre-computation of softmax normalization constants [18] has been proposed as a practical solution to the high computational cost of NN LMs output layer. Following our previous work, in order to compute 4-gram probabilities by using

[3] Relating to the output layer, several techniques exist to reduce its computational cost [8, 15, 18, 2, 12]. Short-list approach [15] is the one used in this work and requires to combine the NN LMs with another LM for words out of the short-list.

NN LMs, lower order NN LMs are needed to compute $N$-gram probabilities for contexts which softmax normalization constant has not been pre-computed. In this way, the model performance decreases but a significant time speed-up is obtained. One drawback of this approach is that it forces the training of several NN LMs for different $N$-gram orders.

## 3 Skipping NNLMs

Since skipping techniques consist of learning several LMs based on different ways of removing words from the past history, a natural way of introducing Skipping NN LMs is by training a sole MLP but replacing random positions in context history by a new special symbol ⟨NONE⟩. There exist $2^{N-1}$ different ways of replacing some of the $N-1$ words by ⟨NONE⟩. After training, $2^{N-1}$ models are available, each one with a fixed pattern of skipped positions. The perplexity (PPL) of each possible skipping model will be evaluated in this paper. Additionally, skipping the proper positions, it is possible to convert a 5-gram NN LM into a 4-gram, 3-gram and so on. Softmax normalization constants can be pre-computed for this unique model using LM training data. When a constant is not found, instead of using a totally different model with lower order, it is possible to use the same model but with skipping positions which reduce its order. We have evaluated this approach, a comparison of PPL results of skipped lower order models and *true* lower order NN LMs will be performed in this paper.

It is worth mentioning that the technique described in [11], also coined as continuous skip-ngram model by their authors, is not related to this work despite the similarity in nomenclature: the main purpose of [11] consists in learning a continuous representation of the lexicon of a task in a way that captures the semantic similarity w.r.t. the task. This representation could be used, for instance, in NN LMs. Their continuous skip-ngram models are log-bilinear classifiers which receive a word at the input layer and try to predict the neighboring (past and future) words.

## 4 Experimental setup

The NN LMs used as baseline for our experiments are fully described in [18, 19]. These models will be extended to construct the Skipping NN LM as described in the previous section. The experiment is conducted on a task based on the transcriptions of the IAM-DB [9] handwritten dataset. The training material to estimate the LMs is taken from LOB [7], Wellington [1] and Brown [3] corpora, which add up to 402K sentences in total. Validation and test sets were taken from IAM-DB corpus (920 and 2781 lines, respectively). Since IAM-DB texts are extracted from LOB corpus, IAM-DB validation and test sentences have been removed from LOB text data. PPL results will be presented for validation and test parts of the IAM-DB.

Backprogation algorithm and L2 regularized cross-entropy loss function are used to train NN LMs and Skipping NN LMs. A projection layer with 256 neurons and a hidden layer with 200 neurons has been used based in previous works. For Skipping NN LMs, the input is stochastically perturbed, introducing the ⟨NONE⟩ symbol in zero or more random positions of the input layer. For every training pattern, the number of skips is sampled from a multinomial distribution. Given a number of skips, the positions of them are uniformly distributed. The multinomial distribution is defined to assign a probability of 50% for no skips at all, whereas the other 50% is distributed over one, two or more skips following a hyperbolic trend (see Table 1).

Table 1: The number of skips is sampled from the following multinomial distributions, given $N$-gram order.

| # skips | bigram | 3-gram | 4-gram | 5-gram |
|---|---|---|---|---|
| 0 | 50% | 50% | 50% | 50% |
| 1 | 50% | 33% | 27% | 24% |
| 2 | – | 16% | 13% | 12% |
| 3 | – | – | 9% | 8% |
| 4 | – | – | – | 6% |

Once the model is trained, $2^{N-1}$ different LMs can be build, having each one a different combination of skips or ⟨NONE⟩ tokens in its input layer. Although one of the main benefits of Skipping NN LMs is to emulate lower order $N$-grams in order to greatly simplify the speed-up technique based on memorizing softmax normalization constants, in this work we focus on: first, investigating these emulation capabilities, and, second, computing the probabilities of the first $N1$ words of a sentence (for higher order $N$-grams) before the complete word history is available.

## 5 Experimental results

First, an evaluation of the PPL trend for validation data depending on the none tokens has been performed (see Figure 2). PPL is computed for all of the $2^{N-1}$ possible skipping positions. To better interpret the results, let us remark that the skipping number is a right-to-left bits mask where "0" indicates no skip and "1" indicates a skip. So, the skip number 7 in a 5-gram refers to the binary representation 1110, meaning that probability of $N$-gram at position $i$ is computed by using word $i-1$ and three ⟨NONE⟩ tokens. The complete Skipping NN LMs models consider the combination of the different LMs associated to the skipping masks. We have performed this combination, for each $N$-gram order, using the `compute-best-mix` tool from SRILM toolkit [17]. The obtained results are very similar and competitive with those of standard NN LMs but they are not able to outperform this baseline.

As can be observed in Figure 2, the Skipping NN LMs with a skipping mask without ⟨NONE⟩ performs as well as standard NN LMs while the presence of

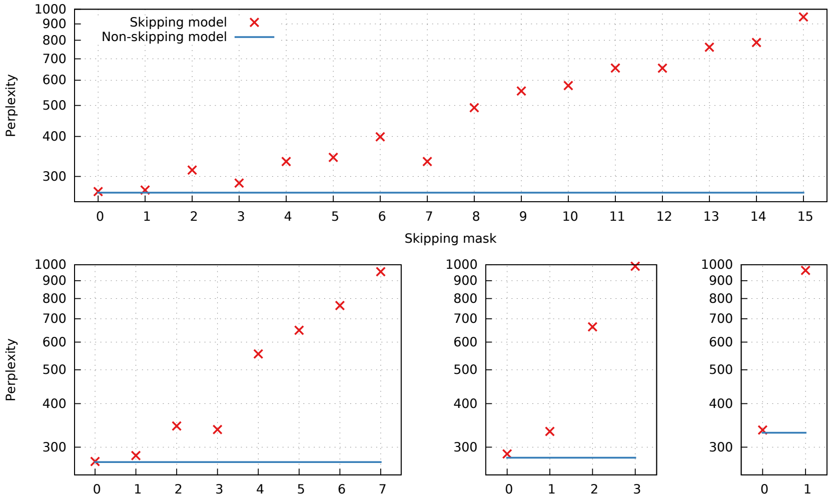


Fig. 2: PPL measured on validation set for Skipping NN LMs varying the skipping mask together with the baseline PPL of the non-skipping NN LM. The vertical axes of the four plots show the PPL and the skipping mask appears on the horizontal axes using the binary notation described in the paper. Upper plot corresponds to 5-grams. Lower plots correspond, from left to right, to 4-grams, trigrams and bigrams.

⟨NONE⟩ downgrades the results and this effect is more pronounced when ⟨NONE⟩ is close to the word to be predicted.

In order to evaluate the future ability of Skipping NN LMs to simplify the work presented at [18], Table 2 shows the PPL obtained by using no skipping NN LMs and Skipping NN LMs where the skipping positions are set to model equal or lower orders. Values from the first row of each table are the PPL values for each one of the NN LMs. The remaining rows contain the values from each of the Skipping NN LMs. The last non-void value of these rows is the PPL for the set without perturbing its input. The previous values are obtained after using skipping configurations that emulate a lower order ngram. For example, the 4-gram value for 5-gram Skipping NN LM is obtained after using the skipping mask that replaces the furthest word from the context by ⟨NONE⟩. To get the trigram value we replace the next word from the context too, and so on.

We can observe that the column values for each table are quite similar. This means that the PPL values for each one of the NN LMs are similar to the ones obtained for the Skipping NN LMs which can compute them using the adequate skipping configuration. Therefore, it is possible to train a single Skipping NN LM to imitate the behaviour of several NN LMs. Backing off for Skipping NN LMs has a simpler training procedure, you need only one model, while for NN LMs you need one model for every $N$-gram order [18].

Table 2: PPL results for IAM-DB validation (left) and test (right). For skip models the PPL has been computed for 0 skips and for skip combinations which simulate a lower order model.

| | Ngram order | | | | |
|---|---|---|---|---|---|
| Model | 1 | 2 | 3 | 4 | 5 |
| 5gr No skip | – | – | – | – | 267 |
| 4gr No skip | – | – | – | 272 | – |
| 3gr No skip | – | – | 280 | – | – |
| 2gr No skip | – | 330 | – | – | – |
| 5gr skip | 946 | 334 | 286 | 272 | 269 |
| 4gr skip | 955 | 337 | 284 | 273 | – |
| 3gr skip | 990 | 333 | 287 | – | – |
| 2gr skip | 963 | 336 | – | – | – |

| | Ngram order | | | | |
|---|---|---|---|---|---|
| Model | 1 | 2 | 3 | 4 | 5 |
| 5gr No skip | – | – | – | – | 309 |
| 4gr No skip | – | – | – | 313 | – |
| 3gr No skip | – | – | 319 | – | – |
| 2gr No skip | – | 376 | – | – | – |
| 5gr skip | 1019 | 378 | 326 | 311 | 309 |
| 4gr skip | 1025 | 383 | 324 | 313 | – |
| 3gr skip | 1045 | 377 | 327 | – | – |
| 2gr skip | 1026 | 381 | – | – | – |

## 6 Conclusions and future work

This work is, to the best of our knowledge, the very first attempt to integrate the well known technique of Skipping $N$-grams into NN LMs. NN LMs are capable of learning distributed representations which might explain that non additional gain is obtained by including the skipping technique. Coming back to the example of *the little boy* which *likes pizza*, from Section 2.1, it is possible that in a large corpus the contexts of words *boy* and *girl* are similar enough to make it possible for the MLP to learn a similar representation for these words so that the effect of skipping is diminished.

On the other side, the capability of Skipping NN LMs to emulate lower order NN LMs makes them very suitable for greatly simplifying the speed-up technique based on pre-computation of softmax normalization constants [18] since these models rely on lower order models when a constant is not found.

As a future work, we plan to investigate the effect of this technique in larger corpora to give more support to the preliminary results presented here and to study the effect of the new LM in the overall error of a recognition system.

## Acknowledgments

This work has been partially supported by the Spanish Government TIN2010-18958.

## References

1. Bauer, L.: Manual of Information to Accompany The Wellington Corpus of Written New Zealand English. Tech. rep., Department of Linguistics, Victoria University, Wellington, New Zealand (1993)

2. Bengio, Y., Senecal, J.S.: Adaptive importance sampling to accelerate training of a neural probabilistic language model. IEEE Transactions on Neural Networks 19(4), 713–722 (2008)
3. Francis, W., Kucera, H.: Brown Corpus Manual, Manual of Information to accompany A Standard Corpus of Present-Day Edited American English. Tech. rep., Department of Linguistics, Brown University, Providence, Rhode Island, US (1979)
4. Goodman, J.T.: A Bit of Progress in Language Modeling - Extended Version. Tech. Rep. MSR-TR-2001-72, Microsoft Research, One Microsoft Way Redmond, WA 98052 (2001)
5. Huang, X., Alleva, F., Hon, H.W., Hwang, M.Y., Lee, K.F., Rosenfeld, R.: The SPHINX-II speech recognition system: an overview. Computer Speech and Language 7(2), 137–148 (1993)
6. Jelinek, F.: Statistical Methods for Speech Recognition. Language, Speech, and Communication, The MIT Press (1997)
7. Johansson, S., Atwell, E., Garside, R., Leech, G.: The Tagged LOB Corpus: User's Manual. Tech. rep., Norwegian Computing Centre for the Humanities, Bergen, Norway (1986)
8. Le-Hai, S., Oparin, I., Alexandre, A., Gauvaing, J.L., Franois, Y.: Structured Output Layer Neural Network Language Model. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). vol. 11, pp. 5524–5527 (2011)
9. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. International Journal of Document Analysis and Recognition 5, 39–46 (2002)
10. Martin, S., Hamacher, C., Liermann, J., Wessel, F., Ney, H.: Assessment of smoothing methods and complex stochastic language modeling. In: Proc. 6th European Conference on Speech Communications and Technology (Eurospeech). vol. 5, pp. 1939–1942 (1999)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
12. Mnih, A., Kavukcuoglu, K.: Learning word embeddings efficiently with noise-contrastive estimation. In: Advances in Neural Information Processing Systems 26, pp. 2265–2273 (2013)
13. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependences in stochastic language modeling. Computer Speech and Language 8(4), 1–38 (1994)
14. Rosenfeld, R.: Adaptative statistical language modeling: A maximum entropy approach. Ph.D. thesis, Carnegie Mellon University (1994)
15. Schwenk, H.: Continuous space language models. Computer Speech and Language 21(3), 492–518 (2007)
16. Siu, M., Ostendorf, M.: Variable n-grams and extensions for conversational speech language modeling. IEEE Trans. Speech and Audio Processsing 8(1), 63–75 (2000)
17. Stolcke, A.: SRILM: an extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). pp. 901–904 (2002)
18. Zamora-Martínez, F., Castro-Bleda, M., España-Boquera, S.: Fast Evaluation of Connectionist Language Models. In: International Work-Conference on Artificial Neural Networks, LNCS, vol. 5517, pp. 33–40. Springer (2009)
19. Zamora-Martínez, F., Frinken, V., España-Boquera, S., Castro-Bleda, M., Fischer, A., Bunke, H.: Neural network language models for off-line handwriting recognition. Pattern Recognition 47(4), 1642 – 1652 (2014)

# A study of the quality of automatic speech recognition in distributed systems

Emilio Granell and Carlos-D. Martínez-Hinarejos

Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain
{egranell, cmartine}@dsic.upv.es

**Abstract.** In the study presented in this paper, we verified how the speech recognition quality in distributed systems depends on the client device. We confirmed a quality reduction on mobile devices when compared with the obtained quality from a simultaneously taken signal from a laptop computer which acts as reference. We analysed the performance of applying a gain modification to the signals and the performance of applying Maximum Likelihood Linear Regression speaker/device adaptation. We also analysed the performance of applying the gain modification before the speaker/device adaptation, and vice versa. Results allow to propose a practical solution to improve speech recognition quality in distributed systems.

**Keywords:** Speech recognition, speech enhancement, mobile devices, distributed system

## 1 Introduction

In the last years, smartphones have experienced an exponential increasing in its processing capabilities, in its available memory, in its connectivity options, and in the type of its embedded sensors. In addition, distributed systems allow to perform more complex tasks by delegating the computational load on the server.

Taking into account that the fastest and most effective way of communication for human beings is speech, we can think about the speech in natural language as one of the most important forms of interaction with smartphones. Two great examples of distributed ASR (Automatic Speech Recognition) systems are the recognisers of Google and Apple (Siri) for Android and iOS, respectively.

However, distributed ASR is only one of the available skills in more ambitious projects such as the *Percepción* project [1], whose main aim is to assist people to interact with the environment using mobile devices as a "sixth sense".

During the development of a distributed ASR system, in particular as partners of the *Percepción* project, we could see how the quality of speech recognition when using mobile devices was significantly degraded with respect to the quality obtained when using a computer. Considering that we used the same speech recogniser, with the same configuration and same models (acoustic, lexical, and language), the main differences that could produce such a degradation in the ASR quality were on the environment, the mobile devices, and the speakers.

In multichannel speech communication systems, the speech signal can be enhanced by using techniques such as blind-matched filtering [2]. This technique allows to estimate the acoustic transfer function of each channel, getting a combined output with a maximised SNR (Signal-Noise Ratio). Our system is single channel, and consequently this technique is not applicable. However, estimating a matched filter to the transfer function of the device would mean a substantial improvement in the quality of speech recognition.

**Table 1.** Client devices used in this study.

| Device | Name | Model | Operating System |
|---|---|---|---|
| Tablet | Samsung Galaxy 2 | GT-P3110 | Android 4.2.2 |
| Smartphone | Samsung Galaxy SIII mini | GT-I8190 | Android 4.4.2 |
| Computer | MacBook Pro | 7.1 Mid-2010 | OS X 10.9.3 |

In the study presented in this paper, we verified that the degradation of the quality of the recognition is device-dependent. We analysed the performance of applying a gain modification to the signals and the performance of adapting the acoustic models to the speaker/device. Finally, based on the results of this study, we proposed a practical solution to improve distributed speech recognition quality.

The rest of the paper is organised as follows: Section 2 presents our hypothesis; Section 3 introduces the experimental framework used to verify our hypothesis; Section 4 explains the performed experiments and the obtained results; finally, Section 5 offers the conclusions.

## 2 Hipothesis

We assume that the distributed ASR system (see Figure 1) is linear and time invariant. Hence, the ASR input signals recorded by client devices $x_d(t)$ can be expressed by the convolution of the speech signal $s(t)$ with the environmental $h_d^e(t)$ and internal $h_d^i(t)$ impulse responses of each client device $d$:

$$x_d(t) = s(t) * h_d^e(t) * h_d^i(t) \tag{1}$$

The Equation (1) is transformed into the following equation in the frequency domain:

$$X_d(f) = S(f) H_d^e(f) H_d^i(f) \tag{2}$$

Since the speech signal produced by users $S(f)$ was the same for each device, the ASR quality differences for each device $d$ were produced only by its impulse response $H_d(f)$:

$$H_d(f) = H_d^e(f) H_d^i(f) \tag{3}$$

Considering Equations (2) and (3), the correlation between the computer signal $X_c(f)$ and the mobile device signals ($X_t(f)$ and $X_s(f)$ for the tablet and the smartphone, respectively) is:

$$X_c(f) = X_t(f) \frac{H_c(f)}{H_t(f)} = X_s(f) \frac{H_c(f)}{H_s(f)} \tag{4}$$

Therefore, our hypothesis is that with an appropriate model, and on the same environmental and speaker conditions, ASR quality with mobile devices must be similar to that obtained with the computer.

## 3 Experimental framework

With the objective of checking our hypothesis, we performed several experiments using a tablet, a smartphone, and a computer as client devices to acquire simultaneously the voice signal. Table 1 shows the list of devices used in this study.

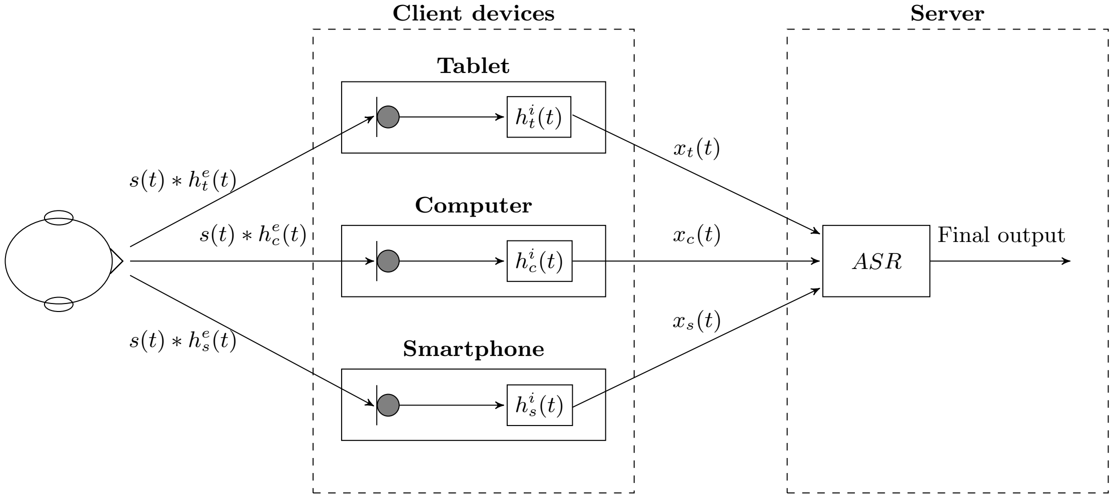


**Fig. 1.** Voice acquisition and distributed ASR scenario.

### 3.1 Voice signal acquisition

The acquisition of the voice signal was performed with a sampling rate of 16 kHz and a 16-bit quantisation, and stored in uncompressed files. For mobile devices, the acquisition was carried out by using an Android application, and in the case of the computer, by using a Java application. Both applications had the same audio recording configuration. Figure 2 shows a picture of the voice acquisition scenario. We tried to get the same speech signal in the three devices in a form of use as natural as possible. For this reason speakers were sat just in front of the computer, having the tablet slightly at left and the smartphone slightly at right of the computer.

### 3.2 Speech features

MFCC (Mel-Frequency Cepstral Coefficients) were extracted from the audio files. The Fourier transform was calculated every 10ms over a window of 25ms of a pre-emphasised signal. Next, 23 equidistant Mel scale triangular filters were applied and the filters outputs were logarithmised. Finally, to obtain the MFCC a discrete cosine transformation was applied. We used the first 12 MFCC and log frame energy with first and second order derivatives as features, resulting in a 39 dimensional vector.

### 3.3 Acoustic modeling

For the training of the acoustic models we used a partition of the Spanish phonetic corpus Albayzin [4]. This corpus consists of a set of three sub-corpus recorded by 304 speakers using a sampling rate of 16 kHz and a 16-bit quantisation. The training partition used in this work included a set of 4800 phonetically balanced utterances, specifically, 200 utterances read by four speakers and 25 utterances read by 160 speakers, with a total length of about 4 hours. A set of 25 HMM (Hidden Markov Models) (23 monophones, short silence, and long silence) was estimated from this corpus by using HTK [7]. Each HMM was formed by 3 states with 64 gaussians each state, giving a total of 4608 gaussians with diagonal covariance matrixes.

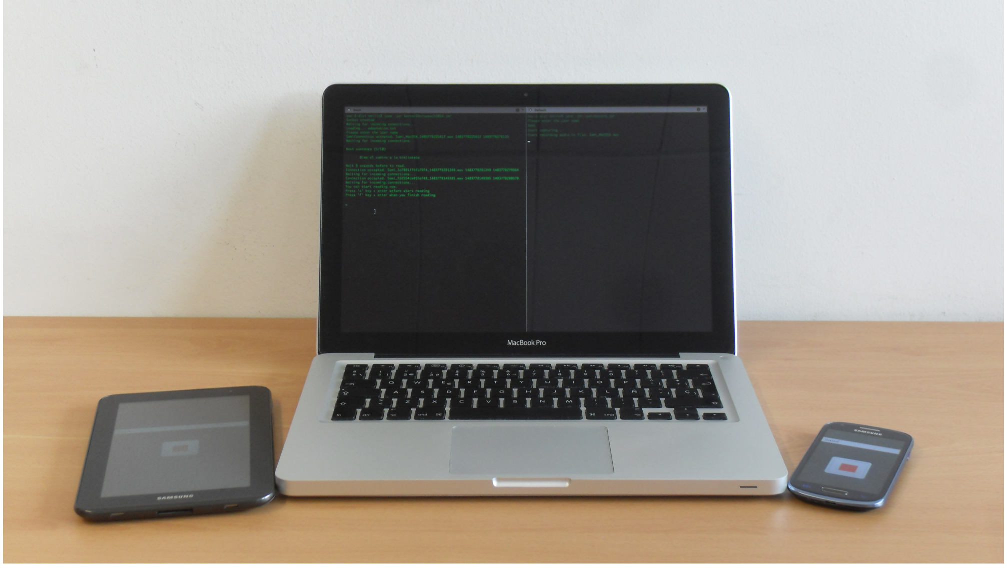


**Fig. 2.** Voice acquisition scenario.

### 3.4 Task, language and lexicon models

The task that was used in this study is an oral interface for interaction in a Smart City environment, specifically, the Smart Campus application of the UJI [1]. This application permits users to find the location of spaces such as buildings, parking areas, waste deposit areas, interior spaces and services in the university campus. This Smart Campus application uses a dialogue manager in order to enhance the quality of speech recognition, and consequently improve the user experience. For each dialogue state the dialogue manager disposes of a different language model. These language models are based on a set of finite-state models that determine a number of categories. For each dialogue state, a set of combinations of these categories was defined as a way to give meaningful sentences in the corresponding state. These combinations have been used for the inference of trigram models that employ the expansion on the defined categories. Additionally, all the task vocabulary was added in order to smooth the language model. These n-grams language models are in ARPA format and the lexicon model in HTK format. For this study we used only the language model from the first dialogue state, which covers all the consults that can be made to the system.

### 3.5 Test set

The test set was generated from 10 sentences related with the selected language model. These 10 sentences are shown in Figure 3. We got the cooperation of 16 different speakers who read the 10 test sentences, giving a total test set of 160 utterances per device (about 10 minutes of speech signal). The set of speakers had the following characteristics:

- They were between 25 and 39 years old.
- They were 3 women, and 13 men.
- 14 of them had higher education, and 2 speakers had secondary education.
- 13 speakers were Spanish, and the other 3 were foreigners.

```
Dime el camino a la biblioteca.
Estoy buscando la pizzeria ágora.
Sitúame la facultad de ciencias sociales y humanas.
¿Dónde está el baño?
Dime donde está el consejo social.
Indícame el camino a un aula docente.
El contenedor de vidrio.
Dime cómo llegar a la escuela de ciencias experimentales.
Quiero llegar a la óptica.
¿Dónde hay un aparcamiento de minusválidos?
```

**Fig. 3.** Test sentences.

**Table 2.** Reference results (WER).

| | Device | | |
|---|---|---|---|
| | Tablet | Smartphone | Computer |
| Without adaptation | 70.9% ± 2.6 | 43.6% ± 3.6 | 32.2% ± 3.1 |
| With adaptation | 63.4% ± 3.2 | 27.1% ± 2.5 | 25.9% ± 2.3 |

– 11 of them had Spanish as mother tongue, but 5 speakers had a different mother tongue (3 had Valencian, 1 had Arabic, and 1 had Finnish).

### 3.6 Speaker/device adaptation

The speaker/device adaptation was successfully used to improve the ASR quality in the UJI Smart Campus application [1]. In this application, significant improvements were achieved by using the MLLR (Maximum Likelihood Linear Regression) technique.

In the present study, we used a cross-validation approach to study the effect of speaker/device adaptation. Thus, we estimated two adapted models for each device and speaker by using HTK [7] with the MLLR technique using 32 regression classes with full transformation matrix. The first one was obtained using the first 5 sentences of the test set, and the second one was obtained using the last 5 sentences. Tests for each set of models were performed with the utterances not employed in the adaptation.

### 3.7 Evaluation metrics

As evaluation metric we used WER (Word Error Rate) between the best hypothesis from decoding processes and the reference. Confidence intervals of 95% were calculated by using the bootstrapping method with 10,000 repetitions [5].

## 4 Experiments

First of all, we obtained the WER reference values for the original signals with and without speaker/device adaptation. As can be observed in Table 2, the ASR quality degradation on mobile devices was confirmed by the reference results. Without adaptation, the WER in the case of the tablet reached 70.9% ± 2.6, and in the case of the smartphone, although the WER was smaller, it

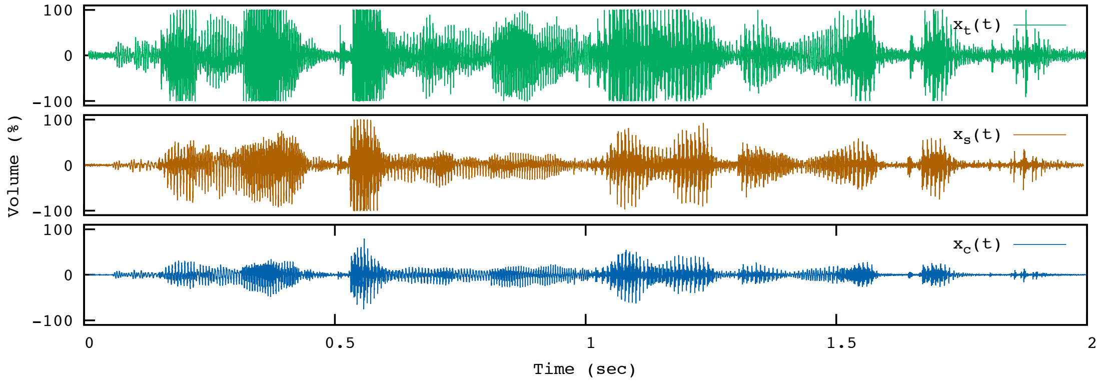


**Fig. 4.** Voice waveform examples.

was still high, remaining at 43.6%±3.6. In both cases, the differences were statistically significant when compared with the 32.2% ± 3.1 of WER obtained for the computer. The speaker/device adaptation produced significant improvements for all devices. The improvement achieved in the case of the tablet was significant, but it was still a high value of WER (63.4% ± 3.2). In the case of the smartphone the improvement stood out above the rest. With an improvement of 18.5 points of WER, the ASR quality for the smartphone sets in the ASR quality level obtained for the computer. To perform all the ASR experiments we used the iAtros recogniser [3].

The next step was to analyse the signals to find the reasons that cause these results. In Figure 4 the speech waveforms of the first sentence for one speaker are shown. The first thing that we can observe is the difference of gain levels and SNR. While for computer ($x_c(t)$) the signal has $38.8dB$ of SNR without saturation, in the worst case (tablet ($x_t(t)$)), we can observe a SNR of $25.3dB$. Moreover, in some points the signal becomes saturated. The same is happening in the case of the smartphone ($x_s(t)$), with less saturation and a better SNR ($33.4dB$). When we compared the spectrum of the signals, we observed a clear difference in the gain levels. Figure 5 presents the spectral analysis of the waveforms showed in Figure 4 using a Hamming window with a size of 512 samples. We used Audacity [8] to measure the SNR, and to do the waveform and frequency analysis.

From the analysis performed, we can conclude that the amplitude is the main difference between the signals. Consequently, we can guess that a modification in the gain of signals could cause improvements in the recognition quality.

### 4.1 Gain adjustment experiment

With a second experiment we checked the effect of modifying the gain from $20dB$ of attenuation to $10dB$ of amplification. Moreover, we checked the added effect of adapting the acoustic model to the speaker/device in two ways; first, by adapting the acoustic model before changing the gain (*Adaptation - Gain*), and second, by modifying the gain before adapting the acoustic model (*Gain - Adaptation*). The gain level modification was performed with the Sox tool [6].

As can be observed in the results shown in Figures 6, 7 and 8, the modification of gain level affects the quality of speech recognition. In the case of the computer (Figure 6), no significant improvement was obtained modifying the gain. However, in the case of the the tablet (Figure 7), the best results were obtained in the range of attenuation between $10dB$ and $20dB$, with

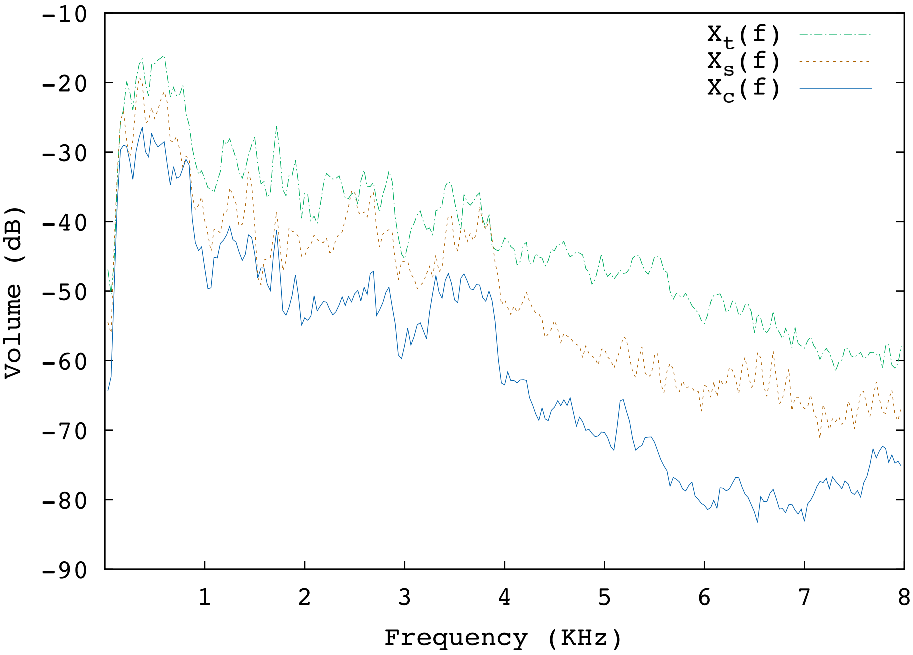


**Fig. 5.** Voice spectrum examples.

**Table 3.** Summary of experiment best results (WER).

| | Device | | |
|---|---|---|---|
| | Tablet | Smartphone | Computer |
| Adaptation - Gain | 44.0% ± 3.4 | 26.6% ± 2.6 | 25.9% ± 2.3 |
| Gain - Adaptation | 28.7% ± 2.7 | 26.5% ± 2.6 | 25.0% ± 2.2 |
| Without adaptation | 38.0% ± 3.0 | 32.9% ± 3.0 | 32.2% ± 3.1 |

38.0% ± 3.0 as best WER value when applying an attenuation of $17dB$ without adaptation. With adaptation, the best WER value was 28.7% ± 2.7 when applying an attenuation of $16dB$ before the acoustic adaptation, and 44.0% ± 3.4 when applying an attenuation of $14dB$ after the adaptation.

For the smartphone (Figure 8) the best results were obtained by applying an attenuation between $0dB$ and $15dB$; in particular, the best result without adaptation was 32.9% ± 3.0 with an attenuation of $8dB$. The effect of applying the adaptation resulted in a WER of 26.5% ± 2.6 when applying an attenuation of $14dB$ before the adaptation, and 26.6% ± 2.6 when applying an attenuation of $3dB$ after the adaptation. Table 3 shows a summary of the best results obtained.

## 5 Conclusions

In this paper we studied the ASR quality for three different client devices on a distributed ASR system. We verified that the ASR quality for mobile devices is lower than the ASR quality

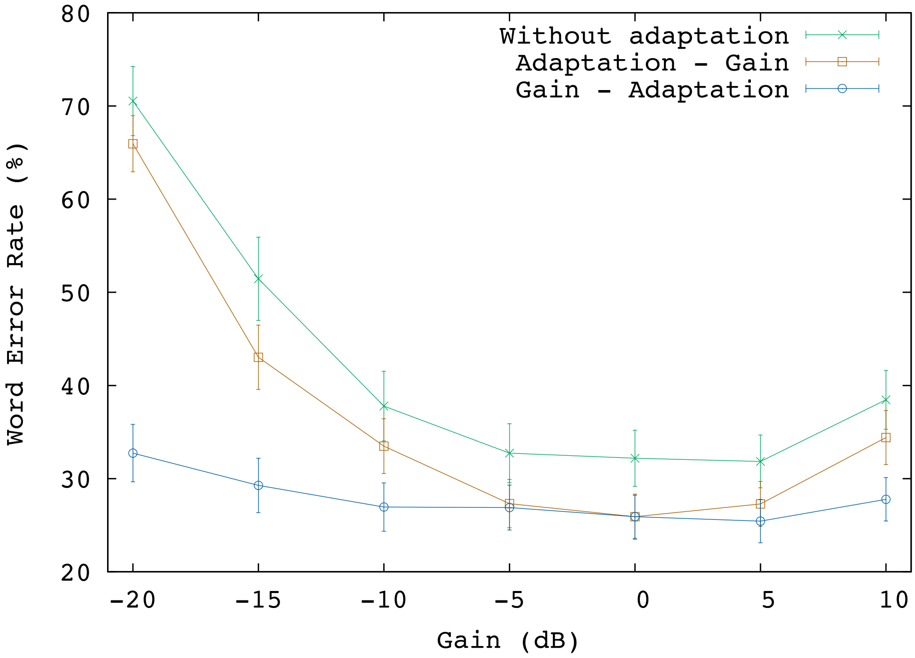


**Fig. 6.** Experiment results for the computer.

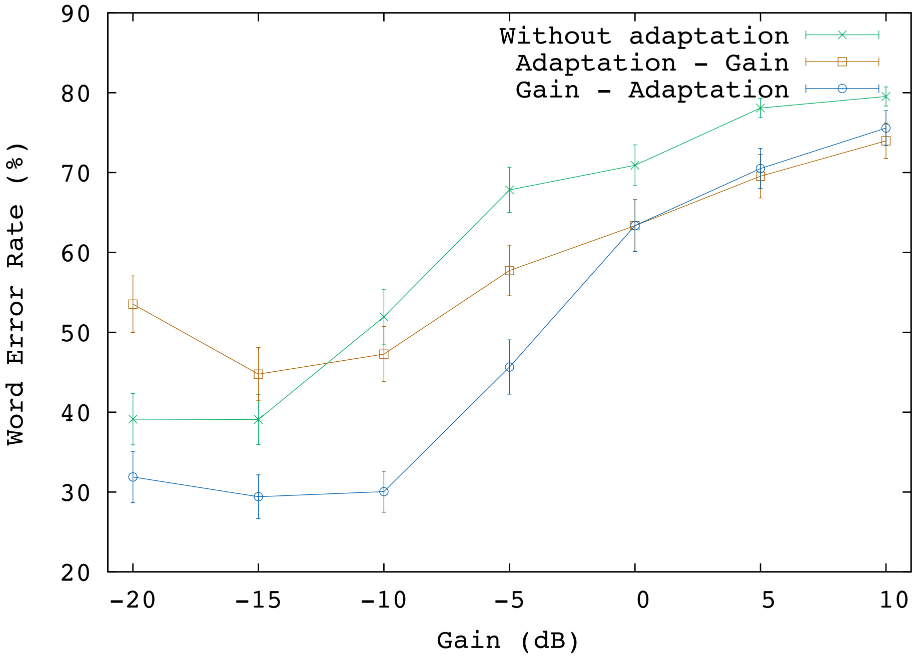


**Fig. 7.** Experiment results for the tablet.

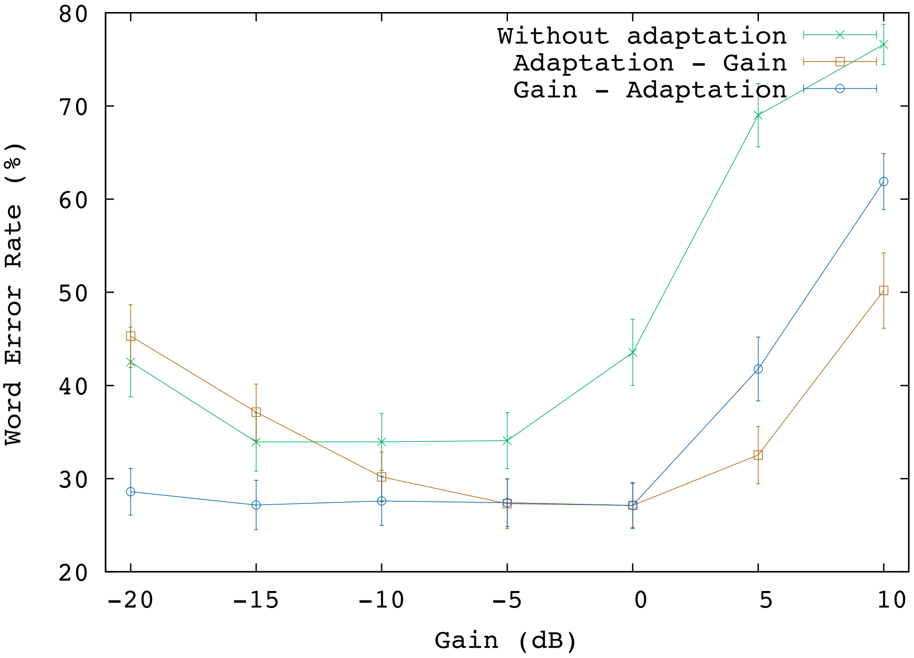


**Fig. 8.** Experiment results for the smartphone.

obtained by using a computer on the same conditions, and through several experiments, we managed to improve significantly the ASR quality for all devices.

The speaker/device adaptation of the acoustic model enhanced the ASR quality for all devices. Applying only the speaker/device adaptation we achieved a level of ASR quality for the smartphone similar to the level obtained for the computer. However, in the case of the tablet the improvement was not enough.

With the gain level adjustment, we achieved an improvement in the ASR quality for both mobile devices that reached the ASR quality level of the computer. In addition, the greatest improvement was obtained when we adapted the acoustic model to the speaker/device after the gain level adjustment.

We leave as an open line of future work the estimation of a device matched-filter. A matched-filter would involve a higher degree of ASR quality improvement.

A first practical application of this study could be used to improve the ASR quality on the UJI Smart Campus application. Currently, the system includes a speaker/device adaptation module that uses the first 5 sentences of the test set to calculate a global diagonal transformation matrix to adapt the acoustic model. Our proposal is to obtain the optimal gain value for the mobile device using the adaptation sentences before estimating the adaptation matrix, and notify this value to the speech recogniser or the mobile device. If this value is communicated to the speech recogniser, it should adjust the signal gain before preprocessing the signal to obtain the cepstral features. Another option is to communicate this value to the client device, in order to adjust the recording gain.

The second option presents several advantages, such as the prevention of the signal saturation and the reduction of the server computing time.

## References

1. E. Granell, C.D. Martínez-Hinarejos, G, Amat, J. Fernández, Á. Arranz, Á. Ramos, J.M. Benedí and A. Sanchis, "Speech Recognition on the *Percepción* Project", submitted to *iberSPEECH 2014*, Las Palmas de Gran Canaria, Spain, November 19-21 2014.
2. S. Stenzel and J. Freudenberger, "Blind-Matched Filtering for Speech Enhancement with Distributed Microphones", *Journal of Electrical and Computer Engineering*, vol. 2012, Article ID 169853, 15 pages, 2012.
3. M. Luján-Mares, V. Tamarit, V. Alabau, C.D. Martínez-Hinarejos, M. Gadea, A. Sanchis, and A.H. Toselli, "iATROS: A speech and handwritting recognition system", *V Jornadas en Tecnologías del Habla (VJTH2008)*, pp. 75–78, Bilbao, Spain. November 12-14, 2008.
4. A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus", *Proceedings of the Third European Conference on Speech Communication and Technology (EuroSpeech'93)*, pp. 175-178, Berlin, Germany, September 22-25, 1993.
5. M. Bisani, H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation", *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, vol. 1, pp. 409–412, Montreal, Canada, May 17-21, 2004.
6. SoX - Sound eXchange, `http://sox.sourceforge.net/Main/HomePage`, 2013.
7. 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 (for HTK version 3.4)", *Cambridge university engineering department*, 2009.
8. Audacity - Free Audio Editor and Recorder, `http://audacity.sourceforge.net`, 2014.

# Experiments on keyword spotting over the Transcrigal Database

Marta Martínez Maquieira and Antonio Cardenal[1]

Multimedia Technology Group, Universidade de Vigo,
ETSE Telecomunicación. Campus Lagoas de Marcosende 36210 Vigo, Pontevedra, Spain
{mmartinez,cardenal}@gts.uvigo.es
http://webs.uvigo.es/gtmuvigo/

**Abstract.** In this paper a set of experiments on keyword spotting over a broadcast news database is presented. The main goal of this work was to understand the advantages and limitations of the basic strategies for keyword spotting described in the literature. For this purpose, a realistic, but controlled scenario was selected: the Transcrigal database. This database is composed of recordings of Broadcast News programs of the local galician television. Using this experimental framework two basic strategies were implemented: direct word recognition, and continuous speech recognition. In both cases several different implementations were tested. We obtained best results using word lattices extracted from N-Best lists which confirms the results described in the literature.

## 1 Introduction

Keyword spotting (KWS) aims at detecting a given list of keywords in a set of speech recordings. It is an interesting task in several scenarios where the objective is to detect a closed set of words or expressions in an audio stream instead of getting an accurate transcription. Keyword spotting systems are used for indexing and retrieval of audiovisual databases or for media monitoring to mention a few examples.

In this paper we present the results of a series of experiments which were conducted to test the performance of several basic KWS algorithms using a controlled experimental framework. Our main objective was to understand the advantages, disadvantages and limitations of each method, and to contrast our conclusions with those found in the literature.

With these objectives in mind, we decided to use the Transcrigal database. This is a database that was compiled and labeled by our research group [4] several years ago. It is composed of recordings of broadcast news programs from the galician local television (TVG), including sport and weather sections. The programs are mainly in galician language, although there is a significative percentage of recordings in spanish language. During the last years, Transcrigal was the test-bed used for the development of our speech recognizer in Galician language. For this work the database was divided into three sets: train, development and test. For the development and test data the word error rate obtained was 32.9 and 29.6 respectively.

The most common approaches for KWS may be classified in three groups: word recognition systems, systems based on large vocabulary continuous speech recognizers (LVCSR) and systems based on phoneme lattices.

Word recognition is the most basic and straightforward method. The idea is to use a speech recognizer with a grammar composed of the desired keywords interconnected using some kind of filler models [6] [5]. To avoid false alarms, word recognition-based systems use some kind of garbage models to perform a statistical hypothesis test. The results of this test are employed as a confidence measure of the word hypothesis to select the desired false alarm/ false rejection tradeoff. Even with this mechanism, word recognizers are reported to have very high false alarm rates, so they are commonly used only in conjunction with more advanced algorithms.

The second method is based on large vocabulary continuous speech recognizers (LVCSR) [10]. The idea is to obtain a complete transcription as good as possible of the audio data. This transcription may be easily searched to find the occurrences of the desired keywords. The main drawback of this method is that the performance of LVCSR degrades heavily when there are acoustic and language mismatches between training and application conditions. Therefore systems based on LVCSR may be unreliable if there is no training data available. A second problem is that LVCSR systems use a closed vocabulary, so keywords not included in the lexicon during the training stage cannot be retrieved.

A more interesting variant of this method is to perform the search using word lattices rather than the raw output of decoder [7]. Word lattices are directed acyclic graphs, where each node represents a single word hypothesis. They are an efficient and compact method for representing a large amount of alternative transcriptions. Searching in word lattices is reported to achieve better accuracy than using the direct transcription [7] [9].

Finally, phoneme lattices [10] are most useful in tasks like spoken term detection in which fast response of the system is a fundamental requirement [2]. The idea is to build lattices of phonemes instead of words which may be efectively searched without a previous knowledge of the keywords. Phoneme lattices are, however, reported to achieve worse results than LVCSR-based systems [10].

In this work we explored the two first mechanisms of KWS described before: word recognizers and continuous speech recognizers, leaving the phoneme lattice for future research. We compared two methods of using LVCSR: direct search in the best transcription and in word lattices. As we will see in the following sections, word recognition offered the worst performance, with a high amount of false alarms for a reasonable rate of keyword detection which confirms the results described in the literature. On the other hand we found that word lattices was the method with better results, with only slightly better performance than direct search.

The remainder of the paper is organized as follows: Section 2 describes the experimental framework used in our experiments and describes the evaluation methods used. Section 3 explains in detail the different systems employed. Section 4 draws the results of the comparison of the different techniques used followed by the discussion of them. Finally Section 5 shows the conclusions extracted from the analysis.

## 2 Experimental Framework

### 2.1 Transcrigal Database

Our keyword spotting systems were tested on a large database of broadcast news from the local Galician television [4]. Each news lasts approximately 60 minutes, and consists of three well separated blocks: news, sports and weather forecast with its corresponding anchorperson.

The news are mainly in Galician language, but also there are non-reporters that speak in other languages as Spanish.

All speech data was digitized into 16 bits with a sampling frequency of 16 kHz. In total, there are near 39 hours of recordings. There are an average of 156 speaker turns for each news. Each speaker turn is divided into many files for which a voice activity detector is used.

For the purpose of the experiment, the database was divided into three independent partitions: train, development and test. Training data were used to obtain acoustic and language models. The development set was utilized for obtaining the optimal point of DET curve while the test set was used to obtain the final results. The composition of each partition may be consulted in Table 1.

To define the keyword set, we selected the most frequent words according to the language model. Therefore there are no out of vocabulary words in this test. We discarded function words to obtain a final list of 139 keywords. The total number of occurrences of keywords in the development and test sets is 1637 and 2268 respectively.

| | **Duration** | **#news** | **#files** |
|---|---|---|---|
| **Train** | 30h 27m 30.07s | 26 | 6152 |
| **Development** | 3h 32m 55.85s | 4 | 834 |
| **Test** | 5h 7m 5.88s | 5 | 1129 |

**Table 1.** Partitions of the Transcrigal database.

### 2.2 UVigo Speech recognizer

The experiments were performed using the UVigo decoder, developed in our research group [1]. To reduce the computation time we did not use in this re-

search our best performing system, which includes acoustic model adaptation and word-graph rescoring. Instead, a faster but less accurate one pass system was applied. The UVigo decoder uses the usual token-passing algorithm, with language model look-ahead and an N-Best re-scoring stage applying an A* stack strategy. The acoustic models are two state demiphones, with 12 Mel Frequency Cepstrum Coefficients plus energy and their delta and acceleration values. We used a bigram based language model (LM) with 60K words, 8M of bigrams and 6M trigrams approximately.
Table 2 shows the perplexity, out of vocabulary (OOV) rate and word error rate for the development and evaluation set.

| | Perplexity | OOV rate | WER |
|---|---|---|---|
| Dev | 184.2 | 2.6 | 32.9 |
| Eval | 218.7 | 2.2 | 29.6 |

**Table 2.** Perplexity, out of vocabulary rate and word error rate for the development and evaluation sets.

### 2.3 Evaluation measures

The performance of a keyword spotting system is measured by the *Receiver Operating Characteristics* (ROC) curve. In this case, it is a plot of the fraction of true positives (TPR) out of the actual positives versus the fraction of false positives (FPR) out of the actual positives. The definitions are:

$$\mathbf{TPR} = \frac{TP}{TP + FN} \tag{1}$$

$$\mathbf{FPR} = \frac{FP}{TP + FN} \tag{2}$$

**TP** is the number of true positives, **FN** represents the number of false negatives and **FP** is the number of false positives. Each point on the ROC curve represents the system efficiency for a specific trade-off between achieving a high true positive rate and a low false positive rate. Our experiments are evaluated using *Figure of Merit* (FOM)[12], which is the average of correct detections over 1 to 10 false alarms per hour. The FOM is calculated as follows:

$$\mathbf{FOM} = \frac{1}{10T} * (p_1 + p_2 + ... + p_N + a * p_{N+1}) \tag{3}$$

where $T$ is the total duration of the test speech in hours (in this case T is 1 hour) and $p_i$ represents the percentage of true hits found before the *i*'th false alarm. $N$ is the first integer greater or equal than *10T - 0,5* and *a= 10T - N* is a factor that interpolates to 10 false alarms per hour.

The system performance was also evaluated by a weighted function of miss and false alarm probabilities. For the NIST STD06 evaluation the primary metric was the *Actual Term Weighted Value* (ATWV). To score a posting list (a list with start, end times, and system detection scores) for a given keyword and detection threshold, entries in the list are matched to reference occurrences using a function that accounts for both temporal overlap between the reference and posting list ofs occurrences and the detection scores assigned by the system. The metric is defined as follows:

$$\mathbf{ATWV} = 1 - \frac{1}{T} * \sum_{t=1}^{T} (P_{miss}(t) + \beta * P_{fa}(t)) \tag{4}$$

$$P_{miss}(t) = 1 - \frac{N_{corr}(t)}{N_{true}(t)}, P_{fa}(t) = \frac{N_{spurious}(t)}{Total - N_{true}(t)} \tag{5}$$

$T$ is the total number of terms. $\beta$ represents the cost/value ratio, in this case is 0.1, thus the value lost by a false alarm is a tenth of value lost for a miss. $N_{corr}$ and $N_{spurious}$ is the number of correct and incorrect term detections,$N_{true}$ is the total number of true term occurrences in the corpus and $Total$ is the duration (in seconds) of the indexed audio corpus.

## 3 Keyword Spotting algorithms

### 3.1 Word recognition

This is the most simple and straightforward method for KWS. A simple word recognizer with a grammar which includes all the needed keywords is applied to the audio stream. The basic procedure is depicted in Figure 1. The top network represents the recognition grammar, composed of a set of N keywords in parallel. Each keyword may be modelled using a single Hidden Markov Model (HMM), or as a sequence of HMMs of subword units [5]. The bottom network of the figure represents the hypothesis test mentioned in Section 1 [9].

In our experiments, the hypothesis test was built using a phonetic recognition of the audio stream using the same acoustic models employed in the main recognizer. For a given word hypothesis $w_i$ with acoustic score $P_w$, we obtained the contrast score $P_a$ adding the scores provided by the phonetic recognizer over the same time period. The final acoustic confidence measure was obtained using the classical expression:

$$ACM = \frac{P_w - P_a}{T} \tag{6}$$

where $T$ is the duration in frames of the word $w_i$.

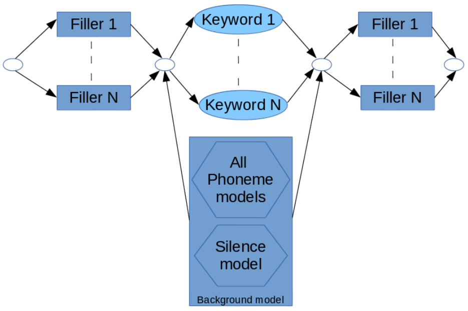


**Fig. 1.** General scheme of a keyword detection system using word recognition.

### 3.2 Continuous speech recognition

Keyword spotting using LVCSR may be easily implementing for searching for the desired keyword on the transcription. The main problem with this method is that there is no possibility of selecting an operation point to minimize false alarms. This is a minor problem if the recognizer is highly reliable, but unfortunately, this is not the usual case. Continuous speech recognizers are known to suffer severe losses in performance when there are mismatches between the training and application conditions. In the experiments described here we explored different possibilities for obtaining a ROC curve and selecting an operation point using the direct transcription. The problem is to identify the reliability of each word hypothesis, or in other words, to find a word confidence measure. This is a well-known problem for which several solutions have been proposed (see for example [3]). The most successful, according to most authors, is to use posterior probabilites computed from word lattices or N-Best lists [11], a strategy which will be described in next section.

In the experiments described in this section we used two different mechanisms to obtain the confidence measures. The first one was to use the same procedure described by Equation 6, using only the acoustic probabilities of the words. In the second experiment we used all the scores given by the decoder, i.e. the combination of the acoustic and LM scores computed as follows:

$$LCM^w = P_a^w + LMW \cdot P_{LM}^w \tag{7}$$

where $P_a^W$ is the acoustic score of the word $w$, $P_{LM}^w$ is the LM score and $LMW$ is the language model weight.

### 3.3 Continuous speech recognition with word lattices

In this experiment, we used the UVigo decoder in conjunction with the lattice tool provided by the SRI-LM toolkit [8] to obtain word lattices with posterior probabilities. We explored two different procedures: to generate the word lattices directly from the decoder; and to extract NBest lists and convert them to word lattices using the SRI-LM toolkit. In both cases the posterior probabilities were employed for plotting the ROC curves and finding the optimum operation point for the development set.

Due to the huge size of the word lattices obtained with the first method, it is necessary to use a pruning level to discard word hypothesis with low posterior probability. As we will see in section 4 we found this pruning level to have an important influence on the system which will be discussed later.

The second strategy used was to build the word lattices from the list of the N best hypothesis generated by the decoder. We used an upper limit of 200 word transcriptions for each audio utterance. In this case, the size of the resulting word lattices made unnecessary the use of posterior pruning.

## 4 Experimental Results

In this section the experimental results for all KWS methods described in section 3 are presented. The experimental procedure was as follows: all experiments were performed over the development set. We selected an operation point to optimize the results in terms of FOM and another one to maximice ATWV. Both operation points were applied to the evaluation set to obtain the final results presented at the end of this section.

Figure 2 shows the results for the experiment using LVCSR. Label *KWS-CRAM* identifies the experiment using the hypothesis test of Equation 6, while *KWS-CRLM LMW* corresponds to the experiments using the sum of the logaritmic acoustic and LM scores and an LM weight of *LMW* (Equation 7).

It can be seen in the figure that *KWS-CRAM* has better performance than *KWS-CRLM* for any LM weight. This is an unexpected result. Confidence measures based on acoustic scores only are known to give poor perfomance in continuous speech recognition tasks. Therefore we expected better performance for KWS also. We think that these results may be somehow biased due to method of selection of the keywords explained in section 2, but further research will be needed to clarify this point.

Figure 3 shows the results of the KWS systems using word lattices. The label *KWS-WL PP* identifies the experiments using word lattices extracted directly from the decoder using a posterior pruning level of *PP*, while *KWS-WLNB* corresponds to the experiment using word lattices extracted from N-Bests lists. We obtained much better results using N-Best lists than using word lattices. Note the effect of the pruning level over the *KWS-WL* experiments. A higher pruning level produces ROC curves with more extension over the x axis, but with lower TPR for a given FPR.

Finally, Figure 4 is a compilation of the best results of Figures 2 and 3. The curve with legend *KWS-WR* was included for comparison purposes and was extracted using the experiment with word recognition. Note that the performance

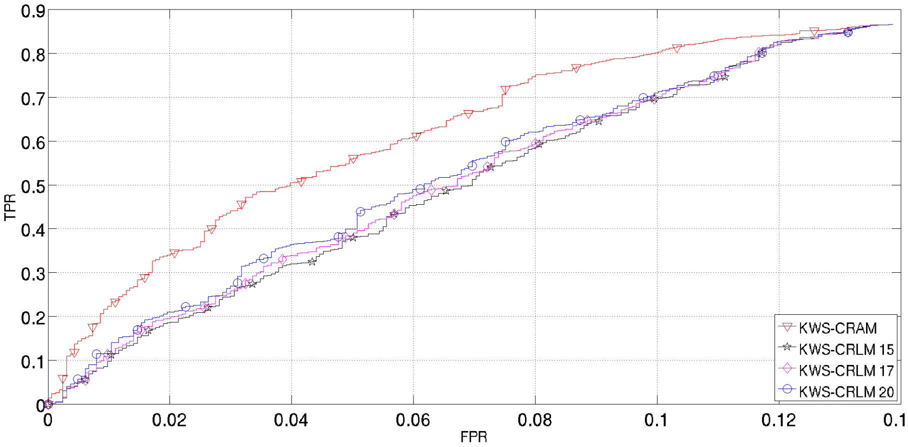


**Fig. 2.** ROC curve for the experiments with LVCSR. **KWS-CRAM**: results using acoustic confidence measures; **KWS-CRLM LMW**: results using acoustic plus language model scores as confidence measures, with LM weight=LMW.

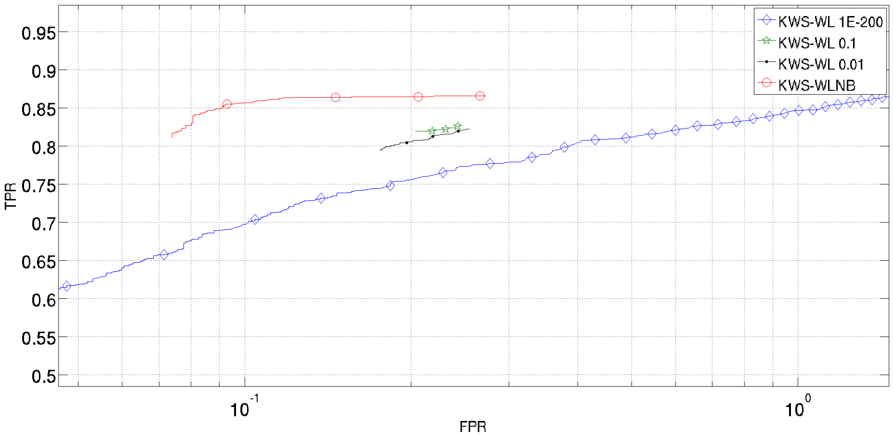


**Fig. 3.** ROC curve for the experiments using word lattices. **KWS-WL PP**: results using word lattices with a posterior pruning of PP; **KWS-WLNB**: results using word lattices extracted from the N-Best lists.

of this method is, as expected, very bad. It can be seen in the figure that the *KWS-WLNB* (word lattices with N-Best) has better TPR than *KWS-CRAM* (LVCSR with acoustic confidence measures) for all values of FPR, but the latter method lets the user select an operating point with a less false positive rate.

Table 3 shows the results, in terms of ATWV and FOM, obtained over the evaluation set, for the three experiments depicted in Figure 4. The results were

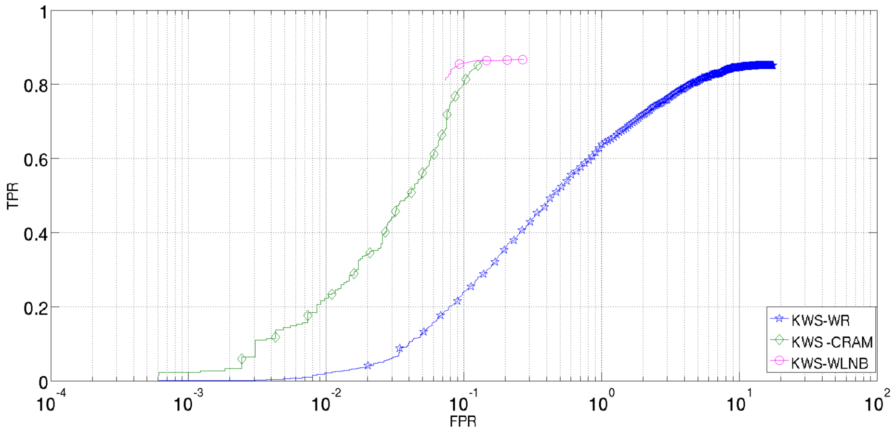


**Fig. 4.** Comparison of the ROC curves for the best systems of Figures 2 and 3 . **KWS-WR** is the curve with the experiment for word recognition.

obtained using the operation point which maximizes ATWV and FOM in the development set. The system *KWS-WLNB* offer better results than *KWS-CRAM* in terms of ATWV also for this data set. However, *KWS-CRAM* performs better in terms of FOM. We think that this effect is due to the fact that in *KWS-WLNB* operation points with less false positive rate than in *KWS-CRAM* may be selected.

| System | ATWV | $P_{miss}$ | $P_{fa}$ | FOM | Computational time |
|---|---|---|---|---|---|
| **KWS-WR** | -0.98 | 0.148 | 0.995% | 26.06 | 6.83 |
| **KWS-CRAM** | 0.75 | 0.115 | 0.022% | 52.73 | 2.03 |
| **KWS-WLNB** | 0.78 | 0.118 | 0.018% | 46.49 | 2.05 |

**Table 3.** ATWV, miss and false alarm probabilities, FOM results on the test data and computational time (measured in real time factor). The experiments was executed in a server with 2xIntel Xeon, 64 bits, 8cores, 2,33GHz, 8GB RAM.

## 5 Conclusions and further work

In this paper we have described a set of initial experiments of keyword spotting over Transcrigal database. Our objective was to identify the advantages and drawbacks of the basic KWS methods.

We have confirmed most of the results that can be found in the literature: the methods based on word recognition offered poor performance, with high false positive rates; the strategies that employ word lattices and posterior probabilities are in general better than those which use the direct transcription provided by the recognizer.

On the other side, the superiority of the acoustic confidence measure, over the use of acoustic and LM probabilities for the continuous speech recognition case, was an unexpected result which deserves more research.

Finally, future work includes the comparison of these algorithms with the method of phoneme lattices and the combination of KWS strategies which, according to several authors [9] may yield to important improvements.

## 6 Acknowledgements

This work has been supported by the European Regional Development Fund, the Galician Regional Government (CN2011/019, "Consolidation of Research Units: AtlantTic Project" CN2012/160), the Spanish Government ("SpeechTech4All Project" TEC2012-38939-C03-01) and the "TecAnDAli Research Network" of the "Consellería de Educación e Ordenación Universitaria, Xunta de Galicia".

## References

1. Docío-Fernández, L., Cardenal-López, A., García-Mateo, C.: TC-STAR 2006 automatic speech recognition evaluation: The uvigo system. In: TC-STAR Workshop on Speech-to-Speech Translation. pp. 145–150 (2006)
2. Fiscus, J., Ajot, J., Garofolo, J., Doddingtion, G.: Results of the 2006 spoken term detection evaluation. In: Proc. SIGIR (2007)
3. Jiang, H.: Confidence measures for speech recognition: A survey. Speech Communication 45, 455 – 470 (2005)
4. Mateo, C.G., Tirado, J.D., Fernández, L.D., Cardenal, A.: Transcrigal: A bilingual system for automatic indexing of broadcast news. In: IV International Conference on Language Resources and Evaluation (2004)
5. Rohlicek, J., Russell, W.: Continuous hidden Markov modeling for speaker-independent word spotting. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 627–630 (1989)
6. Rose, R., Paul, D.: A hidden Markov model based keyword recognition system. In: International Conference on Acoustics, Speech, and Signal Processing. pp. 129–132 (1990)
7. Seide, F., Thambiratnam, K., Peng Yu, R.: Word-lattice based spoken-document indexing with standard text indexers. In: 2008 IEEE Spoken Language Technology Workshop. pp. 293–296 (Dec 2008)
8. Stolcke, A.: Srilm - an extensible language modeling toolkit. In: Proceedings International Conference on Spoken Language Processing. pp. 901–904 (2002)
9. Szoke, I., Motlicek, P., Valente, F.: Improving acoustic based keyword spotting using LVCSR lattices. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 4413–4416 (2012)
10. Szöke, I., Schwarz, P., Matejka, P., Burget, L.: Comparison of keyword spotting approaches for informal continuous speech. In: Interspeech. pp. 633–636 (2005)
11. Wessel, F., Schlüter, R., Macherey, K., Ney, H.: Confidence measures for large vocabulary continuous speech recognition. IEEE Transactions on Speech and Audio Processing 9, 288–298 (2001)
12. Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.4. Cambridge University Press (2006)

# Dynamic Topic-Based Adaptation of Language Models: A Comparison Between Different Approaches

Julian Echeverry-Correa, Beatriz Martínez-González, Rubén San-Segundo,
Ricardo Córdoba, and Javier Ferreiros-López

Speech Technology Group, Universidad Politécnica de Madrid, Spain
{jdec,beatrizmartinez,lapiz,cordoba,jfl}@die.upm.es

**Abstract.** This paper presents a dynamic LM adaptation based on the topic that has been identified on a speech segment. We use LSA and the given topic labels in the training dataset to obtain and use the topic models. We propose a dynamic language model adaptation to improve the recognition performance in 'a two stages' ASR system. The final stage makes use of the topic identification with two variants: the first one uses just the most probable topic and the other one depends on the relative distances of the topics that have been identified. We perform the adaptation of the LM as a linear interpolation between a background model and topic-based LM. The interpolation weight is dynamically adapted according to different parameters. The proposed method is evaluated on the Spanish partition of the EPPS speech database. We achieved a relative reduction in WER of 11.13% over the baseline system which uses a single background LM.

**Keywords:** language model adaptation, topic identification, automatic speech recognition, information retrieval

## 1 Introduction

The performance of a speech recognition system depends significantly on the similarity between the language model (LM) used by the system and the context of the speech that is being recognized. This similarity is even more important in scenarios where the statistical properties of the language fluctuates throughout the time, for instance, in application domains involving spontaneous speech from multiple speakers on different topics. One representative example of this kind of domain is the automatic transcription of *political speeches*. Within this domain, the usage of content words (i.e. those that convey information and have a specific meaning rather than indicating a syntactic function) depends on several factors, such as the topic the speaker is addressing, the style of the speech, the vocabulary used by the speaker and the scenario in which the speech is taking place. Regarding these factors, in this paper we are focusing on studying the identification of the topic and its application in the adaptation of language models. The performance of the speech recognition system will depend, among other elements, on its capacity to update or dynamically adapt the LMs. In this paper we propose a dynamic LM adaptation based on an information retrieval (IR) approach. We used IR techniques for identifying the topics that are related to the content of the speech segment under evaluation. This information enables the system to perform an adaptation of the language model. We explore different approaches for the dynamic language model adaptation. These approaches are based on the interpolation between a background model and topic-based LMs.

## 2 General Overview

In this paper two major tasks can be distinguished: **topic identification** and **dynamic LM adaptation**. Both tasks pursue one common goal, that is improving the performance of an automatic speech recognition system for multitopic speech. We integrate these tasks in 'a two stages' ASR framework. In the first stage, an initial speech recognition of an audio segment is performed using a background LM built from the entire training set. Then, an IR module automatically identifies the topic based on the results of the initial recognition pass. This module uses topic models that have been previously trained for each of the topics available in the database. Using the information provided by the topic identification system and topic-specific LMs, a dynamic adaptation of the background LM is performed. In this paper we present different approaches for the dynamic adaptation of LMs. In the final stage of the framework, the adapted LM is used to re-decode the utterance.

## 3 Related Work

The task of topic identification (TI) falls at the intersection of information retrieval and machine learning systems. In the last years a growing number of statistical learning methods have been applied in TI from these research fields [1]. Common approaches includes Latent Semantic Analysis [2], Rocchio's method [3], Decision Trees [4] and Support Vector Machines [5]. TI has been successfully applied in many contexts and disciplines, ranging from topic detection [6], automated metadata generation [7], document and messages filtering [8] and the recently developed area, sentiment analysis [9], among many other fields of application. Nevertheless it is interesting to review the influence of TI in the field of language model adaptation. Within this field, TI has been used to study the changes that the language experiences when moving towards different domains [10]. In that sense, TI is able to contribute to LM adaptation by adding new sources of information to previously existent models with the objective of enriching them. This leads to a diversity of approaches in the field of LM adaptation that can be distinguished regarding the origin of the new sources of information. Some LM adaptation approaches are based on the specific context of the task that they are addressing. In these approaches, the new data is used to generate a context-dependent LM which is then merged with a static LM. These new sources of information can proceed, for instance, from text categorization systems as in [11], from speaker identification systems [12], from linguistic analysis systems [13] or from the application context itself [14]. Other approaches are based on analysis and extraction of semantic information. Latent Semantic Analysis (LSA) is an example of this type of approach. In [15], the use of LSA is proposed to extract the semantic relationships between the terms that appear in a document and the document itself. More robust techniques in the field of information retrieval, as Latent Dirichlet Allocation (LDA) [16], have also been used for adapting LMs [17]. When using data available online it is possible to find information related to a large variety of topics. In this regard, clustering algorithms have been proposed to group together those elements that share some properties. Topic-based language modeling is an example of this clustering criterion [18, 19].

## 4 Topic Identification

In a broad sense, topic identification is the task of automatically identifying which of a set of predefined topics are present in a document. To perform topic identification some steps must be followed. These steps are: preprocessing, document representation, term weighting and topic modeling and identification.

### 4.1 Preprocessing

The preprocessing stage allows us to to convert both, documents and queries, to a more precise and concise format. This stage has a substantial impact on the success of the topic identification process [20]. Typical preprocessing steps include: structural processing, lexical analysis, tokenization, stopwords removal, stemming and term categorization. We provide a small description of the steps in which we made special considerations:

• *Stopwords removal.* There are several stopwords lists available online for different languages and for general applications in IR systems. However, generic stopwords lists do not contemplate terms, that in fact, are very frequent in domain specific documents. For that reason, we performed the evaluation using two lists: a generic list with 421 stopwords (*List-1*) and a domain specific stopword list (*List-2*), that we created by adding, to the generic list, those terms with an Inverse Document Frequency (IDF) value below a threshold. The IDF measures how common a term is in the whole document collection; we computed it by using a Term-Document Matrix composed of the 1802 documents in the training dataset and the 16528 terms in the word inventory. We performed different experiments on the Development set in order to find the optimal threshold. The lowest topic identification error on this dataset was obtained by setting the threshold to 0.4435, which means removing the terms that appear, at least, in 649 documents. The List-2 has 446 stopwords.

• *Stemming.* This step refers to the transformation of a word to its stem or root form. For this step, we have used the Freeling Toolkit [21]. Due to few errors in the original stemming process, we have modified some of the stemming rules for the Spanish language of the toolkit.

### 4.2 Document Representation

The document representation is based on the widely known bag-of-words model. In this model the relationships between the index-terms and each of the documents in the collection are represented by a Term-Document Matrix, that describes the frequency of occurrence of the index-terms in the documents.

### 4.3 Term Weighting

To improve the capacity of discrimination of the index-terms, weights can be applied to the elements of the Term-Document Matrix by associating the occurrence of an index-term with a weight that represents its relevance with respect to the topic of the document. We have selected the combination of TF (*Term Frequency*) and IDF as the baseline weighting scheme for comparing the results obtained for the topic identification task in this paper. Among the most common weighting schemes, *term entropy* (*te*) is based on an information theory approach and it exploits the distribution of terms over documents [22]. For the index-term $t_i$ in the document $d_j$, it is defined as follows:

$$te_{i,j} = 1 - \sum_{j=1}^{N} \frac{p_{i,j} \cdot \log(p_{i,j})}{\log(N)}, \text{ where } p_{i,j} = \frac{c_{i,j}}{gf_i} \tag{1}$$

Where $c_{i,j}$ represents the term frequency of the index-term $t_i$ in the document $d_j$. $gf_i$ is the global frequency of the index-term $t_i$ measured over the $N$ documents in the collection. This scheme may lead to a log zero calculation if an index-term is not present in a document. It has been suggested to include a smoothing parameter $a$, resulting in $p_{i,j} = (a + c_{i,j})/gf_i$. Indeed, it solves the log zero calculation, but the evaluation that we have performed on the combination of TF and this scheme has shown that it does not significantly improve the TF-IDF baseline weighting scheme. We propose a *pseudo term entropy* calculation based on the *term entropy* formula. Our idea is to assign less weight to the terms that are equally distributed over the documents in the collection and assign more weight to terms that are concentrated in a few documents. In this *pseudo term entropy* the parameter $p_{i,j}$ is calculated as the weighted sum of $c_{i,j}$ and the inverse of $gf_i$.

$$p_{i,j} = \beta \cdot c_{i,j} + \frac{\gamma}{gf_i} \tag{2}$$

The proposed scheme not only solves the log zero problem, but also improves the topic identification accuracy as shown in section 6. We performed different experiments on the Development set in order to adjust the parameters $\beta$ and $\gamma$. For the evaluation proposed in this paper, the best results were obtained by adjusting $\beta = 1.5$ and $\gamma = 2.1$.

### 4.4 Topic Models

In this paper we compare two topic models: the Generalized Vector Model (GVM) and Latent Semantic Analysis (LSA) [2]. Both models represent documents and queries as vectors in a multi-dimensional space, in which the number of dimensions is determined by the number of index-terms in the GVM or the number of latent dimensions in the LSA approach.

In both models, the similarity $sim(\boldsymbol{d}, \boldsymbol{q})$ between a document $\boldsymbol{d}$ and a query $\boldsymbol{q}$ can be computed using the cosine distance. According to this distance, each document is ranked on how close it is to the query. In our approach, we have gathered all documents in the collection belonging to the same topic into one document. We have done the same for all the topics. By doing this, each document represents a distinct topic. So, when computing the similarity between the query and a document, we are actually computing the similarity between the query and a topic.

## 5 Topic-based Language Model Adaptation

Topic-based LM adaptation becomes a strategy to lower the word error rate of the transcription given by the ASR by providing language models with a higher expectation of words and word-sequences that are typically found in the topic or topics of the story that is being analyzed. LM interpolation is a simple and widely used method for combining and adapting language models [23, 24].

### 5.1 Language Model Interpolation

Let us consider probabilistic language modeling and let $P(w|h)$ be the probability of word $w$ given the previous sequence of words $h$. Then, given a background model $P_B(w|h)$ and a topic-based model $P_T(w|h)$ it is possible to obtain a final model $P_I(w|h)$, to be used in the second decoding pass, as

$$P_I(w|h) = (1 - \lambda)P_B(w|h) + \lambda P_T(w|h) \tag{3}$$

where $\lambda$ is the interpolation weight between both models, which has to fulfill the condition $0 \leq \lambda \leq 1$. The topic-based LM is generated by combining several topic-specific LMs $P_t(w|h)$ in general. In our case, the background model, as well as the topic-specific models are static models. They are trained once and remain unchanged during the evaluation. The topic-based LM could be either static or dynamic. It depends on the adaptation scheme followed, as we will see later in this paper. This model, as well as the final model $P_I(w|h)$, are generated during the evaluation of each audio segment.

### 5.2 Interpolation Schemes

Two questions arise at this point. How to generate the topic-based model $P_T(w|h)$? and, how to determine the interpolation weight $\lambda$ with the background model? For solving these questions, we propose different approaches:

- **Hard approach.** In this approach, the topic-based LM $P_T(w|h)$ is built by considering only one of the topic-specific language models ($P_t(w|h)$). This model is selected as the one related to the topic ranked in the first position by the TI system. For estimating the interpolation weight $\lambda$ we define a distance measure $\delta$ between this LM and the background LM. In this approach, our hypothesis is that the greater the distance between both models, the greater the contribution of the topic specific model to the final one. This distance is computed by considering the average difference in the unigram probabilities of both models.

$$\delta_T = \frac{1}{N} \sum_{\forall w_i \in P_T} |P_T(w_i) - P_B(w_i)| \tag{4}$$

Where $N$ is the number of unigrams in the topic-based LM $P_T(w|h)$. To ensure the interpolation weight fulfills the condition $0 \leq \lambda \leq 1$, we include the summation of the distances of all the topic-specific LMs to the background model as a normalization constant. Then, the interpolation weight is computed as the relative distance between $\delta_T$ and this normalization constant.

$$\lambda = \frac{\delta_T}{\sum_{j=1}^{n} \delta_j} \tag{5}$$

Where $n$ is the number of topics and $\delta_j$ the distance of the *j-th* topic-specific LM to the background LM.

- **Soft-1 approach.** In this case, instead of using only one specific-topic LM for generating the topic-based LM, this model is built on a dynamic basis by the interpolation of a different number of topic-specific LMs. The **Soft-1 approach** tries to gather the dynamic of the specific-topic models $P_t(w|h)$ depending on the similarity of the audio segment to each of the topics. By doing this, more relevance is given to the topics ranked in the first positions by the TI system. The topic-based LM is then computed as follows:

$$P_T(w|h) = \alpha_1 P_{t_1}(w|h) + \alpha_2 P_{t_2}(w|h) + \cdots + \alpha_k P_{t_k}(w|h) \tag{6}$$

where $k$ is the number of models considered for obtaining the topic-based LM. The interpolation weight $\alpha_i$ is calculated as the normalized value of the similarity measure of the TI system.

$$\alpha_i = \frac{sim(\boldsymbol{d}_i, \boldsymbol{q})}{\sum_{j=1}^{k} sim(\boldsymbol{d}_j, \boldsymbol{q})} \tag{7}$$

The interpolation weight $\lambda$ between the background LM and the topic-based LM was set experimentally in this case.

- **Soft-2 approach.** This approach is similar to the previous one, but instead of setting $\lambda$ experimentally, we have computed it by weighting the relevance of the topic-specific LMs according to the cosine distance. That is:

$$\lambda = \sum_{i=1}^{k} \frac{sim(\boldsymbol{d}_i, \boldsymbol{q})}{\sum_{j=1}^{k} sim(\boldsymbol{d}_j, \boldsymbol{q})} \cdot \frac{\delta_i}{\sum_{j=1}^{k} \delta_j} \tag{8}$$

In Soft-1 and Soft-2 approaches, we have considered two additional possibilities: a) to create the topic-based LM using all the topic-specific LMs, that is by setting $k$ as the total number of topics, and b) to create the topic-based LM by selecting the 10 topics with higher positions in the TI ranking.

# 6 Experimental Evaluation

Our evaluation focuses in two aspects: the evaluation of the topic identification approach and the evaluation of the dynamic language model adaptation by means of evaluating the performance of the speech recognition system. Before discussing the results obtained, we describe the dataset used for the evaluation.

## 6.1 Dataset

We have used the Spanish partition of the EPPS Database (*European Parliament Plenary Sessions*) of the TC-STAR Project to study the performance of the proposed system. Due to the fact that the training dataset of the database is the only one that includes distinct labels for the topics, we used it for training, development and evaluation purposes. The topics have been manually labeled according to the main discussion subject of each session [25]. We believe that identifying the topic on short sentences can be ambiguous because few words do not provide semantic information about the topic that is being addressed. For that reason we decided to perform the evaluation over segments of audio with a length no less than a minute. We extracted these segments from turns of intervention of just one speaker. By this criterion, we obtained 252 audio segments for the evaluation. Some details of the corpus: The language of the corpus is Spanish. There are both male and female speakers (approx. 75% - 25% distributed). The domain of the corpus is political speeches. Training set is composed of 21127 sentences grouped in 1802 speaker turns of intervention. Development set is composed of 2402 sentences grouped in 106 speaker turns. The lexicon size is 16.5k words and the Test set is composed of 3738 sentences grouped in 252 speaker interventions. Each of the speaker interventions belongs to one of 67 different topics. We also use the EUROPARL [26] text database for training both background and topic-specific LMs.

## 6.2 Topic Identification Evaluation

For the topic identification task, the initial performance of the system was obtained by using the Generalized Vector Model, a classic TF-IDF weighting scheme and a general domain stopwords list (SW *List-1*). We will use this configuration as the baseline to discuss the improvements in the different approaches that we have applied. We compared the two different lists of stopwords. We also compared different weighting schemes and the influence of preprocessing stages like stemming in the topic identification error. Table 1 shows the results obtained in topic identification using both GVM and LSA approaches.

**Table 1.** Topic Identification error (T.ID. error) using GVM and LSA topic models approaches

| Topic identification approach | T.ID. error (%) |
|---|---|
| GVM + TF-IDF + SW (*List-1*) | 35.32 ± 5.90 |
| GVM + TF-IDF + SW (*List-2*) | 34.13 ± 5.85 |
| GVM + TF-IDF + SW (*List-2*) + Stemming | 36.11 ± 5.93 |
| GVM + TF-Entropy + SW (*List-2*) | 34.52 ± 5.87 |
| GVM + TF-PseudoEntropy + SW (*List-2*) | 33.33 ± 5.82 |
| LSA + TF-IDF + SW (*List-1*) | 32.94 ± 5.80 |
| LSA + TF-IDF + SW (*List-2*) | 30.95 ± 5.70 |
| LSA + TF-IDF + SW (*List-2*) + Stemming | 32.14 ± 5.76 |
| LSA + TF-Entropy + SW (*List-2*) | 29.76 ± 5.64 |
| **LSA + TF-PseudoEntropy + SW (*List-2*)** | **27.38 ± 5.50** |

In general, LSA outperforms the Generalized Vector Model. In both topic models, the combination of TF and *pseudo term entropy* (TF-PseudoEntropy) reduces the topic identification error when compared to TF-*entropy* and to TF-IDF weighting schemes, nevertheless this reduction is not statistically significant. For both models, Stemming does not significantly contribute in error reduction. The criterion that we followed for creating the *List-2* of stopwords contributes in most of the cases in reducing topic identification error. The best combination of parameters is obtained for the LSA model, using the *List-2* of stopwords and weighting the terms with TF-PseudoEntropy scheme. This configuration presents a relative improvement of 22.48% when compared to the baseline approach.

### 6.3 Dynamic LM Evaluation

For the evaluation of the dynamic LM adaptation we have used the best configuration of parameters obtained in the previous section. The initial performance of our baseline system (i.e. without the dynamic LM adaptation) achieved a WER of 21.75. In Table 2 the results of the speech recognition performance when using the proposed approaches for the dynamic LM adaptation are compared. Although there is no significant difference between the **Soft-1** and the **Soft-2** approaches when comparing both variants (all topics and top-10), there is, in fact, a significant difference between the results obtained by the **Soft-1 - top 10** and the **Hard** approach, and even better results can be found when compared to the baseline approach. In general, with this soft integration we manage to reduce 11.13% of the initial WER.

## 7 Conclusions

In this paper we have presented a framework for dynamic language model adaptation based on topic identification. The results in the ASR task have

**Table 2.** Comparison between the word error rate obtained for different LM adaptation approaches

| LM Adaptation approach | WER | Relative Improvement |
|---|---|---|
| Baseline (no adaptation) | 21.75 ± 0.26 | |
| Hard | 19.90 ± 0.25 | 8.51 |
| Soft 1 - all | 19.61 ± 0.25 | 9.84 |
| Soft 1 - top 10 | **19.33 ± 0.25** | **11.13** |
| Soft 2 - all | 19.65 ± 0.25 | 9.66 |
| Soft 2 - top 10 | 19.50 ± 0.25 | 10.34 |

shown that a small but statistically significant improvement in word error rate can be obtained by the adaptation strategy that has been proposed. Adapting the LM by taking only into consideration the closest topic, improves the baseline performance, but does not take advantage of all the sources of information available. The proposed criterion for selecting stopwords and the proposed weighting scheme have contributed in reducing the topic identification error.

**Acknowledgements.** The work leading to these results has received funding from TIMPANO (TIN2011-28169-C05-03), and MA2VICMR (Comunidad Autonóma de Madrid, S2009/TIC-1542) projects. It has also been supported by the European Union under grant agreement number 287678. Julian Echeverry-Correa also acknowledges the support from Colciencias and from the Universidad Tecnológica de Pereira, Colombia.

## References

1. Sebastiani, F. “Machine learning in automated text categorization”. ACM Computing Surveys (CSUR), 34(1):1–47, 2002.
2. Deerwester, S. et al. “Indexing by latent semantic analysis”. Journal of the American Society for Information Science, 41 (6): 391–407, 1990.
3. Rocchio, J. “Relevance Feedback in Information Retrieval”, in G. Salton [Ed], The SMART retrieval system: experiments in automatic document processing. Prentice-Hall, Inc., 1971.
4. Lewis, D., and Ringuette, M. “A comparison of two learning algorithms for text categorization”. In Proc. of 1994 Symposium on Document Analysis and Information Retrieval, pages 81–93. 1994.
5. Joachims, T. “Text categorization with Support Vector Machines: Learning with many relevant features”. Machine Learning: ECML-98. Springer Berlin Heidelberg, pages 137–142. 1998.
6. Qiu, Y. “A keyword based strategy for spam topic discovery from the internet”. In Proc. of 2010 Fourth International Conference on Genetic and Evolutionary Computing (ICGEC)., pages 260–263. 2010.
7. Cheng, N. and Chandramouli, R. and Subbalakshmi, K.P. “Author gender identification from text”. Digital Investigation, 8 (1):78–88, 2011.
8. Günal, S. et al. “On feature extraction for spam e-mail detection”. Multimedia Content Representation, Classification and Security. Springer Berlin Heidelberg, 2006

9. Maks, Isa and Vossen, Piek. "A lexicon model for deep sentiment analysis and opinion mining applications". Decision Support Systems, 53: 680–688, 2012.
10. Bellegarda, J. "Statistical language model adaptation: review and perspectives". Speech communication, 42 (1): 93–108, 2004.
11. Seymore, K. and Rosenfeld, R. "Using story topics for language model adaptation". In Proc. of EUROSPEECH, 1997.
12. Nanjo, H. and Kawahara, T. "Unsupervised language model adaptation for lecture speech recognition". In ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.
13. Liu, Y. and Liu, F. "Unsupervised language model adaptation via topic modeling based on named entity hypotheses". In Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2008., pages 4921–4924, 2008.
14. Lucas-Cuesta, J. et al. "On the dynamic adaptation of language models based on dialogue information". Expert Syst. Appl., 40 (4): 1069–1085, 2013.
15. Bellegarda, J."Exploiting latent semantic information in statistical language modeling". Proceedings of the IEEE, 88 (8): 1279–1296, 2000.
16. Blei, D. and Ng, A. and Jordan, M. "Latent dirichlet allocation". Journal of Machine Learning Research, 3: 993–1022, 2003.
17. Chien, J.T. and Chueh, C.H. "Dirichlet class language models for speech recognition". Audio, Speech, and Language Processing, IEEE Transactions on, 19 (3): 482–495, 2011.
18. Florian, R. and Yarowsky, D. " Dynamic nonlocal language modeling via hierarchical topic-based adaptation". In Proc. of the ACL, pages 167–174, 1999.
19. Iyer, R. and Ostendorf, M. "Modeling long distance dependence in language: Topic mixtures versus dynamic cache models". Speech and Audio Processing, IEEE Transactions on, 7 (1): 30–39, 1999.
20. Uysal, A. and Günal, S. "The impact of preprocessing on text classification". Information Processing and Management, 50: 104–112, 2014.
21. Padró, L. and Stanilovsky, E. "Freeling 3.0: Towards Wider Multilinguality". Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey, May 2012. ELRA.
22. Dumais, S. " Improving the retrieval of information from external sources". Behavior Research Methods, Instruments, & Computers, 23 (2): 229–236, 1991.
23. Federico, M. and Bertoldi, N. "Broadcast news LM adaptation over time". Computer Speech & Language, 18 (4): 417–435, 2004.
24. Chiu, H. and Chen, B. "Word topical mixture models for dynamic language model adaptation". In Proc. of IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2007, volume 4, pages 169–172, 2007.
25. Mostefa, D. and Hamon, O. and Moreau, N. and Choukri, K. "Evaluation Report for the Technology and Corpora for Speech to Speech Translation (TC-STAR Project). Deliverable N. 30"
26. Koehn, P. "Europarl: A Parallel Corpus for Statistical Machine Translation". In Proc. of the 10th Conference on Machine Translation (MT Summit'05), 2005.

# Using Automatic Speech Transcriptions in Lecture Recommendation Systems

A. Pérez-González-de-Martos, J.A. Silvestre-Cerdà, M. Rihtar,
A. Juan, and J. Civera

MLLP, DSIC, Universitat Politècnica de València (UPV)
{aperez,jsilvestre,ajuan,jcivera}@dsic.upv.es
Jožef Stefan Institute (IJS)
matjaz.rihtar@ijs.si

**Abstract.** One problem created by the success of video lecture repositories is the difficulty faced by individual users when choosing the most suitable video for their learning needs from among the vast numbers available on a given site. Recommender systems have become extremely common in recent years and are used in many areas. In the particular case of video lectures, automatic speech transcriptions can be used to zoom in on user interests at a semantic level, thereby improving the quality of the recommendations made. In this paper, we describe a video lecture recommender system that uses automatic speech transcriptions, alongside other relevant text resources, to generate semantic lecture and user models. In addition, we present a real-life implementation of this system for the VideoLectures.NET repository.

**Keywords:** recommender systems, automatic speech recognition, video lectures

## 1 Introduction

Online multimedia repositories are rapidly growing and being increasingly recognised as key knowledge assets. This is particularly true in the area of education, where large repositories of video lectures and Massive Open Online Courses (MOOCs) are becoming a permanent feature of the learning paradigm in higher education. A well-known example of this is the VideoLectures.NET repository, which currently includes more than 18,000 educational videos covering different topics of science.

These repositories are being subtitled in several languages in order to make them accessible to speakers of different languages and to people with disabilities [4, 21]. The lack of efficient solutions to meet this need is the motivation behind the European project transLectures [15, 19], which aims at developing innovative, cost-effective solutions for producing accurate transcriptions and translations for large video repositories. Transcriptions and translations of video lectures are the basis from which numerous other technologies can be derived.

For instance, digital content management applications such as lecture categorisation, summarisation, automated topic finding, plagiarism detection and lecture recommendation.

This latter has become essential due to the significant growth of video lecture repositories. Users are often overwhelmed by the amount of lectures available and may not have the time or knowledge to find the most suitable videos for their learning requirements. Up until recently, recommender systems have mainly been applied in areas such as music [8, 10], movies [2, 22], books [11] and e-commerce [3], leaving video lectures largely to one side. Only a few contributions to this particular area can be found in the literature, most of them focused on VideoLectures.NET [1]. However, none of them has explored the possibility of using lecture transcriptions to better represent lecture contents at a semantic level.

In this paper we describe a content-based lecture recommender system that uses automatic speech transcriptions, alongside lecture slides and other relevant external documents, to generate semantic lecture and user models. In Section 2 we give an overview of this system, focusing on the text extraction and information retrieval process, topic and user modeling and the recommendation process. In Section 3 we address the dynamic update of the recommender system and the required optimisations needed to maximise the scalability of the system. The integration of the system presented in Sections 2 and 3 into VideoLectures.NET, carried out as part of the PASCAL Harvest Project La Vie, is described in detail in Section 4. Finally, we close with some concluding remarks, in Section 5.

## 2 System Overview

Fig. 1 gives an overview of the recommender system. The left-hand side of the figure show the topic and user modeling procedure, which can be seen as the training process of the recommender system. To the right we see the recommendation process. The aim of topic and user modeling is to obtain a simplified representation of each video lecture and user. The resulting representations are stored in a recommender database. This database will be exploited later in the recommendation process in order to recommend lectures to users.

As shown in Fig. 1, every lecture in the repository goes through the topic and user modeling process, which involves three steps. The first step is carried out by the text extraction module. This module comprises three submodules: ASR (Automatic Speech Recognition), WS (Web Search) and OCR (Optical Character Recognition). As its name suggests, the ASR submodule generates an automatic speech transcription of the video lecture. The WS submodule uses the lecture title to search for related documents and publications on the web. The OCR submodule extracts text from the lecture slides, where available. The second step takes the text retrieved by the text extraction module and computes a *bag-of-words* representation. This bag-of-words representation consists of a simplified text description commonly used in natural language processing and information retrieval. More precisely, the bag-of-words representation of a given

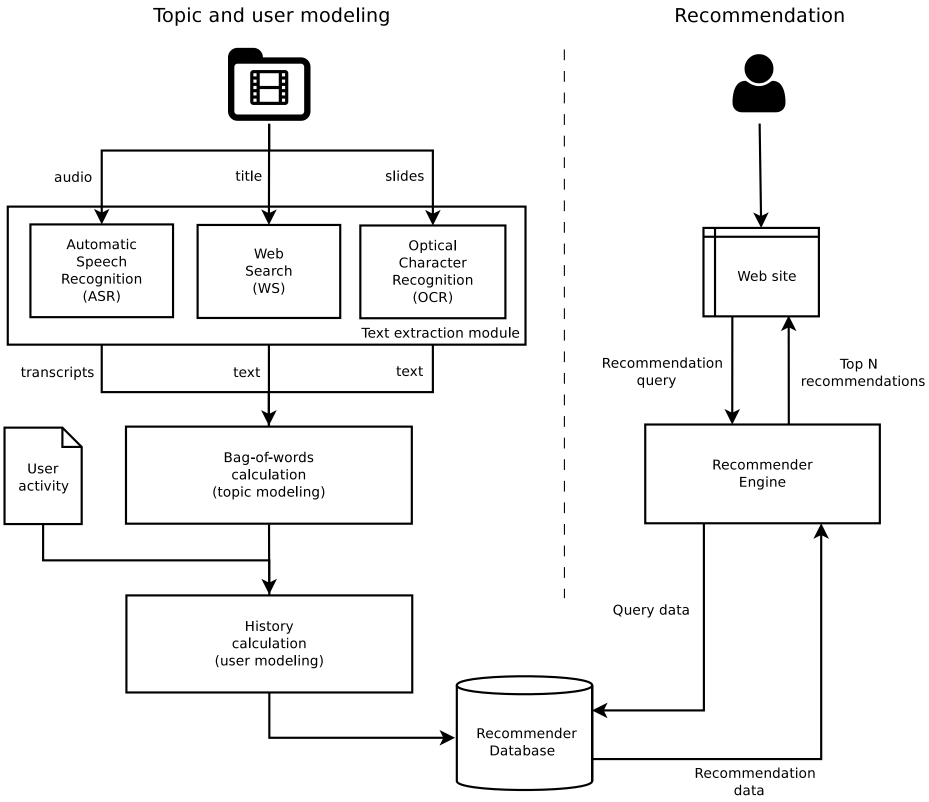


**Fig. 1.** System overview.

text is its vector of word counts over a fixed vocabulary. Finally, in the third step, lecture bags-of-words are used to represent the users of the system. That is, each user is represented as the bag-of-words computed over all the lectures the user has ever seen.

When the topic and user modeling process ends, the recommender database is ready for exploitation by the recommender engine (see the right-hand side of Fig. 1). This engine uses recommendation features to calculate a measure $s$ of the suitability of the recommendation for every ($u$, $v$, $r$) triplet, where $u$ refers to a particular user, $v$ is the lecture they are currently viewing and $r$ is a hypothetical lecture recommendation. In recommender systems, this is usually referred to as the *utility function* [13]. Specifically, it indicates how likely it is that a user $u$ would want to watch lecture $r$ after viewing lecture $v$. For instance, this utility function can be computed as a linear combination of recommendation features:

$$s(u,v,r) = \boldsymbol{w} \cdot \boldsymbol{x} = \sum_{n=1}^{N} w_n \cdot x_n \tag{1}$$

where $\boldsymbol{x}$ is a feature vector computed for the triplet ($u$, $v$, $r$), $\boldsymbol{w}$ is a feature weight vector and $N$ is the number of recommendation features. In this work, the following recommendation features were considered:

1. *Lecture popularity:* number of visits to lecture $r$.
2. *Content similarity:* weighted dot product between the lecture bags-of-words $v$ and $r$ [6].
3. *Category similarity:* number of categories (from a predefined set) that $v$ and $r$ have in common.
4. *User content similarity:* weighted dot product between the bags-of-words $u$ and $r$.
5. *User category similarity:* number of categories in common between lecture $r$ and all the categories of lectures the user $u$ has watched in the past.
6. *Co-visits:* number of times lectures $v$ and $r$ have been seen in the same browsing session.
7. *User similarity:* number of different users that have seen both $v$ and $r$.

Feature weights $\boldsymbol{w}$ can be learned by training different statistical classification models, such as support vector machines (SVMs), using positive and negative ($u$, $v$, $r$) recommendation samples.

The most suitable recommendation $\hat{r}$ for a given $u$ and $v$ is computed as follows:

$$\hat{r} = \underset{r}{\operatorname{argmax}}\, s(u,v,r) \tag{2}$$

However, in recommender systems the most common practice is to provide the user the $M$ recommendations $r$ that achieve the highest utility values $s$, for instance, the first 10 lectures.

## 3 System Updates and Optimisation

Lecture repositories are rarely static. They may grow to include new lectures, or have outdated videos removed. Also, users' learning progress or interactions with the repository influence the user models. The recommender database must therefore be constantly updated in order to include the new lectures added to the repository and update the user models. Furthermore, the addition of new lectures to the system might lead to changes to the bag-of-words (fixed) vocabulary. Any variation to this vocabulary involves a complete regeneration of the recommender database. That said, changes to the vocabulary may not be significant until a substantial percentage of new lectures has been added to the repository.

Two different update scenarios can be defined: the incorporation of new lectures and updating the user models, on the one hand, and the redefiniton of the bag-of-words vocabulary, including the regeneration of both the lecture and user bags-of-words, on the other. We will refer to these scenarios as *regular update* and *occasional update*, respectively, after the different periodicities with which they are meant to be run.

- *Regular update*: The regular update is responsible for including the new lectures added to the repository and updating the user models with the last user activity, both in the recommender database. As its name suggests, this process is meant to be run on a daily basis, depending on the frequency with which new lectures are added to the repository, since new lectures cannot be recommended until they have been processed and included in the recommender database.
- *Occasional update*: As mentioned in Section 2, lecture bags-of-words are calculated under a fixed vocabulary. Since there is no vocabulary restriction on the text extraction process, we need to modify the bag-of-words vocabulary as new lectures are added to the system. The occasional update carries out the process of updating this vocabulary, which involves recalculating both the lecture and user bags-of-words.

In order to maximise the scalability of the system, while also reducing the response time of the recommender, the features *Content similarity*, *Category similarity*, *Co-visits* and *User similarity* described in Section 2 are precomputed for every possible lecture pair and stored in the recommender database. Then, during the recommendation process, the recommender engine loads the values of these features, leaving the computation of features *User content similarity* and *User category similarity* until runtime. The decision to calculate the features *User content similarity* and *User category similarity* at runtime was driven by the highly dynamic nature of the user models, in contrast to the lecture models, which remain constant until the bag-of-words vocabulary is changed.

## 4 Integration into VideoLectures.NET

The proposed recommendation system was implemented and integrated into the VideoLectures.NET repository during the PASCAL2 Harvest Project *La Vie*

(*Learning Adapted Video Information Enhancer*) [12]. Said integration is discussed here across five subsections. First, we describe the VideoLectures.NET repository, in Section 4.1. In Section 4.2 we give a brief overview of the transLectures project, as part of which transcriptions of sufficient accuracy as to be usefully deployed were generated for lectures in this repository. Next, we address topic and user modeling from video lecture transcriptions and other text resources, in Section 4.3. In Section 4.4 we describe how recommender feature weights were learned from data collected from the existing VideoLectures.NET recommender system. Finally, we present our evaluation of the system in Section 4.5.

### 4.1 The VideoLectures.NET Repository

VideoLectures.NET [20] is a free and open access repository of video lectures mostly filmed by people from the Jožef Stefan Institute (JSI) at major conferences, summer schools, workshops and other events from many fields of science. It collects high quality educational content, recorded to high quality, homogeneous standards. The portal is aimed at promoting science, the exchange ideas and knowledge sharing by providing high quality didactic contents not only for the scientific community, but also the general public. VideoLectures.NET has so far published more than 18,000 educational videos. Relevant details regarding the repository can be found in Table 1.

**Table 1.** Basic statistics on the VideoLectures.NET repository (June 2014)

| | |
|---|---:|
| Number of videos | 18,824 |
| Total number of authors | 12,252 |
| Total duration (in hours) | 11,608 |
| Average lecture duration (in minutes) | 37 |

### 4.2 transLectures

The generation of accurate speech transcriptions for the VideoLectures.NET repository was carried out as part of the European research project transLectures [19]. transLectures aims to develop a set of tools for the automatic generation of quality transcriptions and translations for large video lecture repositories. At the scientific level, the goals of transLectures are to advance the state-of-the-art in model adaptation (to the domain, to the speaker, and using title searches and text data extracted from the presentation slides) and intelligent human-machine interaction, both as means of efficiently improving the end quality of the automatic transcriptions and translations generated.

The English subset of the VideoLectures.NET repository was automatically transcribed using the transLectures-UPV Toolkit [18]. The recommender system was able to access the transcriptions via the transLectures Platform API [16, 17].

### 4.3 Topic and User Modeling

The first step in generating lecture and user models involved collecting textual information from different sources. In particular, for VideoLectures.NET, the text extraction module gathered textual information from the following sources:

- transLectures speech transcriptions.
- Web search-based textual information from Wikipedia, DBLP and Google (abstracts and/or articles).
- Text extracted from lecture presentation slides (PPT, PDF or PNG using Optical Character Recognition (OCR)).
- VideoLectures.NET internal database metadata.

Next, the text extraction module output was used to generate lecture bags-of-words for every lecture in the repository. These bags-of-words, as mentioned in Section 2, were calculated under a fixed vocabulary that was obtained by applying a threshold to the number of different lectures in which a word must appear in order to be included. By means of this threshold, vocabulary size is significantly reduced, since uncommon and/or very specific words are disregarded. Once defined, term weights were calculated using *term frequency-inverse document frequency* (td-idf), a statistical weighting scheme commonly used in information retrieval and text mining [9]. Specifically, tf-idf weights are used to calculate the features *Content similarity* and *User content similarity.* Finally, the VideoLectures.NET user activity log was parsed in order to obtain values for the feature *Co-visits* for all possible lecture pairs, as well as a list of lectures viewed per user. This list was used together with the lectures bags-of-words to generate the users bags-of-words and categories. These, in turn, were used to calculate *User content similarity* and *User category similarity*, respectively, as well as *User similarity* for all possible lecture pairs. In a final step, all this data was stored in the recommender database in order to be exploited by the recommender engine in the recommendation process.

### 4.4 Learning Recommendation Feature Weights

Once the data needed to compute recommendation feature values for every possible $(u, v, r)$ triplet in the repository was made available, the next step was to learn the optimum feature weights $\boldsymbol{w}$ for the calculation of the utility function shown in Equation 1. To this end, an SVM classifier was trained using data collected from the existing VideoLectures.NET naïve recommender system (based only on keywords extracted from the lecture titles). Specifically, every time a user clicked on any of the 10 recommendation links provided by this recommender system, 1 positive and 9 negative samples were registered. SVM training was performed using the SVM$^{light}$ open-source software [7]. The optimum feature weights were those that obtained the minimum classification error over the recommendation data.

### 4.5 Evaluation

Although there are many different approaches to the evaluation of recommender systems [14, 5], it is difficult to state any firm conclusions regarding the quality of the recommendations made until they are deployed in a real-life setting. The La Vie project therefore provided an ideal evaluation framework, being deployed across the official VideoLectures.NET site. The strategy followed for the objective evaluation of the La Vie recommender was to compare it against the existing VideoLectures.NET recommender by means of a *coin-flipping* approach. Specifically, this approach consisted of logging user clicks on recommendation links provided by both systems on a 50/50 basis and comparing the total number of clicks recorded for each system.

The results did not show any significant differences between the two recommenders in terms of user behaviour. This can be explained by the fact that user-click count alone is not a legitimate point of comparison for recommendation quality. For instance, random variables not taken into account might influence how users respond to the recommendation links provided. As an alternative, we can compare the rank of the recommendations clicked by users within each system. Specifically, for each recommendation clicked by a user in either system, we can compare how the same recommendation ranked in the other system. This might be a more appropriate measure for comparing the recommendations in terms of suitability. However, additional data need to be collected in order to carry out this alternative evaluation. This data is currently being collected and future evaluation results will be obtained following this rank comparison approach.

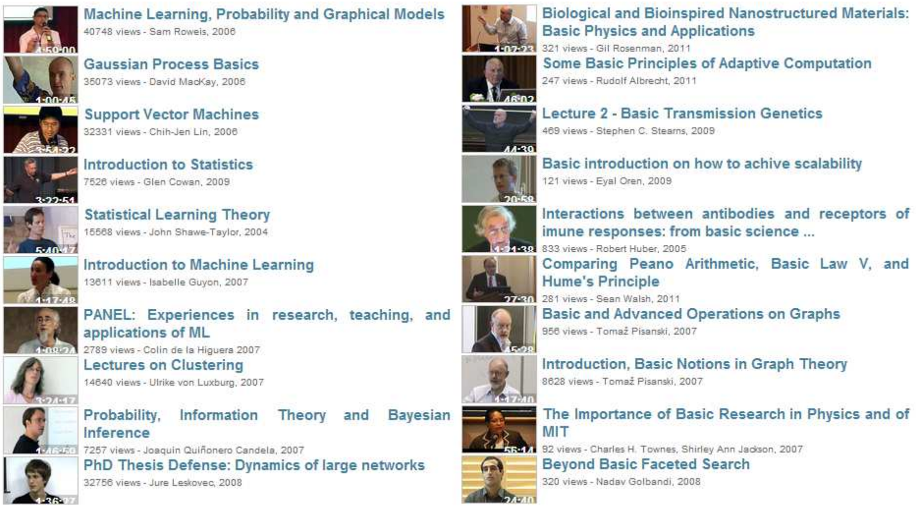
Despite the lack of objective evidence for assessing the comparative performance of the La Vie system, subjective evaluations indicate that the proposed recommender system provides better recommendations than the existing VideoLectures.NET recommender. Fig. 2 shows recommendation examples from both systems for a new user viewing a random VideoLectures.NET lecture. Although recommendation suitability is a subjective measure, La Vie recommendations seem to be more appropriate in terms of content similarity.

## 5 Conclusions

In this paper we have shown how automatic speech transcriptions of video lectures can be exploited to develop a lecture recommender system that can zoom in on user interests at a semantic level. In addition, we have described how the proposed recommender system has been particularly implemented for the VideoLectures.NET repository. This implementation was later deployed in the official VideoLectures.NET site.

The proposed system could also be extended for deployment across more general video repositories, provided that video contents are well represented in the data obtained by the text extraction module.

By way of future work we intend to evaluate the recommender system using other evaluation approaches that measure the suitability of the recommendations



Machine Learning, Probability and Graphical Models
40748 views - Sam Roweis, 2006

Gaussian Process Basics
35073 views - David MacKay, 2006

Support Vector Machines
32331 views - Chih-Jen Lin, 2006

Introduction to Statistics
7526 views - Glen Cowan, 2009

Statistical Learning Theory
15568 views - John Shawe-Taylor, 2004

Introduction to Machine Learning
13611 views - Isabelle Guyon, 2007

PANEL: Experiences in research, teaching, and applications of ML
2789 views - Colin de la Higuera 2007

Lectures on Clustering
14640 views - Ulrike von Luxburg, 2007

Probability, Information Theory and Bayesian Inference
7257 views - Joaquin Quiñonero Candela, 2007

PhD Thesis Defense: Dynamics of large networks
32756 views - Jure Leskovec, 2008

Biological and Bioinspired Nanostructured Materials: Basic Physics and Applications
321 views - Gil Rosenman, 2011

Some Basic Principles of Adaptive Computation
247 views - Rudolf Albrecht, 2011

Lecture 2 - Basic Transmission Genetics
469 views - Stephen C. Stearns, 2009

Basic introduction on how to achive scalability
121 views - Eyal Oren, 2009

Interactions between antibodies and receptors of imune responses: from basic science ...
833 views - Robert Huber, 2005

Comparing Peano Arithmetic, Basic Law V, and Hume's Principle
281 views - Sean Walsh, 2011

Basic and Advanced Operations on Graphs
956 views - Tomaž Pisanski, 2007

Introduction, Basic Notions in Graph Theory
8628 views - Tomaž Pisanski, 2007

The Importance of Basic Research in Physics and of MIT
92 views - Charles H. Townes, Shirley Ann Jackson, 2007

Beyond Basic Faceted Search
320 views - Nadav Golbandi, 2008

**Fig. 2.** On the left, La Vie system recommendations for a new user after viewing "Basics of probability and statistics" VideoLectures.NET lecture. On the right, recommendations offered by VideoLectures.NET's existing system.

more accurately, such as the aforementioned recommendation rank comparison. In addition, it is our intention to perform several analysis on the importance of the speech transcription with respect to other variables regarding recommendations quality.

**Acknowledgments.** The research leading to these results has received funding from the PASCAL2 Network of Excellence under the PASCAL Harvest Project La Vie, the EU 7th Framework Programme (FP7/2007-2013) under grant agreement no. 287755 (transLectures), the ICT Policy Support Programme (ICT PSP/2007-2013) as part of the Competitiveness and Innovation Framework Programme (CIP) under grant agreement no. 621030 (EMMA), the Spanish MINECO Active2Trans (TIN2012-31723) research project, and by the Spanish Government with the FPU scholarship AP2010-4349.

## References

1. Antulov-Fantulin, N., Bošnjak, M., Znidaršic, M., Grcar, M.e.a.: Ecml-pkdd 2011 discovery challenge overview. Discovery Challenge (2011)
2. Carrer-Neto, W., Hernández-Alcaraz, M.L., Valencia-García, R., García-Sánchez, F.: Social knowledge-based recommender system. application to the movies domain. Expert Systems with Applications 39(12), 10990–11000 (2012)
3. Castro-Schez, J.J., Miguel, R., Vallejo, D., López-López, L.M.: A highly adaptive recommender system based on fuzzy logic for b2c e-commerce portals. Expert Systems with Applications 38(3), 2441–2454 (2011)

4. Fujii, A., Itou, K., Ishikawa, T.: Lodem: A system for on-demand video lectures. Speech Communication 48(5), 516 – 531 (2006)
5. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS) 22(1), 5–53 (2004)
6. Joachims, T.: A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Tech. rep., DTIC Document (1996)
7. Joachims, T.: Svmlight: Support vector machine. SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund 19(4) (1999)
8. Lee, S.K., Cho, Y.H., Kim, S.H.: Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations. Information Sciences 180(11), 2142–2155 (2010)
9. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
10. Nanopoulos, A., Rafailidis, D., Symeonidis, P., Manolopoulos, Y.: Musicbox: Personalized music recommendation based on cubic analysis of social tags. Audio, Speech, and Language Processing, IEEE Transactions on 18(2), 407–412 (2010)
11. Núñez-Valdéz, E.R., Cueva Lovelle, J.M., Sanjuán Martínez, O., García-Díaz, V., Ordoñez de Pablos, P., Montenegro Marín, C.E.: Implicit feedback techniques on recommender systems applied to electronic books. Computers in Human Behavior 28(4), 1186–1193 (2012)
12. PASCAL Harvest Programme. http://www.pascal-network.org/?q=node/19
13. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM 40(3), 56–58 (1997)
14. Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: Recommender systems handbook, pp. 257–297. Springer (2011)
15. Silvestre, J.A., del Agua, M., Garcés, G., Gascó, G., Giménez-Pastor, A., Martínez, A., de Martos, A.P.G., Sánchez, I., Martínez-Santos, N.S., Spencer, R., Miró, J.D.V., Andrés-Ferrer, J., Civera, J., Sanchís, A., Juan, A.: translectures. In: Proceedings of IberSPEECH 2012 (2012)
16. Silvestre-Cerdà, J.A., Pérez, A., Jiménez, M., Turró, C., Juan, A., Civera, J.: A system architecture to support cost-effective transcription and translation of large video lecture repositories. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2013. pp. 3994–3999 (2013), http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6722435
17. The transLectures-UPV Team: The transLectures Platform (TLP). http://translectures.eu/tlp
18. The transLectures-UPV Team: transLectures-UPV toolkit (TLK) for Automatic Speech Recognition. http://translectures.eu/tlk
19. UPVLC and XEROX and JSI-K4A and RWTH and EML and DDS: translectures. https://translectures.eu/ (2012)
20. Videolectures.NET: Exchange ideas and share knowledge. http://www.videolectures.net/
21. Wald, M.: Creating accessible educational multimedia through editing automatic speech recognition captioning in real time. Interactive Technology and Smart Education 3(2), 131–141 (2006)
22. Winoto, P., Tang, T.Y.: The role of user mood in movie recommendations. Expert Systems with Applications 37(8), 6086–6092 (2010)

# Detection of sexist language by using text classification techniques

Carlos-D. Martínez-Hinarejos[1] and Ruxandra Dumitru[2]

[1] Pattern Recognition and Human Language Technologies Research Center - Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia, Spain
[2] Escola Tècnica Superior d'Enginyeria - Universitat de València, Avinguda de la Universitat, s/n, 46100, Burjassot, Spain

**Abstract.** Sexist language is the use of language that reinforces differences between women and men. Sexist language has a strong social influence, since propagates beliefs that get reflected in social behaviour, and therefore must be avoided in order to guarantee equal opportunities for women and men. This is specially important in public institutions documents. Nowadays, detecting and correcting sexist language in official documents is a hand-made task. However, this task could benefit of the use of Text Classification techniques in order to detect sexist expressions and allow their correction. We present the use of several text classifiers in the detection of sexist language in Spanish official documents. Results show that the automatic detection guarantees a low rate of sexist expressions in the final document (about 10%) with an admissible extra review work.

## 1 Introduction

Language is a tool of daily and massive use in modern cultures, and is a mainstream influence in human thinking and behaviour, making concepts and ideas to be assimilated in a conscious and (most times) unconscious way. Language is strongly linked to culture, and allows to build reality and truth. Thus, language is as well a powerful manipulation tool, as Aldous Huxley said in his famous sentence "Words are magical in the way they affect the minds of those who use them".

From this viewpoint, we can understand the importance of sexist language in our culture. Sexist language is usually defined as the use of speech that reinforces and perpetuates gender stereotypes and status difference between women and men [6]. Sexist language has strong implications on the development of social sexist attitudes, as some studies reveal [31]. Thus, avoiding sexist language may lead to a more equal-opportunity society since many sexist attitudes would disappear by properly managing the language. The use of language that avoids sex implications leads to the so-called inclusive language.

Of course, sexist language problem is highly dependant on the language. Low-sexed languages such as English present sexist treatment in only a few cases: use of third person personal and possessive pronouns (he, she, his, her), some typically used nouns that presuppose gender (fireman, waitress, mankind), and some unfortunate sentences that reinforce stereotypes (e.g., "After the nursing student graduates, she must face a difficult state board examination", but not all nursing students are women, or "Wives of senior government officials are required to report any gifts they receive", but not all senior government officials are men) [11].

However, strong-sexed languages present more problems. This is the case of the Spanish language, where not only nouns and pronouns are sexed, but articles and adjectives as well. Thus, sexist language is more present in this type of languages, where it is more difficult to find neutral nouns and adjectives in the daily use of the language.

Anyway, in the last years some advances have been achieved to avoid sexist language, such as style recommendations (English usage recommendations are available in [24], Chapter 8, and [34] defines recommendations for Spanish in university environment). Moreover, many political actions have been taken in order to improve equal opportunities, including inclusive use of language (e.g., creation of *Ministerio de Igualdad* by the Spanish government in 2008, now included in the *Ministerio de Sanidad*), although many decisions have been controversial from the linguistic point of view (e.g., proposing the use of inexistent words or adopting ungrammatical structures).

In any case, changing social uses is a long and difficult task that must be led by public institutions, since they must be models on upholding laws, and equal opportunities laws are present in many countries. Thus, official language must avoid sexist terminology and guarantee the use of inclusive language.

In this framework, it is of particular importance reviewing official texts in order to avoid sexist expressions, by changing them into inclusive expressions. In this task, Machine Learning techniques, and more specifically Text Classification (TC) techniques, could be applied to automatically detect whose sentences of a document are sexist and offer them to the reviewer in order to be changed. There exists one computer application, called Themis [32], that allows for this correction task in Spanish texts, but no details on how sexist language is detected are known. Thus, to the best of our knowledge, the task of detecting sexist language by using TC techniques has not been addressed until now, although it has been a topic of high interest in social science [30, 31, 20].

In this work, we are addressing the problem of the identification of sexist language in Spanish official documents by using TC techniques; this identification would be used in the framework of the task of correcting the contents of official documents by a human expert. Therefore, the assessment measures must take into account both the number of sexist expressions that could appear in the final document and the corrector effort. Several types of models were used and compared. Section 2 shows a review on the usual text classifiers and makes an overview of the features of those that we used. Section 3 describes the data corpus and its features. Section 4 details the experiments and the corresponding results. Section 5 offers the main conclusions and depicts possible future work lines on this topic.

## 2 Review of text classification techniques

As far as we know, no previous work on automatic detection of sexist language based on TC techniques is available. Thus, in this section we give an overview of the most popular TC techniques. General TC techniques are described in [19], Chapter 16 and [28]. These techniques include non-parametric distance-based methods (also known as exampled based methods, such as $K$-Nearest Neighbour), Naive Bayes models [18], Rocchio's method [25], Decision Trees [23], regression methods such as Linear Least Square Fit [35], linear classifiers inferred via the Perceptron algorithm [26], Support Vector Machines [14], and Neural Networks [27].

In our case, we centered in the techniques described below, along with the most popular works that employed the corresponding model.

### 2.1 $K$-Nearest Neighbour classifier

The $K$-Nearest Neighbour ($K$-NN) classifier is a classical classifier based on the distance between the representation of the labelled objects (prototypes) and the object to be classified; it employs a voting scheme such that the unknown object is classified into the class with most votes in the set of the nearest $K$ prototypes according to a given definition of distance.

$K$-NN has been applied as baseline or comparing system in TC tasks; for example, for the classical Reuters task [13, 16, 12]. In general, this classifier shows a good performance in TC, but it presents the drawback of its high time complexity [28].

### 2.2 Naive Bayes models

The Naive Bayes model relies on the assumption that the different features that compose the objects are independent among them; thus, the conditional probability is assumed to be the product of the conditional probabilities for each feature (i.e., for feature vector $X = x_1 x_2 \ldots x_D$ and class $C$, $P(X|C) = \prod_{i=1}^{D} P(x_i|C)$).

The two most popular versions of the Naive Bayes models are the multinomial and the Bernoulli classifier; the first one is used when documents are represented as counts of words (bag-of-words), whereas the second is used when documents are represented by the presence or absence of the words (binary features). Many works have used Naive Bayes classifiers; for example, to classify e-mail contents in spam and no-spam [36], or for the Newsgroup and the Reuters tasks [15]. In general, these methods are considered a quite simple but effective technique in the TC field.

### 2.3 Discriminant Linear Functions

Discriminant Linear Functions are discriminant functions that present a linear form, i.e., $g_c(x) = \sum_{i=1}^{D} w_i x_i + w_0$, where $D$ is the data dimensionality and $w_i$ are the weights ($w_0$ is a special weight called bias). These weights can be optimised for a training set by using the Perceptron algorithm [26].

A few previous works used the Perceptron algorithm and the inferred Discriminant Linear Functions in TC; for example, for the Reuters task [21, 7]. In general, Perceptron-based classifiers have obtained good performance in many experiments [28].

### 2.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) are non-probabilistic models based on obtaining the widest separation between the samples of two classes by projecting them in another space of representation; the projection function is called kernel. Points closest to the border between the classes constitute the Support Vectors that give name to the technique and allow classification in each class.

SVM have been applied in many TC works: basic SVM outperform other classifiers on the Reuters and Ohsumed tasks [13]; active learning on SVM has been used to improve classification on Reuters and Newsgroups tasks [33]; string kernels for SVM have been used as well on the Reuters task [17].

## 3 Experimental data

The data set used for the experiments was built from excerpts of different Spanish public institutions documents, since it is specially relevant in these official institutions to avoid sexist expressions (according to the non-sex discrimination principle stated in article 14 of Spanish Constitution [10]). The different texts were extracted from:

- Boletín Oficial del Estado (BOE) [3]: Spanish official newspapers that publishes all laws and public dispositions of the Spanish government.
- Boletín Oficial del País Vasco (BOPV) [4]: official newspaper for the public dispositions of the Basque Country government.
- Boletín Oficial de Castilla y León (BOCyL) [2]: official newspaper for the public dispositions of the Castilla y León autonomic government.
- Diari Oficial de la Comunitat Valenciana (DOCV) [8]: official newspaper for the public dispositions of the Comunitat Valenciana.
- Grants calls and official norms of the Universitat de València.

The data was extracted in the form of sentences, i.e., the units to be considered for classification are sentences, not whole documents or shorter word sequences. Although the classifier will not indicate which part makes the sentence sexist, sentence is a short enough unit to allow a fast detection of the sexist expression. The total number of selected sentences is 2000, with a vocabulary of 7238 words (including punctuation marks and numbers). The average length per sentence is of 38.7 words. Each sentence was manually labelled by an only human expert as sexist or non-sexist according to criteria described in [22], giving a total of 1100 sexist sentences and 900 non-sexist sentences. Some samples of sentences are shown in Figure 1.

In order to reduce the vocabulary size and avoid excessive data sparsity, a dimensionality reduction [29] was applied to the original corpus. This reduction was applied in four different steps, giving as a result four different corpora:

**Corpus A** Original corpus without punctuation marks.
**Corpus B** Corpus A without prepositions.
**Corpus C** Corpus B without conjunctions.
**Corpus D** Only nouns, pronouns, articles, and adjectives from the original corpus; these categories where detected by using a POS-tagger.

Table 1 shows the different features of the five corpora (original and the four derived) that were used in the experimental part. Feature extraction consisted of encoding each sentence as vectors of size equal to vocabulary size, where in each position is encoded the number of occurrences of the word corresponding to that position in the vocabulary (i.e., a bag-of-words representation).

Las personas interesadas dispondrán de un formulario para cumplimentar la solicitud en soporte informático . NOSEXIST
Todos tienen derecho a conocerlos y usarlos . SEXIST
Se consolidará y ampliará la oferta de cursos y seminarios en materias relacionadas con la igualdad de oportunidades en los planes de formación del personal de la Administración de la Generalitat , tanto en número de ediciones como incorporando nuevas acciones formativas . NOSEXIST
Mejorar la formación de los responsables en la gobernanza y gestión de dichas entidades , en especial en las nuevas herramientas e instrumentos de gestión . SEXIST

**Fig. 1.** Some examples of sentences in the corpus, along with their label. Sexists expressions are "Todos" (instead of "Todas las personas") for the second sentence, and "los responsables" (instead of "las personas responsables") in the fourth sentence.

**Table 1.** Text corpora features.

| Feature / Corpus | Original | A | B | C | D |
|---|---|---|---|---|---|
| Vocabulary size | 7238 | 7297 | 7277 | 7105 | 5041 |
| Avg. sentence length | 38.7 | 35.5 | 29.0 | 26.0 | 21.4 |

## 4 Experiments and results

In this section we present the experimental protocol we followed for the different techniques, the assessment measures, and the obtained results.

### 4.1 Experimental protocol

All the experiments were performed using a cross-validation approach on the datasets, using 5 partitions (i.e., 1600 sentences for training and 400 sentences for test). With respect to the parameters of the classifiers:

- $K$-NN: it used Euclidean distance and values from $K$=1 to $K$=50 (in steps of 5)
- Naive Bayes multinomial classifier: employed Laplace smoothing, with smoothing values equal to $2^{-i}$, for $i = 0, \ldots, 40$
- Naive Bayes Bernoulli classifier: apart from encoding in binary the feature vector, employed simple truncation smoothing, with the same values that the multinomial classifier
- Discriminant Linear Function (Perceptron): the Perceptron algorithm was employed with a maximum number of 100 iterations, margin ($b$) values 0, 0.5, and 1, and learning factor ($\alpha$) values 0.1, 0.5, and 1
- SVM: employed different kernels (linear, polynomial, radial, sigmoid) and cost values from 1 to $10^9$

All the techniques were implemented in Octave [9] using its standard features, except for SVM, were the `libSVM` library port to Octave was employed [5].

### 4.2 Assessment measures

The final framework for this task would be situated in an environment where a human corrector has to detect and change sexist sentences by non-sexist sentences. Therefore, the evaluation measures have to take into account two main factors: how many sentences must be revised by the corrector and how many sexist sentences would not be corrected.

The first factor is the percent of sentences classified as sexist; however, we are only interested in reviewing those that are really sexist, and the misclassified (non-sexist classified as sexist) is the extra work the reviewer has to do; thus, we will use this measure (which is usually known as False Positive - FP) for assessment of the effort. The second factor is the percent of sexist sentences classified as non-sexist (which is usually known as False Negatives - FN).

Apart from these basic measures, other measures can be used:

- Classification error rate (CER): percent of sentences misclassified (equivalent to the sum of False Positive and False Negatives).
- Precision, recall, and F-measure: classic Information Retrieval measures that have been extensively used in TC as well (see for example [28]).
- Receiver Operating Characteristic (ROC): curve that presents the relation between False Positive Rate (FPR) and True Positive Rate (TPR) for a confidence threshold.
- Area under ROC (AROC): area covered by a ROC curve with respect to the area covered by the baseline ROC curve (identity function).

Results in Subsection 4.3 will reflect these measures for the different classifiers. Along with results, confidence intervals of 95% were computed by using the bootstrapping technique, as described in [1], by using 10,000 repetitions.

### 4.3 Results

Table 2 shows the initial best results, at CER level, of the different classifiers presented in Section 2. Best results for each technique and corpus usually present the same parameters: multinomial employs smoothing parameter equal to $2^{-11}$ (except for corpus D, with $2^{-13}$), Bernoulli uses smoothing of $2^{-10}$ or $2^{-11}$, Perceptron employs $b = 0$ or $b = 1$, with $\alpha = 0.1$ (except for corpus D, with $\alpha = 0.5$), and SVM use cost equal to 1000 and sigmoid kernel (except for corpus B, that obtains the best result with a radial function). Only $K$-NN classifier presented an irregular behaviour, but results for different $K$ values in the same corpus did not present significant differences (except in a few cases). Confidence intervals for CER are in all cases around $\pm 1.8$.

From these initial result we can extract some preliminary conclusions. The first conclusion is that, taking into account that the *a priori*-based classifier will give a CER of 45% (classifying all sentences as sexist), all the presented classifiers improve significantly this baseline. Another conclusion is that the $K$-NN classifier does not work well for this task, since it presents a CER which is significantly higher in all cases with respect to other classifiers. Among the other classifiers, differences are not significant except in a few cases (e.g., Bernoulli and SVM for original corpus).

In global, Perceptron-based classifier presents the best results (except for the original corpus), with best result for corpus B (26.7 CER). Nevertheless, as previously said, there are no significant differences in Perceptron results with respect to the multinomial

**Table 2.** Initial best results (%) for each classifier and corpus at CER level, along with False Positive (FP) and False Negative (FN) results.

| Corpus | Original | | | A | | | B | | | C | | | D | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier | CER | FP | FN | CER | FP | FN | CER | FP | FN | CER | FP | FN | CER | FP | FN |
| $K$-NN | 40.6 | 30.5 | 10.1 | 40.2 | 23.3 | 16.9 | 39.4 | 30.2 | 9.2 | 39.0 | 29.6 | 9.4 | 40.1 | 33.8 | 6.3 |
| NB mult | 28.6 | 18.8 | 9.8 | 28.7 | 18.4 | 10.3 | 28.6 | 17.8 | 10.8 | 28.0 | 17.7 | 10.3 | 28.1 | 17.6 | 10.5 |
| NB Bern | 30.4 | 17.9 | 12.5 | 30.0 | 18.1 | 11.9 | 30.5 | 18.9 | 11.6 | 30.1 | 18.3 | 11.8 | 29.7 | 17.9 | 11.8 |
| Perceptron | 27.4 | 15.6 | 11.8 | 26.8 | 15.0 | 11.8 | 26.7 | 14.7 | 12.0 | 27.0 | 16.6 | 10.4 | 27.4 | 15.7 | 11.7 |
| SVM | 26.9 | 15.9 | 11.0 | 27.4 | 15.7 | 11.7 | 27.2 | 15.9 | 11.3 | 27.3 | 16.1 | 11.2 | 27.5 | 16.2 | 11.3 |

**Table 3.** Precision (P), recall (R), and F-measure (F) results (ratios). Classifiers are the same than those of Table 2.

| Corpus | Original | | | A | | | B | | | C | | | D | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier | P | R | F | P | R | F | P | R | F | P | R | F | P | R | F |
| $K$-NN | 0.59 | 0.82 | 0.69 | 0.62 | 0.69 | 0.65 | 0.60 | 0.83 | 0.70 | 0.61 | 0.83 | 0.70 | 0.58 | 0.88 | 0.71 |
| NB mult | 0.71 | 0.82 | 0.76 | 0.71 | 0.81 | 0.76 | 0.71 | 0.80 | 0.75 | 0.72 | 0.81 | 0.76 | 0.72 | 0.81 | 0.76 |
| NB Bern | 0.70 | 0.78 | 0.74 | 0.70 | 0.78 | 0.74 | 0.70 | 0.79 | 0.74 | 0.70 | 0.78 | 0.74 | 0.71 | 0.79 | 0.74 |
| Perceptron | 0.73 | 0.79 | 0.76 | 0.74 | 0.79 | 0.76 | 0.74 | 0.78 | 0.76 | 0.73 | 0.81 | 0.77 | 0.73 | 0.79 | 0.76 |
| SVM | 0.73 | 0.80 | 0.76 | 0.73 | 0.79 | 0.76 | 0.73 | 0.79 | 0.76 | 0.73 | 0.80 | 0.76 | 0.73 | 0.79 | 0.76 |

or the SVM classifier results, and only a few with respect to the Bernoulli classifier. Apart from that, differences among the different corpora are not significant for a given classifier, which shows the robustness of the techniques even for very sparse spaces.

Results on FN (that is one important measure since it gives the final percent of sexist sentences that will remain in the corrected corpus) show a better behaviour for $K$-NN and Naive Bayes multinomial classifier (FN<10), but at the expense of having a higher load of revision work. This is especially noticeable for $K$-NN, with FP values around 30%. In the case of the multinomial classifier a better balance is found, and it could be considered as a reasonable option (it guarantees around 10% of no corrected sexist sentences with an extra revision of around 18%).

Table 3 shows the precision, recall, and F-measure results for that classifiers that achieved best CER for each corpus and set of parameters. These results confirm that $K$-NN presents the lowest performance and that Perceptron-based and SVM classifiers have the best behaviour, with small differences with respect to multinomial classifier.

A final assessment measure that can be obtained is based on confidence classification for each classifier, by using ROC curves. In this case, the critical event is classification as non-sexist, since the critical error is classifying a sexist sentence as non-sexist (FN event). Therefore, confidence measure is taken in terms on how confident is classified a sentence as non-sexist, and is computed for each sentence based on the difference between the probability the model assigns to each class for that sentence. Differences are normalised to obtain the confidence measure in rank [0-1] and TPR (non-sexist correctly classified) and FPR (sexist incorrectly classified) are computed for the different values of the confidence measures obtained for each classifier.

Figure 2 shows the ROC curves associated to the best combination of classifier-corpus for each type of classifier. These results confirm that $K$-NN is the classifier with

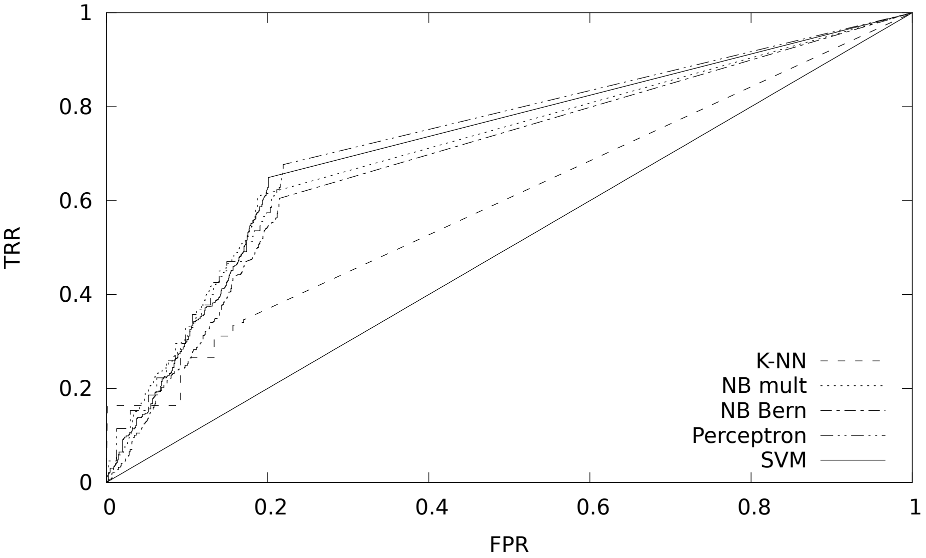


**Fig. 2.** ROC curves for best classifier of each type based on CER results: $K$-NN corpus C, multinomial corpus C, Bernoulli corpus D, Perceptron corpus B, SVM original corpus. Diagonal line shows the baseline.

lowest performance, whereas Perceptron-based classifier shows the best behaviour, very similar to that of SVM. Naive Bayes-based classifiers show a bit lower performance; in those two classifiers, again multinomial classifier shows a bit better performance than the Bernoulli classifier. To complete these results, AROC measure was calculated for best classifiers of each type for each corpus; results are shown in Table 4 and confirm the same conclusions that were obtained by using CER and F-measure.

## 5 Conclusions and future work

In this work we have presented the detection of sexist language in Spanish official documents by using TC techniques. This detection is presented in the framework of official texts correction, where sexist sentences must be detected and changed into non-sexist sentences. A variety of text classifiers were tested and compared. The Perceptron-based classifier was the one that presented a lowest classification error. However, Naive Bayes multimodal classifier provided a lower number of False Negative results (i.e., sexist documents that are not detected and remain in the final revised corpus) at expense of a bit higher reviewing effort (False Positives). Preprocess techniques applied to the corpus did not reveal a significant influence. CER results were contrasted with other measures, such like precision-recall and ROC-AROC, giving similar conclusions.

Future work could be directed to improve the performance of the classification by employing other classifiers (e.g., decision trees, Hidden Markov Models, Conditional Random Fields), by improving general feature extraction (e.g., using bigrams or trigrams instead of a bag-of-word representation, using stopwords), by extracting more

**Table 4.** AROC values for the best classifiers of each type in each corpus.

| Corpus | Original | A | B | C | D |
|---|---|---|---|---|---|
| Classifier | | | | | |
| $K$-NN | 0.59 | 0.62 | 0.59 | 0.60 | 0.57 |
| NB mult | 0.69 | 0.69 | 0.69 | 0.70 | 0.70 |
| NB Bern | 0.68 | 0.68 | 0.68 | 0.69 | 0.69 |
| Perceptron | 0.72 | 0.72 | 0.72 | 0.71 | 0.72 |
| SVM | 0.71 | 0.71 | 0.71 | 0.71 | 0.71 |

specific features (e.g., gender detection, POS categories), or by enriching the corpus with more examples. This work could be considered as a first step on using TC for social-aware tasks; thus, not only sexist language could be detected, but racist or discriminatory language in general. A more advanced step could be the classification of text contents according to age scales, i.e., determining if a text has contents that are suitable for all ages, for children older than 7, only for adults, etc. The application of this last task would be immediate for web page filtering and parental control when accessing web pages. Apart from that, extension to other languages and types of documents (not only official documents, but other social impacting documents such as newspapers, advertisements, etc.) is a must in order to extend the sexist language detection to all social layers. Finally, in the framework we proposed, sexist sentences must be converted into non-sexist sentences, and Machine Translation techniques could be used to make proposals to the reviewer in order to speed-up the whole process.

## Acknowledgements

Work partially supported by Spanish MEC under project Active2Trans (TIN2012-31723), and by GVA under project PROMETEOII/2014/030.

## References

1. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in asr performance evaluation. In: Proceedings of ICASSP'04. vol. 1, pp. 409–412 (May 2004)
2. BOCyL: Boletín Oficial de Castilla y León (2014), `http://bocyl.jcyl.es/`
3. BOE: Agencia Estatal Boletín Oficial del Estado (2014), `http://www.boe.es/`
4. BOPV: Boletín Oficial del País Vasco (2014), `http://www.lehendakaritza.ejgv.euskadi.net/r48-bopv2/es/bopv2/datos/Ultimo.shtml`
5. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
6. Crawford, M.: Gender and language. In: R., U. (ed.) Handbook of the psychology of women and gender, pp. 228–244. Wiley, New York, NY (2001)
7. Dagan, I., Karov, Y., Roth, D.: Mistake-Driven learning in text categorization. In: EMNLP-97. pp. 55–63 (August 1997)
8. DOCV: Diari Oficial de la Comunitat Valenciana (2014), `http://www.docv.gva.es/`

9. Eaton, J.W., Bateman, D., Hauberg, S.: GNU Octave version 3.0.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform (2009), http://www.gnu.org/software/octave/doc/interpreter, ISBN 1441413006
10. España: Constitución Española (1978), BOE núm. 311 de 29 de diciembre de 1978, http://www.boe.es/buscar/pdf/1978/BOE-A-1978-31229-consolidado.pdf
11. Hacker, D.: The Bedford Handbook. Bedford/St. Martin's, 6th edn. (2002)
12. Jiang, S., Pang, G., Wu, M., Kuang, L.: An improved k-nearest-neighbor algorithm for text categorization. Expert Syst. Appl. 39(1), 1503–1509 (Jan 2012)
13. Joachims, T.: Text categorization with suport vector machines: Learning with many relevant features. In: Proceedings of ECML '98. pp. 137–142. Springer-Verlag (1998)
14. Joachims, T.: A statistical learning learning model of text classification for support vector machines. In: Proceedings of SIGIR '01. pp. 128–136. ACM, New York, NY, USA (2001)
15. Kim, S.B., Han, K.S., Rim, H.C., Myaeng, S.H.: Some effective techniques for naive bayes text classification. IEEE Trans. on Knowledge and Data Eng. 18(11), 1457–1466 (2006)
16. Lam, S., Lee, D.L.: Feature reduction for neural network based text categorization. In: Proceedings of DASFAA-99. pp. 195–202 (1999)
17. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. J. Mach. Learn. Res. 2, 419–444 (Mar 2002)
18. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
19. Manning, C., Schütze, H.: Foundations of statistical natural language processing. MIT Press (1999)
20. Mills, S.: Language and Sexism. Cambridge University Press (2008)
21. Ng, H.T., Goh, W.B., Low, K.L.: Feature selection, perceptron learning, and a usability case study for text categorization. In: Proceedings of SIGIR '97. pp. 67–73. ACM (1997)
22. Quilis-Merín, M., Albelda-Marco, M., Cuenca, M.J.: Guía de uso para un lenguaje igualitario. Tech. rep., Unitat d'Igualtat de la Universitat de València (2012)
23. Quinlan, J.R.: Induction of decision trees. Machine learning 1(1), 81–106 (1986)
24. Random-House: Random House English Language Desk Reference. Random House (1995)
25. Rocchio, J.: Relevance Feedback in Information Retrieval, pp. 313–323. Prentice Hall Inc. (1971)
26. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65(6), 386–408 (1958)
27. Ruiz, M.E., Srinivasan, P.: Hierarchical neural networks for text categorization (poster abstract). In: Proceedings of SIGIR '99. pp. 281–282. ACM, New York, NY, USA (1999)
28. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (Mar 2002)
29. Sebastiani, F.: Text categorization. In: Text Mining and its Applications to Intelligence, CRM and Knowledge Management. pp. 109–129. WIT Press (2005)
30. Sheridan, E., Committee, I.S.S.a.R.: Sex Stereotypes and Reading: Research and Strategies. International Reading Association (1982)
31. Swim, J., Mallett, R., Stangor, C.: Understanding subtle sexism: Detection and use of sexist language. Sex Roles 51(3/4), 117–128 (2004)
32. Themis: (2014), http://www.themis.es/
33. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2, 45–66 (Mar 2002)
34. UNED: Guía de lenguaje no sexista. Tech. rep., Oficina de Igualdad UNED (2012)
35. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of SIGIR '99. pp. 42–49. ACM, New York, NY, USA (1999)
36. Zhang, H., Li, D.: Naïve bayes text classifier. In: GrC. pp. 708–711 (2007)

# Improving accessibility for hearing impaired on board the airplane with ASR

Isabel Vázquez Rufino, Ascensión Gallardo-Antolín, Carmen Peláez Moreno, and Belén Ruiz Mezcua

Signal Theory and Communications Department, Computer Science Department
Universidad Carlos III de Madrid (Leganés, Madrid, Spain)
http://www.uc3m.es

**Abstract.** The recent improvements in Automatic Speech Recognition (ASR) systems, both in performance and computational efficiency, allow us to develop new tools for the welfare of society. One of the social sectors that can highly benefit from these systems are the hearing impaired since they could obtain transcripts of live speech in an increasing number of environments. Still, the specific conditions present on board the airplane pose several challenges. This paper introduces a tool to improve the integration and autonomy of people with hearing disabilities in airplane cabins through a system which provides captions of the crew's announcements to passengers in a readable format similar to subtitle.

**Keywords:** Accessibility, ASR, noise, hearing disability, airplane, tool

## 1 Introduction

Creating accessibility directives has long been associated solely to physical environments such as architectural barriers. However, today this concept is evolving, taking on more types of disabilities, and covering a wide range of actions that must be taken into account in the social and political strategies in order to achieve universal accessibility.

People with sensory disabilities are those affected by hearing or sight impairments (sometimes both). These people face activities limitations and social participation restrictions due to socio-environmental factors. Access to information and communication are among their basic needs some of which could be answered by developing new solutions based on automatic speech recognition technology.

According to data from the World Health Organization (WHO), more than 5% of the world's population (about 360 million people) have hearing loss, 32 of them are children [9]. As an example, there are around one million of people with hearing impairment in Spain, according to the National Statistics Institute (Instituto Nacional de Estadstica –INE–) [6].

Currently, unassisted aid[1] for people with hearing disabilities wishing to travel by plane is limited to very specific places in airport terminals. In these posi-

[1] Help without the assistance of another person.

tions magnetic induction loops are used to guide them throughout the airport[3]. Once they are on-board however, the aid disappears and every communication with them.

The tool presented in this paper is designed to fill this gap by providing captions of the crew's announcements to passengers. It integrates an ASR system [10], transcribing the crew's utterances and displaying them as captions both in the own airplane displays and as a web service accessible in personal devices such as cellular phones or tablets through an intranet.

Figure 1 describes the tool service. A simple graphic user interface helps the crew configure the tool and passengers access and read the captions in their own personal devices.

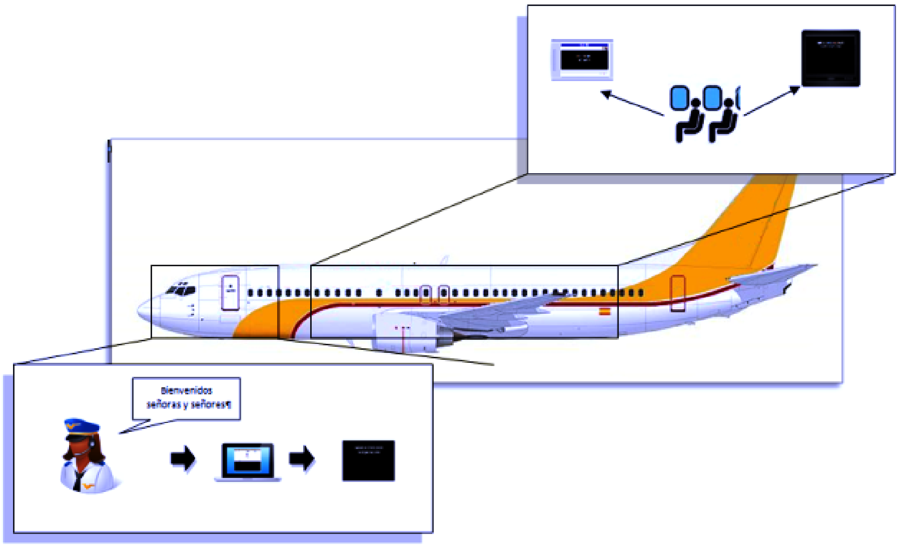


**Fig. 1.** Service description (Source:[11], p. 17)

This paper is organized as follows: Section 2 presents the challenges faced by ASR-based tool for assistance to people with hearing impairments on-board the airplane outlined in Section 1. Section 3 describes the experimental set-up designed to test the system. Experiments and results are presented in Section 4 and finally Section 5 draws some conclusions.

## 2 Challenges to overcome

This tool deals with continuous speech in two different languages (English and Spanish) and is adapted to cabin noises. In particular, the following requirements were taken into account:

1. Speaker dependent: specific profiles are created for the individual members of the crew since the performance was very affected by this feature.
2. Noise robustness: noise on board an aircraft is specific of this environment.
3. Continuous speech: crew announcements are produced in natural continuous speech.
4. Large generic vocabulary: though part of the vocabulary is specific and certainly some kind of adaptation would be of help in this case, we have concentrated in the previous challenges for the moment, leaving this issue for future work and thus employed a large generic vocabulary.

As we said before, cabin noise is an important limitation for the development of these kind of systems. Most of the important messages are announced during take-off and landing. Measurements of noise inside cabins indicate that they range between 60 and 65 dBA before take-off, 80-85 dBA during the flight and 75-80 dBA when landing[7]. Nonetheless if the microphone is properly positioned (close to the mouth of the speaker) and has a noise-cancelling mechanism the noise effects would be considerably lowered.

A classical mismatch problem appears when the acoustic models have been trained in a clean environment and the actual environment in which the system should operate is noisy as depicted in Figure 2.

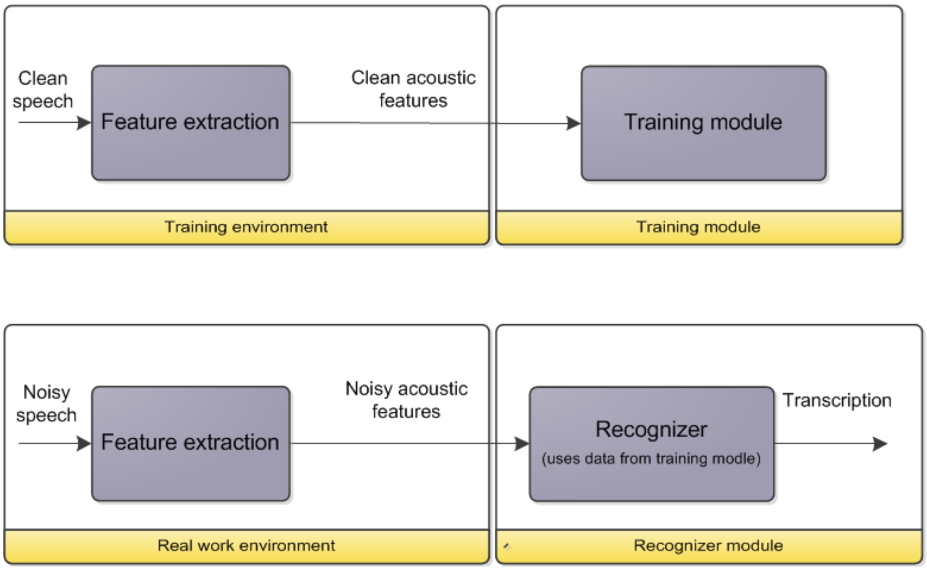


**Fig. 2.** Problem of robustness in ASR

Several preliminary experiments were carried out to enhance the input signal by noise removal and filtering techniques and also by modifying the input features by using Spectral Subtraction procedures but the fact that the speech engine[2] chosen is a proprietary software posed important limitations to the application and success of these techniques. Instead, we opted for a speaker dependent model adapted also to noise.

## 3 Experimental setup

### 3.1 Databases

Since the goal is to use a large vocabulary continuous speech system we chose a testing database materials from an audio book. In particular, we used dictated speech for about 15 minutes per chapter with volume and speakers' tones changes.

These characteristics are very similar to the actual situation in which the tool would be used, since the crew's speeches are scripted and tone and volume depend on the stress and fatigue to which they are subject.

The audio book selected was *Alice's Adventures in Wonderland by Lewis Carrol (1832 - 1898).* Two audio recordings were used[3]:

- BookA: is the audio book read by the same speaker, in US English. [1]
- BookB: is the audio book where each chapter was read by a different speaker, in US English. [2]

All audio files were resampled at a frequency of 16 kHz and converted to .WAV format.

The book was split into 12 chapters, 2 of them used for training (see Table 1) and the rest, for testing the tool (see Table 2). Also, recordings of aircraft cabin noise were collected and employed to produce noisy versions of the previous materials at different SNRs. In particular, all the audio was mixed with aircraft cabin noise at different Signal Noise Ratio (SNR) (see Table 3), using the MATLAB [8] tool Add Noise [12].

### 3.2 Acoustic modeling

The ASR tool allows to create and adapt acoustic models in both English and Spanish. Five acoustic models were used to assess the tool:

1. Acoustic Model #1: US English model, with large vocabulary and not adapted to any speaker.
2. Acoustic Model #2: US English model, with large vocabulary and adapted to the speaker using the first and fourth chapters (about 24 minutes) of BookA.

[2] Dragon NaturallySpeaking©2014 Nuance Communications, Inc., was chosen due to its ability to perform speaker adaptation, the availability of a large expandable vocabulary in both English and Spanish

[3] A detailed description is included in *PFC: Herramienta de transcripción automática de los comunicados al pasaje aéreo.* Chapter 4 [11].

**Table 1.** Audio for training

| **Train: BookA** | |
|---|---|
| Chapter | Length (min:seg) |
| 1 | 11:20 |
| 4 | 12:46 |
| Total | 24:07 |

**Table 2.** Audio for test

| **Test: BookB** | | | **Test: BookA** |
|---|---|---|---|
| Chapter | Length (min:seg) | Gender | Length (min:seg) |
| 1 | 10:27 | M | 11:20* |
| 2 | 11:59 | F | 11:10 |
| 3 | 17:08 | F | 08:46 |
| 4 | 19:20 | M | 12:46* |
| 5 | 13:06 | F | 11:49 |
| 6 | 12:39 | M | 14:45 |
| 7 | 17:01 | M | 11:47 |
| 8 | 13:03 | M | 13:02 |
| 9 | 14:45 | M | 13:11 |
| 10 | 22:17 | F | 12:51 |
| 11 | 09:56 | F | 10:17 |
| 12 | 12:38 | F | 11:43 |
| Total | 2:54:19 | | 2:23:27 |

Chapters marked with * were not used for test.

**Table 3.** SNR levels for test

| **SNR (dB)** | -18 | -12 | -6 | -3 | 3 | 6 | 12 | 18 |
|---|---|---|---|---|---|---|---|---|

3. Acoustic Model #3: US English model, with large vocabulary and adapted to noise and speaker using the first and fourth chapters (about 24 minutes) of BookA with added noise at SNR = -6 dB.
4. Acoustic Model #4: Spanish model, with large vocabulary and adapted to speaker with chapter first and fourth of BookA (in Spanish version).
5. Acoustic Model #5: Spanish model, with large vocabulary and adapted to speaker with chapter first and fourth of BookA (in Spanish version) mixed with noise at SNR = -9 dB.

## 4 Experiments and results

In order to characterize the tool, the following tests were performed with the audio materials described in Section 3.2:

1. Speaker independent test (Id. 1): clean input speech (without aircraft cabin noise) and Acoustic Model #1.
2. Speaker dependent test (Id. 2): clean input speech (without aircraft cabin noise) and Acoustic Model #2.
3. Noise robustness test (Id. 3): noisy input speech (with additive aircraft cabin noise at different SNR ratios) and Acoustic Model #1.
4. Noise model adaptation test (Id. 4): noisy input speech (with additive aircraft cabin noise at different SNR ratios) and model adapted with a noisy speech at SNR = -6dB (Acoustic Model #3).

A final simulation where the Spanish models with adaptation to speaker (Acoustic Models #4 and #5) are tested with clean or noisy live speech as will be further explained.

Also, as said before, preliminary tests were also carried out with noise prefiltering using Spectral Substraction (SS) [13], but results were not encouraging.

Results are shown in terms of the Word Recognition Rate (WRR) which was computed by using NISTAlign [4] from CMUSphinx [5]. This tool allows to compare two texts (the original one and the automatic transcription obtained with the ASR system) and returns as a result the percentage of correct words once the alignment between texts is performed.

– Table 4 presents results from test Id 1 and Id 2. As can be seen, speaker adaptation plays an important role by highly improving the performance on BookA. A marginal improvement is also observed on BookB.

**Table 4.** Test 1 and 2 results summary

| **Test 1 and 2 results summary (% WRR)** | | |
|---|---|---|
| **Id** | | **Mean** |
| **1** | **BookA** | 70.73 |
| | **BookB** | 51.26 |
| **2** | **BookA** | 85.01 |
| | **BookB** | 52.49 |

– Table 5 shows results from test Id 3 and Id 4 presented as the mean of the 10 chapters for each SNR and the total grand mean. Again the model adapted to speaker and noise improves the performance but only up to SNR = 3 dB. Above this SNR, the adaptation is detrimental which should be taken into account.

**Table 5.** Test 3 and 4 results summary

**Test 3 and 4 results summary (% WRR)**

| **Id** | **SNR (dB)** | | | | | | | | | **Mean** |
|---|---|---|---|---|---|---|---|---|---|---|
| | -18 | -12 | -6 | -3 | 3 | 6 | 12 | 18 | clean | |
| **3** | 14.91 | 34.49 | 60.12 | 67.88 | 76.26 | 78.66 | 81.34 | 82.65 | 83.86 | 64.69 |
| **4** | 21.05 | 49.81 | 67.09 | 71.56 | 76.71 | 78.02 | 79.13 | 79.12 | 76.87 | 66.60 |

Clean: The audio book was not mixed with noise.

- A final simulation was run in the laboratory where a speaker with a microphone headset read a guided welcome message. The noisy conditions where recreated by placing a loudspeaker playing the noise recorded inside the cabin (see Figure 3). Table 6 and Figure 4 show the results obtained.

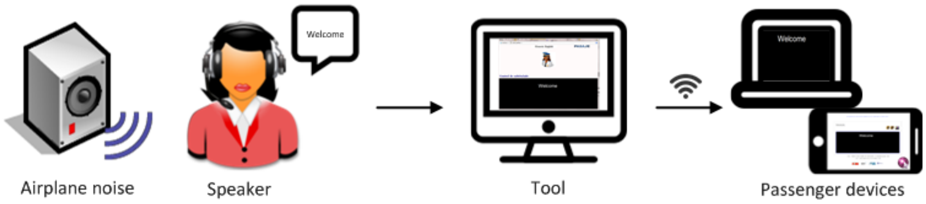


**Fig. 3.** Simulation

**Table 6.** Simulation summary

**Simulation results (% WRR)**

| **Simulation** | **Case** | **WRR** |
|---|---|---|
| 1 | Clean speech - Acoustic Model #4 | 97% |
| 2 | Clean speech - Acoustic Model #5 | 0% |
| 3 | Noisy speech - Acoustic Model #4 | 42% |
| 4 | Noisy speech - Acoustic Model #5 | 88% |

As shown in Table 4, Table 5 and Figure 4 the test results were conclusive. First, as expected, an adapted speaker model will perform better than a speaker independent one. Second, the model adapted to the environment in which it will be used (matched conditions), will get better results.

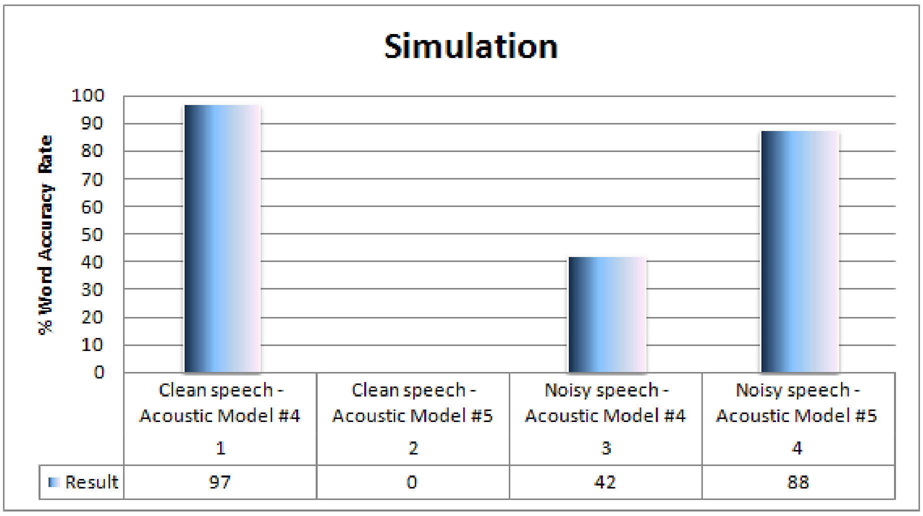


**Fig. 4.** Simulation results (% WRR)

## 5 Conclusions

A tool based on ASR to achieve the integration and autonomy of people with hearing disabilities in the aviation industry has been presented. It provides transcriptions of the crews announcement to airline passengers. Some features of these systems and their current limits, due to the noisy environment in which they operate, have been discussed.

This study has been performed on artificially distorted signals, so it would be desirable to test the performance of the system on a real scenario.

Anyway, there is a need to develop more new accessibility systems that will provided more accessible and autonomous environments for all people.

## 6 Acknowledgments

This work has been funded by the "*Cátedra EADS - Fundación ADECCO para la integración laboral de personas con discapacidad en entornos aeronáuticos*"[4] at Universidad Carlos III de Madrid in association with Centro Español del Subtitulado y la Audiodescripción (CESyA).

## References

1. L. Carroll. *Alice's Adventures in Wonderland*. Internet Archive, 2013. Read by Cory Doctorow. Available at `http://archive.org/details/AliceInWonderlandReadByCoryDoctorow`.

[4] Cátedra EADS - Fundación ADECCO for the employment of people with disabilities in aeronautical environments.

2. L. Carroll. *Alice's Adventures in Wonderland.* Librivox, 2013. Read by LibriVox Volunteers. Available at `https://librivox.org/alices-adventures-in-wonderland-by-lewis-carroll`.
3. CERMI. Discapacidad. AENA instala 250 equipos de inducción magnética para orientar a las personas sordas en los aeropuertos. *Semanal.CERMI*, 2012. Available at: `http://semanal.cermi.es/noticia/AENA-instala-equipos-inducci\%C3\%B3n-magn\%C3\%A9tica-personas-sordas-aeropuertos.aspx`.
4. CMUSphinx. *NISTAlign.* Carnegie Mellon University. Available at: `http://cmusphinx.sourceforge.net/sphinx4/javadoc/edu/cmu/sphinx/util/NISTAlign.html`.
5. CMUSphinx. *Open Source Toolkit For Speech Recognition.* Carnegie Mellon University. Information available at: `http://cmusphinx.sourceforge.net/`.
6. Subdirección General de Difusión Estadística INE. Encuesta sobre discapacidades, deficiencias y estado de salud. *INEbase*, 2000. Available at: `http://www.ine.es/prodyser/pubweb/discapa/disctodo.pdf`.
7. H. Kurtulus Ozcan and S. Nemlioglu. In-cabin noise levels during commercial aircraft flights. In *Journal of the Canadian Acoustic Organization*, volume 34. Canadian Acoustics, 2006.
8. MathWorks. *MATLAB.* Mathworks, 1994-2013. Information available at: `http://www.mathworks.es/`.
9. OMS. Deafness and hearing loss. *OMS Media centre*, no. 300, February 2013. Available at: `http://www.who.int/mediacentre/factsheets/fs300/en/index.html`.
10. S. Preeti and K. Parneet. Automatic speech recognition: A review. In *International Journal of Engineering Trends and Technology*, volume 4, page 132. February 2013.
11. Isabel Vázquez Rufino. *PFC: Herramienta de transcripción automática de los comunicados al pasaje aéreo.* Universidad Carlos III de Madrid, March 2014.
12. K. Wojcicki and K. Florian. *Add Noise.* MATLAB, 2011-2013. BSD License. Available at: `http://www.mathworks.com/matlabcentral/fileexchange/32136-add-noise/content/addnoise/addnoise.m`.
13. E. Zavarehei. *SBoll79.* MATLAB, 2005. No BSD License. Available at `http://www.mathworks.com/matlabcentral/fileexchange/7675-boll-spectral-subtraction/content/SSBoll79.m`.

# A comparative study between generative and discriminative statistical models for unsegmented dialogue annotation

Carlos-D. Martínez-Hinarejos and José-Miguel Benedí

Pattern Recognition and Human Language Technologies Research Center - Universitat Politècnica de València, Camino de Vera, s/n, 46022, Valencia, Spain

**Abstract.** Dialogue systems employ the dialogue strategy to define is behaviour in their interaction with users. The dialogue strategy is usually based on models whose parameters are estimated from dialogues annotated in terms of Dialogue Acts. Therefore, dialogue annotation is necessary to obtain dialogue systems, but manual annotation is hard to achieve and automatic annotation is desirable to obtain at least a draft annotation. The annotation problem can be formulated as an statistical optimisation problem on a sequence of turns. Some previous annotation works assumed the segmentation of turns into relevant subsequences (segments), but this segmentation is not usually available. Probabilistic annotation can be based on different statistical models. In this work, we compare the performance of two different paradigms: generative and discriminative models. These models are applied and compared in unsegmented dialogue annotation of two dialogue corpora of different nature.

## 1 Introduction

Dialogue systems are computer systems that interact with human users by means of dialogue. Dialogue systems are usually directed to solve a given task. In the last decade, there have been many projects with the aim of developing a dialogue system, such like *Companions* [18], *Classic* [13], or *PARLANCE* [9], among others. The core component of a dialogue system is the dialogue manager, which decides how to manage the user input and which output and actions must be produced. These decisions are taken by using the dialogue strategy, that uses user input, previous development of the dialogue (dialogue history), and other data sources (e.g., database related to the system task).

There are many proposals for determining the dialogue management, but they are usually divided into two classes by their approach: rule-based and data-based (or statistical). Rule-based approximation uses a set of rules, usually manually defined, that govern the dialogue system. Data-based case employs a statistical model for deciding the system response. Rule-based needs a human expert and is hard to create or adapt the rules. In contrast, data-based estimates automatically the strategy, but needs large labelled sets. Currently, data-based approach is the most usual in dialogue systems.

The concept of annotation that arises in the data-based approach consists of applying a set of labels that models the discourse structure during the dialogue. These labels represent the current state of the dialogue and the actions that can be performed by a dialogue system at each point of the interaction. Consequently, each label is associated to a communicative function or a dialogue situation. Thus, the models become defined in terms of the set of labels. Since only the dialogue relevant information is represented in the label (and consequently used by the model), the models become simpler.

There have been many proposals to define the set of labels, but the most widely accepted proposal is annotation based on Dialogue Acts (DAs). A DA is defined as a label that represents the meaning of a dialogue segment (or *utterance* [15]), which is any subsequence in a dialogue turn that has a relevant paper in the dialogue process.

The final aim of the annotation is to estimate the parameters of the dialogue models from a set of dialogues. Therefore, the first step is to annotate the training data with the set of DAs; thus, given a set of dialogues where segmentation is usually not available but the whole dialogue transcription is available, segments must be identified and DA labels must be properly assigned to each segment. Manual annotation of dialogues is hard and time-consuming; therefore, employing statistical models (estimated from a manually annotated subset of dialogues) for obtaining a draft annotation of the corpus is an interesting option in order to reduce annotation effort.

Most proposals on statistical models for dialogue annotation are based on generative models. For example, in [15], an approach based on generative models - Hidden Markov Models (HMM) and N-grams - is used on a segmented version of the SwitchBoard corpus [8]. These models are easy to estimate, but they impose severe restrictions on the nature of the probability distribution. In the last years, discriminative models have gained attention for many Natural Language Processing tasks due to its less restrictive nature and the advances on its estimation. For example, in [12], a discriminative model - Conditional Random Fields (CRF) [16] - is proposed for annotating chat dialogues.

In any case, most of this works assume that segmentation of dialogue turns into segments is available [15, 17], which is not realistic in many cases. In this work we compare the performance of generative models (HMM and N-grams) and discriminative models (CRF) for the annotation of unsegmented dialogues in two dialogue corpora of different nature and features. Section 2 presents the general statistical framework along with the generative and discriminative approximations to dialogue annotation. Section 3 describes the used corpora (SwitchBoard [8] and Dihana [3]). Section 4 details the experiments and results. Section 5 offers the conclusions and final remarks.

## 2 Statistical dialogue annotation models

Given a dialogue with $T$ turns, $W_1^T$ defines the sequences of words (input) for each turn, whereas $U_1^T$ defines the corresponding DAs sequences (output). The dialogue annotation problem could be seen as an optimisation problem that can be formulated as:

$$\widehat{\mathcal{U}} = \operatorname*{argmax}_{U_1^T} \Pr(U_1^T|W_1^T) \tag{1}$$

This problem could be solved by using a model, which could be generative or discriminative, of the probability. Generative models model the joint probability, i.e.: $\Pr(U_1^T, W_1^T)$, which via Bayes rule can be used in the maximisation of the posterior probability. However, discriminative models directly model the posterior probability. The following subsections detail how generative (HMM) and discriminative (CRF) models fit to the dialogue annotation framework.

### 2.1 Generative models: Hidden Markov Models

Equation (1) can be defined through a generative model, which models $\Pr(W_1^T, U_1^T)$ instead of $\Pr(U_1^T|W_1^T)$, because the optimisation is independent of $\Pr(W_1^T)$. Then, by using the Bayes rule, Equation (1) can be reformulated as:

$$\widehat{\mathcal{U}} = \underset{U_1^T}{\operatorname{argmax}} \Pr(W_1^T, U_1^T) = \underset{U_1^T}{\operatorname{argmax}} \Pr(W_1^T|U_1^T) \Pr(U_1^T) \tag{2}$$

By factorising in terms of turns $t = 1, \ldots, T$ and assuming that words of turn $t$ ($W_t$) do not depend on previous turns words ($W_1^{t-1}$), Equation (2) can be approximated by:

$$\widehat{\mathcal{U}} \approx \underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^{T} \Pr(W_t|U_1^t) \Pr(U_t|U_1^{t-1}) \tag{3}$$

Let $W_t = w_1 w_2 \ldots w_l$ be the word sequence of turn $t$, where $l$ is the sequence length. We can decompose $W_t$ into $r$ segments as $W_t = w_1^l = w_{s_0+1}^{s_1} w_{s_1+1}^{s_2} \ldots w_{s_{r-1}+1}^{s_r}$, where $s = (s_0, s_1, \ldots, s_r)$ is the segment representation and $s_k$ is the index of the segment $k$ of $W_t$. Then, we rewrite the probability distributions of Equation (3) in terms of this 'hidden' segmentation $s$ and the number of segments $r$, by expressing $U_t = u_1^r$ and $W_t = w_{s_0+1}^{s_1} w_{s_1+1}^{s_2} \ldots w_{s_{r-1}+1}^{s_r}$, providing as final expression:

$$\widehat{\mathcal{U}} \approx \underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^{T} \sum_{r, s_1^r} \prod_{k=1}^{r} \Pr(w_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}) \Pr(u_k|u_1^{k-1}, U_1^{t-1}) \tag{4}$$

where it is assumed that words of segments previous to $k$ ($w_1^{s_{k-1}}$) do not affect words in current segment $k$ ($w_{s_{k-1}+1}^{s_k}$).

For the final implementation of Equation (4), some assumptions are taken: (1) the probability of $w_{s_{k-1}+1}^{s_k}$ only depends on the DA $u_k$, (2) the probability of the DA $u_k$ only depends on the $n-1$ previous DA, (3) only the optimal segmentation is taken (i.e., summation on $r, s_1^r$ is changed by a maximisation). Then, the final model is:

$$\widehat{\mathcal{U}} \approx \underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^{T} \max_{r, s_1^r} \prod_{k=1}^{r} \Pr(w_{s_{k-1}+1}^{s_k}|u_k) \Pr(u_k|u_{k-n+1}^{k-1}) \tag{5}$$

In this implementation, $\Pr(w_{s_{k-1}+1}^{s_k}|u_k)$ is modelled by a Hidden Markov Model and $\Pr(u_k|u_{k-n+1}^{k-1})$ by an N-gram of degree $n$.

Estimation of the parameters of this model would be obtained with the training set of annotated dialogues, and a Viterbi process would be applied to unlabelled test dialogues in order to obtain their draft annotation. The experiments presented in Section 4 will study the performance of this model for the available data.

### 2.2 Discriminative models: Conditional Random Fields

Equation (1) can be directly modelled by using discriminative models that provide the estimation of the conditional probability. Following a notation similar to that presented in [16], this can be done by using Linear Chain Conditional Random Fields (CRF):

$$\Pr(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_{\tau=1}^{\mathcal{T}} \exp\left(\sum_{k=1}^{K} \theta_k f_k(y_\tau, y_{\tau-1}, \boldsymbol{x}_\tau)\right) \tag{6}$$

Here, $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output sequences of size $\mathcal{T}$ and $Z(\boldsymbol{x})$ is a normalisation factor that guarantees the definition of a proper probability. $f_k$ $(k = 1, \ldots, K)$ is the set of feature functions that associate inputs and/or outputs and form the actual distribution probability; usually, these functions are known as bigram features, since relates output in $\tau$ and in $\tau - 1$. $\theta_k$ are weights for the $K$ probability distributions of $f_k$.

In the case of dialogue annotation, input sequences are the words of the dialogue $\mathcal{W}$ and output sequences are DA labels $\mathcal{U}$. Thus, following the turn notation of Equation (2), expression in Equation (6) could be adapted by using $\boldsymbol{y} = U_1^T$, $\boldsymbol{x} = W_1^T$, and using productories that decompose the original one in terms of turns and words in turns.

In order to use word based features, a redefinition of the sequence of DA $U_1^T$ is needed. For a given turn $t$, we have the DA sequence $U_t = U = u_1 u_2 \ldots u_{r_t}$ and the word sequence $W_t = W = w_1 w_2 \ldots w_{l_t}$. In this case, we introduce in $U$ a metasymbol $\#$ that represents the empty output for those words $w_i$ that are not final words of a segment (i.e., no DA is attached to that word). Therefore, $U$ gets transformed into $V = v_1 v_2 \ldots v_{l_t}$, with $v_i = \#$ when $w_i$ is not a final word of a segment and $v_i = u_j$ when $w_i$ is the final word of the $j$-th segment of turn $t$.

Consequently, the final CRF model for DA annotation of a dialogue is provided by:

$$\hat{\mathcal{U}} = \underset{U_1^T}{\operatorname{argmax}} \Pr(U_1^T | W_1^T) = \underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^{T} \prod_{i=1}^{l_t} \exp\left(\sum_{k=1}^{K} \theta_k f_k(v_i, v_{i-1}, \boldsymbol{w}_i)\right) \tag{7}$$

Notice that normalisation factor $Z(W_1^T)$ disappears since it does not affect the maximisation. Moreover, words are represented by feature vectors ($\boldsymbol{w}_i$), and therefore more data than only the word could be given as input (e.g., word position, POS-tagging, etc.).

With the selection of a set of feature functions and estimating the weights with training dialogues, a Viterbi process could be applied for obtaining a draft annotation of unlabelled dialogues. Experiments in Section 4 will show the performance of this discriminative model and compare it with the proposed generative model.

## 3 Experimental data

In this section we present the data that we used in the experiments. Two different corpora are introduced: the SwitchBoard corpus [8] and the Dihana corpus [3]. These two corpora present dissimilar and complementary features that allow to check the performance of the models in different situations. The corpora differ respect to the vocabulary size, the set of DA, the nature of the interaction, the semantic restrictions, etc. A summary of their features is presented in Table 1.

### 3.1 SwitchBoard corpus

The SwitchBoard corpus [8] is a popular corpus of telephone conversations in English between two human speakers that discuss about general defined topics. The corpus

**Table 1.** Summary of the features of the SwitchBoard and the Dihana corpus. For Dihana, U means user turns and S system turns.

| Corpus | SwitchBoard | Dihana |
|---|---|---|
| Language | English | Spanish |
| Nature | Human-human | Human-computer |
| Semantically restricted | No | Yes |
| Number of dialogues | 1155 | 900 |
| Number of turns | 115,000 | 6280 U + 9133 S |
| Vocabulary size | 42,000 | 900 |
| Annotation scheme | SWBD-DAMSL | IF-DIHANA |
| Number of DA labels | 42 | 248 |

| Spk | Seg | Transcription | Lab |
|---|---|---|---|
| S1 | S1-1 | Yeah, | aa |
| | S1-2 | to get references and that, | sd |
| | S1-3 | so, but, uh, | % |
| | S1-4 | I don't feel comfortable about leaving my kids in a big day care centre, simply because there's so many kids and so many <sniffing> <throat_clearing> | sd |
| S2 | S2-1 | I think she has problems with that, too. | sd |

**Fig. 1.** Example of SwitchBoard annotation. The meaning of the labels is statement-non-opinion (sd), uninterpretable (%) and agree/accept (aa).

recorded spontaneous speech, with frequent overlaps and interruptions between the speakers, and with several spontaneous speech phenomena (such as hesitations, non-linguistic sounds, etc.) and background noises. A total of 1155 conversations (approximately 115,000 turns) were recorded and manually transcribed. Vocabulary size is about 42,000 words. Total recorded speech signal is about 95 hours.

Switchboard dialogues were annotated with the SWBD-DAMSL [11] scheme, that comprises a total number of 42 different labels. The labels cover communicative functions such as statement, question, etc., but with associated subtypes (e.g., statement-opinion and statement-non-opinion). The manual labelling gave an inter-annotator agreement with a Kappa value of 0.8 [11]. An annotation example is given in Figure 1.

SwitchBoard has become in the last years a sort of standard corpus to evaluate annotation models. This is demonstrated by the different published works that used SwitchBoard to evaluate their proposed models, such as [15, 17, 14]. Some of these previous works do not give details on the corpus preprocessing or experimental framework.

In our case, the preprocessing of the SwitchBoard corpus removed interruptions and overlaps by joining the separated pieces of turns (consequently, all the segments in the turn remain in a single speaker interaction), all the punctuation marks were separated as single words, and all the words were transcribed to lowercase[1].

[1] Corpus available at `www.dsic.upv.es/~cmartine/research/resources.html`.

| Spk | Seg | Transcription | | |
|---|---|---|---|---|
| | | Level 1 | Level 2 | Level 3 |
| S | S1 | Welcome to the railway information system. How may I help you? | | |
| | | Open | Nil | Nil |
| U | U1 | I want to know the departure times from DEP-TW(Valencia) | | |
| | | Que | Dep-h | Org |
| | U2 | to Madrid | | |
| | | Que | Dep-h | Dest |
| | U3 | arriving on DATE(May the 15th of 2004). | | |
| | | Que | Dep-h | Day |
| S | S2 | Do you want to leave on DATE(Sat, May the 15th of 2004)? | | |
| | | Conf | Day | Day |
| U | U4 | Yes. | | |
| | | Accept | Day | Nil |

**Fig. 2.** An excerpt of an annotated dialogue (translated from Spanish into English) from the Dihana corpus. *Nil* denotes the absence of information. Words in capital letters denote categories.

### 3.2 Dihana corpus

The Dihana corpus [3] is a set of 900 telephone dialogues in Spanish between a human user and a Wizard-of-Oz (WoZ) environment [6], in order to simulate a human-computer interaction. These dialogues were directed by the definition of scenarios in a task related to railway information (timetables and fares) for long-distance trains in Spain. The only restrictions during the acquisition were defined by the scenarios, and no other syntactical, lexical, or semantic restriction was applied in the interaction.

The corpus acquired conversations from 225 speakers (153 male and 72 female) that presented small Spanish dialectal variants. The total number of turns in the corpus is 6280 for the users and 9133 for the system, with a vocabulary of approximately 900 words. The total amount of acquired signal is about 5.5 hours. All the dialogues were manually transcribed, including special annotation marks for spontaneous speech phenomena. No overlaps or interruptions were present in the corpus. The spontaneous speech marks were removed for obtaining the final version of the corpus to be annotated.

The set of DA labels that was defined for the annotation of Dihana is an adaptation of the Interchange Format (IF) used for dialogue annotation [7], by adapting its three different levels (speech act, concept, and argument) to the Dihana task [1]. This set is composed of a total of 248 different labels (152 for user turns and 95 for system turns). An annotation example is provided in Figure 2.

In contrast with SwitchBoard, Dihana is not such a sort of standard dialogue corpus, but it presents several interesting features for the implementation of actual dialogue systems: it is a medium-size corpus, task-oriented, with vocabulary limited to the task. In conclusion, Dihana is a useful corpus to complement the conclusions obtained with SwitchBoard and check the models in a real dialogue system framework

The preprocessing of the Dihana corpus includes lowercase transcription, separation of punctuation marks, adding a speaker mark (U for user and S for system) to each word, and a categorisation into task categories such as town names, hours, dates, etc.[2].

[2] Corpus available at `www.dsic.upv.es/~cmartine/research/resources.html`.

| Reference | B | \| | Z | Z | Z | \| | K | K | | K | \| | B | \| | Q | Q | \| | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| System | Z | \| | Z | Z | Z | | Z | Z | \| | B | | B | \| | Q | Q | \| | Error computation |
| *Lenient* | × | | ✓ | ✓ | ✓ | | × | × | | × | | ✓ | | ✓ | ✓ | | 4 Err/10 Ref = 40% |
| *Strict* | × | | × | × | × | | × | × | | × | | × | | ✓ | ✓ | | 8 Err/10 Ref = 80% |
| SegDAER | | $S_1$ | | | | D | | | $S_2$ | | $S_2$ | | C | | | C | (1D + 2S)/(2C + 1D + 2S) = 60% |

**Fig. 3.** An example on how to calculate the different assessment measures. Reference and system show the DA labels present in the reference and given by the system (B, Z, K, and Q represent DA labels, | represents segment limits). In *Lenient* and *Strict* measures, × means error and ✓ correct. In SegDAER, $S_k$ means substitution (two $S_k$ with the same $k$ value represent the same substitution), D deletion, and C correct.

## 4 Experiments and results

In this section we present and compare the results provided by the generative model (HMM-based model) and the discriminative model (CRF) for dialogue annotation.

Assessment measures for the annotation task are the *Lenient* and *Strict* measures (as defined in [2]) and the SegDAER measure:

- *Lenient*: calculates the number of words with incorrect DA label divided by the total number of words.
- *Strict*: calculates the number of words with incorrect DA label or incorrect segmentation divided by the total number of words (the difference with *Lenient* is that takes into account if the word is in the correct segment).
- SegDAER (Segmentation and DA Error Rate) calculates the edition errors between sequences composed of DA labels joined to their position.

Figure 3 presents an example on how calculate the different measures. *Lenient* and *Strict* are standard assessment measures in dialogue annotation. SegDAER is proposed to avoid a feature common to *Lenient* and *Strict*, which is that they consider all the words of the segment affected by the label and the boundaries; however, in an annotation framework, labels can be considered to be at the end of the segment (and they would be interpreted to affect all the words of the segment), and correcting them (in position or in value) would imply edit operations on the label, not on the total sequence of words.

Our experimental framework is based on using a cross-validation approach. In the case of the SwitchBoard corpus, we used 11 partitions of 105 dialogues. In the case of the Dihana corpus, since it presents different features with respect to size, vocabulary, etc., the cross-validation partitions are 5 partitions of 180 dialogues[3].

Since the experiments are for dialogue annotation, input for the models are whole dialogues. Therefore, the search is performed from the first word in the first turn of the dialogue up to the last word of the last turn of the dialogue. Only turn boundaries are supposed to be available and they are taken into account to force output labels at the end of each turn, since we consider that DA labels do not span between different turns. No other information on segmentation is given at the input of the system.

[3] Partitions available at `www.dsic.upv.es/~cmartine/research/resources.html`.

**Table 2.** *Lenient*, *Strict*, and SegDAER SwitchBoard and Dihana results with HMM and different DA $n$-grams, and with CRF. Confidence intervals lower than $\pm 0.2$ in SwitchBoard and than $\pm 0.6$ in Dihana.

| | | HMM 2gr | HMM 3gr | CRF |
|---|---|---|---|---|
| SwitchBoard | *Lenient* | 30.9 | 31.0 | 38.5 |
| | *Strict* | 79.7 | 79.5 | 63.7 |
| | SegDAER | 60.7 | 60.5 | 48.9 |
| Dihana | *Lenient* | 13.2 | 13.1 | 24.4 |
| | *Strict* | 61.4 | 61.5 | 25.0 |
| | SegDAER | 33.7 | 34.5 | 26.5 |

**Table 3.** DAER and SegER results for best HMM and CRF for SwitchBoard and Dihana.

| | DAER | | SegER | |
|---|---|---|---|---|
| | HMM | CRF | HMM | CRF |
| SwitchBoard | 55.8 | 45.9 | 40.7 | 20.3 |
| Dihana | 15.3 | 26.3 | 25.6 | 3.6 |

The HMM experiments tested different values for the degree of the DA $n$-gram (from 2 to 5). The CRF experiments were performed by using CRF++[4] with CRF-L2 algorithm, $C = 1$, and $\eta = 5 \cdot 10^{-3}$; features were the word and whether it is final or not; the template file was that given with CRF++ for the Base-NP task; this template file includes as feature functions unigrams based on the word and on the DA label (spanning from current position up to two positions backward and forward) and a bigram based function. To check the statistical significance, 90% confidence intervals were calculated using bootstrapping with 10,000 repetitions for all the experiments [4].

Experiments produced the results presented in Table 2, where only bigrams and trigrams with HMM are shown since they were the experiments that provided better results for that model. In all cases, CRF presents significantly better results in terms of *Strict* and SegDAER, but results are worse than those of HMM for the *Lenient* measure.

Error analysis was directed to detect the types of errors that the models committed; basically, errors could be caused by wrong DA labels or by wrong segmentation. The metrics for these errors could be derived from SegDAER. When considering only DA labels (which is known as DAER, Dialogue Act Error Rate), errors on wrong DA label assignment are computed. By only considering segmentation boundaries (which is known as SegER, Segmentation Error Rate), errors on wrong segmentation are computed. Results in Table 3 show DAER and SegER for the best HMM model and for the CRF model for both corpora.

As can be seen, in the SwitchBoard corpus both models produce most part of the errors due to wrong DA label assignment; although segmentation errors in this corpus are high as well, they are lower than DA errors. CRF model presents the best errors for the two types of errors, with a dramatic decrease of segmentation errors.

[4] `http://code.google.com/p/crfpp/`

However, in Dihana corpus, which is task oriented and presents a more regular DA sequence structure, errors in the generative model are mainly due to segmentation errors, whereas in the discriminative model DA label assignment errors are dominant. Moreover, the CRF model presents a very accurate segmentation (which could explain the high decrement of the *Strict* measure shown in Table 2), but an error rate in DA label assignment higher than that produced by the HMM-based model.

In general, we can conclude that for the annotation task, CRF discriminative model provides better results than the HMM-based generative model. This conclusion is similar for two corpora with different features, and it can be considered as general enough, although experiments with other corpora would be necessary to confirm it. The difference in the corpora does not affect on the quality of the segmentation that CRF obtains (in any case better than for HMM), but with a high number of different labels, CRF presents an accuracy lower than HMM for assigning the DA label, possibly due to the higher number of parameters that has to estimate.

## 5 Conclusions and future work

In this work we presented a comparison on the annotation of dialogues in terms of DA by using a generative and a discriminative approach. The generative model was based on Hidden Markov Models (HMM) and N-grams. The discriminative model was based on Conditional Random Fields (CRF). Results show that the annotation based on CRF provides better results that any other combination for the *Strict* and SegDAER measures, while for the *Lenient* measure (which is not as sensitive for segmentation errors as *Strict* and SegDAER), HMM provides better results. Anyway, using discriminative models and more specifically CRF seems a better alternative than using generative models. This discriminative model produces better segmentations, but with respect to DA assignment, its accuracy depends on the nature of the corpus and labels (in a general corpus provides less error than the HMM-based model, but for a task-oriented corpus provides a higher error). However, it must be taken into account that HMM were fine tuned with respect to training and decoding parameters (e.g., the Grammar Scale Factor which modules the influence of the N-gram), whereas CRF was used directly without much optimisation of the different parameters and features, using values recommended in the literature.

Future work will be directed to explore the performance of these models on other corpora such like ICSI meeting [10] or AMI [5], which are more difficult since they involve multy party dialogues. Since the presented models show a different behaviour with respect to the kind of errors they produce, its combination would be an interesting research way, in order to obtain a system with the advantages of both types of models. Finally, a more in-depth experimentation with CRF would be desirable in order to tune the different parameters and features involved in the annotation process.

## Acknowledgements

Work partially supported by European Union under project CasMaCat (FP7 - 287576), by Spanish MEC under projects STraDA (TIN2012-37475-C02-01) and Active2Trans

(TIN2012-31723), and by GVA under projects PROMETEOII/2014/030 and AMIIS - ISIC/2012/004.

## References

1. Alcácer, N., Benedí, J.M., Blat, F., Granell, R., Martínez, C.D., Torres, F.: Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In: SPECOM. pp. 583–586 (2005)
2. Ang, J., Liu, Y., Shriberg, E.: Automatic dialog act segmentation and classification in multiparty meetings. In: ICASSP '05. vol. 1, pp. 1061–1064 (2005)
3. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. In: Fifth LREC. pp. 1636–1639. Genova, Italy (2006)
4. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in asr performance evaluation. In: ICASSP'04. vol. 1, pp. 409–412 (2004)
5. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The ami meeting corpus: A pre-announcement. In: Proceedings of MLMI'05. pp. 28–39. Springer-Verlag (2006)
6. Fraser, M., Gilbert, G.: Simulating speech systems. Comp. Speech Lang. 5, 81–99 (1991)
7. Fukada, T., Koll, D., Waibel, A., Tanigaki, K.: Probabilistic dialogue act extraction for concept based multilingual translation systems. In: Proc. of ICSLP. vol. 6, pp. 2771–2774 (1998)
8. Godfrey, J., Holliman, E., McDaniel, J.: Switchboard: Telephone speech corpus for research and development. In: Proc. ICASSP-92. pp. 517–520 (1992)
9. Hastie, H.: Parlance: Probabilistic adaptive real-time learning and natural conversational engine (2012), https://sites. google. com /site /parlanceprojectofficial/home
10. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI Meeting Corpus. In: Proceedings ICASSP '03. vol. 1, pp. I–364–I–367 vol.1 (2003)
11. Jurafsky, D., Shriberg, E., Biasca, D.: Switchboard swbd-damsl shallow- discourse-function annotation coders manual - draft 13. Tech. Rep. 97-01, U. of Colorado Institute of Cognitive Science (1997)
12. Kim, S.N., Cavedon, L., Baldwin, T.: Classifying dialogue acts in multi-party live chats. In: Proc. of the 26th PACLIC. pp. 463–472 (2012)
13. Lemon, O.: The classic project: Computational learning in adaptive systems for spoken conversation (2011), http://www.classic-project.org/
14. Martínez-Hinarejos, C.D., Benedí, J.M., Granell, R.: Statistical framework for a spanish spoken dialogue corpus. Speech Communication 50, 992–1008 (2008)
15. Stolcke, A., Coccaro, N., Bates, R., Taylor, P., van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M.: Dialogue act modelling for automatic tagging and recognition of conversational speech. Computational Linguistics 26(3), 1–34 (2000)
16. Sutton, C., McCallum, A.: An introduction to conditional random fields. Foundations and Trends in Machine Learning 4(4), 267–373 (2012)
17. Webb, N., Wilks, Y.: Error analysis of dialogue act classification. In: Proceedings of the 8th TSD. pp. 451–458 (2005)
18. Wilks, Y.: Companions: Intelligent, persistent, personalised interfaces to the internet. (2006), http://www.companions-project.org

# Processing Annotated TMX Parallel Corpora

Rui Brito[1], José João Almeida[1], and Alberto Simões[2]

[1] Centro de Ciências e Tecnologias da Computação
Universidade do Minho, Braga, Portugal
{ruibrito,jj}@di.uminho.pt
[2] Centro de Estudos Humanísticos
Universidade do Minho, Braga, Portugal
ambs@ilch.uminho.pt

**Abstract.** In the later years the amount of freely available multilingual corpora has grown in an exponential way. Unfortunately the way these corpora are made available is very diverse, ranging from simple text files or specific XML schemas to supposedly standard formats like the XML Corpus Encoding Initiative, the Text Encoding Initiative, or even the Translation Memory Exchange formats. In this document we defend the usage of Translation Memory Exchange documents, but we enrich its structure in order to support the annotation of the documents with different information like lemmas, multi-words or entities. To support the adoption of the proposed formats, we present a set of tools to manipulate the different formats in an agile way.

**Keywords:** parallel corpora, annotated corpora, TMX

## 1 Introduction

Multilingual corpora [7] are very rich resources. They have been used for very different tasks like training machine translation software [5,4], extracting bilingual resources [18,8,19] or information retrieval [9,10].

Unfortunately there is no widely used standard to share parallel corpora at their raw level or with part-of-speech annotation. Some corpora are made available in specific XML formats together with simple programs to process them. Some others are made available in formats like the Text Encoding Initiative (TEI) or the XML Corpus Encoding Initiative (XCES). Unfortunately these standards are not flexible enough for the tasks they are being used, and therefore each user expand and/or interpret the standard by their will [17].

In this document we present a set of extensions to the Translation Memory Exchange (TMX) format to store annotated multilingual corpora. Our main guideline was that these formats should be easy to process using standard XML parsers, following the TMX schema, but not making it awkwardly difficult to parse. Instead of just describing the format, we will show a set of tools ready to process them. These tools are available as Open-Source Software and can be used and bettered by any user.

First, in section 2, we will briefly discuss the available formats for encoding parallel corpora. Then, in section 3, we will detail the *annotated translation memory exchange* (aTMX) and the *partially lemmatized translation memory exchange* (plTMX) formats, including some examples. Follows section 4 that presents the tools used to produce these formats, and section 5 that explains how to use our toolkit to process these formats. Finally, section 6 draws some conclusions and points different evolution directions.

## 2 Parallel Corpora Encoding Formats

There are a few standards to encode parallel corpora. The main problems [17] with these standards are the lack of documentation and evolution:

- The *Text Encoding Initiative* (TEI) is not devoted specifically for this purpose, and its way to encode parallel corpora is not versatile: parallel corpora are usually encoded in two different files, one for each language, and then a mapping file. This makes its processing error prone.
- The *XML Corpora Encoding Standard* (XCES) is outdated, unmaintained and incomplete. There are some researchers that still release their corpora in this format but, as the standard is silent regarding a lot of details, researchers tune the format to their will, making it hard to process.
- The *Translation Memory Exchange* format is quite simple to encode translation memories. As a sentence aligned parallel corpus can be seen as a translation memory this format has been used by some projects to encode parallel corpora. Nevertheless, it does not support, natively, any kind of mark-up to annotate the corpus.
- The *XML Localization Interchange File Format* (XLIFF) is specially used to store software localization translations. Just like TMX, it can be abused to store parallel corpora, but the XML overhead is bigger than using TMX.

Given the status of these formats there are some adaptations, just like ours, to known standards. For example, Forcada [3] proposes an idea similar to our, but extending the TMX tags at their limits. Although this gives extra flexibility to the annotation process, it makes it extremely difficult to keep track of the annotation. Also, the addition of XML tags for each word makes the document huge. Note that if a raw TMX for a parallel corpus can take up to 3 Gigabytes, adding annotations to each word using standard XML tags can make the file 3 to 5 times bigger.

Another Achilles' heel for the wide use of these formats is the lack of tools prepared to their manipulation.

## 3 The Annotated TMX Format

Annotated corpora can be powerful tools for developing and evaluating linguistic theories [6], forging a path for greater linguistic understanding and rigour. Annotations may include structural mark-up, part-of-speech (PoS) tagging, parsing, and numerous other representations.

### 3.1 Basic Format

As discussed previously the simpler formats that are being used are TMX and XLIFF. The first one is more known and therefore, there are more tools that deal correctly with it. This resulted in choosing TMX as the base format for our work. Figure 1 shows a two translation memory excerpt of a TMX file.

```
<tmx version="1.4">
<header creationtool="po2tmx" creationtoolversion="1.9.0"
            segtype="sentence" adminlang="en" srclang="en"/>
<body>
 <tu>
  <tuv xml:lang="EN">
   <seg>Display dialog boxes from shell scripts</seg>
  </tuv>
  <tuv xml:lang="PT">
   <seg>Apresentar caixas de diálogo a partir de scripts de consola</seg>
  </tuv>
 </tu>
 <tu>
  <tuv xml:lang="EN"> <seg>Type your password</seg>    </tuv>
  <tuv xml:lang="PT"> <seg>Introduza a sua senha</seg> </tuv>
 </tu>
</body>
</tmx>
```

**Fig. 1.** Example of a TMX file with two translation units.

The next decision is how to annotate the text inside each one of the TMX translation units. Our main goal when discussing this issue was to reduce the overhead of the annotation. With this in mind, and given that a lot of researchers use the Open Corpus Workbench [2] to encode their corpora we defined the *Annotated Translation Memory Exchange* format (aTMX) as a sort of fusion between the formats of both TMX and CWB. With this fusion we eliminate the need for an XML entry in each text line and another in each tag, making this format very economic. Figure 2 shows the annotated TMX for the translation units shown in Figure 1.

Note that, given the column-oriented approach, where each column represents a layer, it allows the user to add desired level of annotation. The most common columns are *word*, *POS* and *lemma*, as they are the usual output of taggers. Syntactical anotation (treebank-like) can be easily used adding one or more columns for labeled dependency graphs or similar. The same approach is used in the CoNLL[3] data format, MaltParser [11], and others.

[3] CoNLL is the Conference on Computational Natural Language Learning, that often includes shared tasks, where data is made available in a specific format.

```
<tu>
  <tuv xml:lang="en"><seg><![CDATA[  <s>
    Display       display       NN
    dialog        dialog        NN
    boxes         box           NNS
    from          from          IN
    shell         shell         NN
    scripts       script        NNS
  </s>  ]]></seg></tuv>

  <tuv xml:lang="pt"><seg><![CDATA[  <s>
    Apresentar    apresentar    VMN0000
    caixas        caixa         NCCP000
    de            de            SPS00
    diálogo       diálogo       NCMS000
    a             a             SPS00
    partir        partir        VMN0000
    de            de            SPS00
    scripts       scripts       NCMP000
    de            de            SPS00
    consola       consola       NCFS000
  </s>  ]]></seg></tuv>
</tu>
```

**Fig. 2.** Translation unit from a aTMX file.

### 3.2 Region Annotation

One of the big problems with corpora annotation is the way XML forces tags to be properly nested. So, when annotations nest clearly, the proposed approach allows the use of user-defined tags. For example, Figure 3 shows how one can annotate multi-word expressions.

Note that the sentence tags (‘*s*’) and the multi-word expression tags (‘*mwe*’) are inside a CDATA section. This means they will be completely ignored by any TMX parser. But in the other hand, after retrieving the CDATA contents, they can be fed up to a XML parser for further processing. For other types of annotations, that can not be properly nested, different CQP layers (columns) can be used.

## 4 Input Tools

To produce an annotated TMX we need a tool to process the TMX file, and another one to produce annotations for each language segment. To process TMX files we use `XML::TMX` [1], a Perl module that is ready to deal with big TMX files whose Data Object Model (DOM) does not fit into memory. For the annotation we conduced several experiments with two different tools: *Apertium-Tagger* [15] and *FreeLing* [13,14].

```
<tu>
  <tuv xml:lang="en"><seg><![CDATA[  <s>
    <mwe lema="text_view" pos="NP">
      Text      text       NN
      View      view       NN
    </mwe>
  </s>  ]]></seg></tuv>

  <tuv xml:lang="pt"><seg><![CDATA[  <s>
    <mwe lema="vista_de_texto" pos="NP00000">
      Vista     ver        VMP00SF
      de        de         SPS00
      Texto     texto      NCMS000
    </mwe>
  </s>  ]]></seg></tuv>
</tu>
```

**Fig. 3.** Extract from a aTMX file with multi-word annotation.

The approach for each of these taggers is slightly different.

– The API for *FreeLing* is available to be used in Perl [16] which allows to use one or more languages at the same time. Therefore, the TMX is processed one translation unit at a time, where each language is fed to the language tagger (algorithm 1). This approach is useful for any tool that allows the use through an API.
– For the use of *Apertium-Tagger*, the TMX is processed previously, creating two different files, one for each language. These files are processed independently by the tagger, and then joined together in the resulting TMX file (algorithm 2). This approach is useful for external tools that do not export a simple API.

Note that meta-information is stored in the TMX header `prop` elements, like with columns and tags are present in the current file, for each specific language.

## 5 Output Tools

To make a specific format usable by third-parties it is very important to release software that can be used with the formats. In this section we present three tools that process annotated TMX files and produce different type of resources:

– codify the multilingual corpora into Open Corpus Workbench (OCWB);
– produce partially lemmatized translation memories (plTMX);
– extract probabilistic translation dictionaries (PTDs) taking into account words morphological information.

**Algorithm 1:** Tagging process using a library.

```
$langs \leftarrow \text{langs}(TMX)$;
foreach $segment \in TMX$ do
    foreach $l \in langs$ do
        $seg_l \leftarrow \text{select}_l(segment)$;
        $tagged_l \leftarrow \text{tag}_l(seg_l)$;
    $\text{rebuildTU}(tagged)$
```

**Algorithm 2:** Tagging process using an external tool.

```
$langs \leftarrow \text{langs}(TMX)$;
foreach $segment \in TMX$ do
    $id \leftarrow id + 1$;
    foreach $l \in langs$ do
        $seg_l \leftarrow \text{select}_l(segment)$;
        $\text{save}(id, seg_l, file_l)$;
foreach $l \in langs$ do
    $\text{tag}_l(file_l)$;
foreach $id \in IDs$ do
    foreach $l \in langs$ do
        $segment_l \leftarrow \text{fetch}(id, file_l)$;
    $\text{saveTU}(segments)$;
```

### 5.1 Exporting to CWB

The format used to annotate the corpora was taken from the OCWB format. This allows the direct importation of the annotated corpora into it. The Perl module `XML::TMX::CWB`[4] include a method to import a translation memory (being it annotated or not) into OCWB, and allowing the choice to import some specific languages only. This process includes the encoding of each language corpus and then the alignment import for every language pair.

The module also supports the inverse operation, exporting the OCWB into an annotated TMX file.

### 5.2 Computing Lemmatized Dictionaries

One useful resource extracted from multilingual dictionaries are word alignments, like the ones extracted by Giza++ [12] or the Probabilistic Translation Dictionaries (PTD) extracted by NATools [18].

We are specially interested in the probabilistic translation dictionaries. These dictionaries compute relationships between words from the two languages that comprise a parallel corpus. Statistically, it is expected that this relationship maps words from a source language to their translations in a target language. A standard PTD entry is presented in Figure 4.

There are two big problems when computing PTD. The first one is related to certain linguistic constructs, like the use of auxiliary verbs, where the statistical nature of the algorithm will create relationships between the auxiliary verbs. The second problem is related to unbalanced morphology complexity. For example, in Portuguese (and in most of the romance languages) a verb produces easily more than a hundred forms, but in English it will produce just half a dozen. This kind

[4] Available in `https://metacpan.org/pod/XML::TMX::CWB`

$$\textit{imaginar}\begin{cases} \textit{image} & : 57.75\ \% \\ \textit{(none)} & : \ \ 3.99\ \% \\ \textit{imagining} & : \ \ 3.64\ \% \\ \textit{fathom} & : \ \ 3.63\ \% \\ \textit{wondered} & : \ \ 3.18\ \% \\ \textit{picture} & : \ \ 2.74\ \% \\ \textit{imagined} & : \ \ 2.54\ \% \\ \textit{conceive} & : \ \ 1.84\ \% \end{cases} \qquad \textit{imagine}\begin{cases} \textit{imaginar} & : 48.89\ \% \\ \textit{ideia} & : \ \ 4.15\ \% \\ \textit{imagina} & : \ \ 3.85\ \% \\ \textit{suponho} & : \ \ 3.85\ \% \\ \textit{imaginava} & : \ \ 3.79\ \% \\ \textit{imagine} & : \ \ 2.31\ \% \\ \textit{sabia} & : \ \ 1.55\ \% \\ \textit{imagino} & : \ \ 1.53\ \% \end{cases}$$

**Fig. 4.** Two example of entries from a standard PTD, generated from a TMX file. The first example has Portuguese as source language and English as target language. The second example maps English to Portuguese.

of relation will create lots of relations between a single form in English to a lot of Portuguese forms, with each of these relations having a very low probability.

To help in this alignment we can use annotated TMX. Given that this format includes annotations we can take advantage of them to reduce ambiguity and reinforce asymmetrical relations. This can be done at different levels:

- It is possible to use only lemmas. In this situation the huge amount of forms of verbs is not a problem, given they will be all replaced by the infinitive form. This will happen similarly for other word categories.
- Together with the lemmas we can add portions of its part-of-speech. For example, adding a prefix to specify the word category (noun, verb, adverb, adjective, etc), and therefore obtain translations for words when used in different syntactic contexts.
- Also, we can use the idea of partially lemmatized translation memories (that will be discussed in the next section) to obtain a mix of standard and lemmatized PTD.

As an application example consider the construction of a bilingual verbs dictionary bootstrapped by parallel corpora. Consider the following process:

1. Produce an annotated TMX file from a standard TMX file;
2. Collapse each word entry to a token that saves its part-of-speech;
3. Use NATools to extract a pair of probabilistic translation dictionaries;
4. Filter the resulting dictionaries to include only verbs.

The result of applying this process to a literary corpus is shown in Figure 5.

$$\textit{v\_imaginar}\begin{cases} \textit{v\_imagine} & : 44.47\ \% \\ \textit{v\_wonder} & : 10.81\ \% \\ \textit{v\_think} & : \ \ 4.41\ \% \\ \textit{v\_suppose} & : \ \ 0.76\ \% \\ \textit{v\_sense} & : \ \ 0.70\ \% \\ \textit{*v\_have} & : \ \ 0.58\ \% \end{cases} \qquad \textit{v\_imagine}\begin{cases} \textit{v\_imaginar} & : 59.24\ \% \\ \textit{v\_supor} & : \ \ 2.29\ \% \\ \textit{v\_ver} & : \ \ 2.19\ \% \\ \textit{v\_pensar} & : \ \ 2.08\ \% \\ \textit{v\_descobrir} & : \ \ 0.14\ \% \\ \textit{*v\_ir} & : \ \ 0.04\ \% \end{cases}$$

**Fig. 5.** Probabilistic Translation Dictionary of verbs.

In fact, step two of this process can be useful for different tasks. We can collapse each word information in different ways, like above, adding a part-of-speech mark to the lemma. The next section introduces the concept of plTMX, a TMX file whose words are special tokens.

### 5.3 Partial Lemmatized Translation Memories

Sometimes it is useful to convert an annotated TMX to something more simple that can be processed easily as if translation units were traditional sentences, but keeping some morphological information. We tackled this problem defining the concept of *partially lematized* translation memories (plTMX). These translation memories follow exactly the TMX standard, but instead of including simple words, or even the CQP annotation syntax, it includes tokens that mangle together words or lemmas, and some details of part of speech. Figure 6 shows the translation unit from Figure 2 as a partially lemmatized translation unit.

```
<tuv xml:lang="en">
  <seg>v_display adj_dialog n_box from adj_shell n_script</seg>
</tuv>
<tuv xml:lang="pt">
  <seg>v_apresentar n_caixa de n_diálogo p_a_partir_de n_script de
      n_consola</seg>
</tuv>
```

**Fig. 6.** Example of a partially lemmatized translation unit.

In the example verbs, names and adjectives were replaced by the pattern $\langle pos + \_ + lemma \rangle$. Remaining words were kept unchanged. Of course that the way these substitutions are chosen depends highly on the specific purpose of the experiment.

Figure 5 presents a PTD extracted from a plTMX. Compare the result from the previous unprocessed PTD. In this situation we have a quite strong relation between the verb *imaginar* and *imagine*, instead of the several weak relations of all the verb forms. Also note that, removing verb forms gave space to other interesting word to appear.

These resources can be used to bootstrap monolingual verb dictionaries as well. Consider the composition of a PTD, that maps the Portuguese language into the English language, with the PTD that maps the English language into the Portuguese language. This process creates a pseudo-probabilistic set of synonyms. The information associated with *imagine, imaginar* is presented in Figure 7.

Starting form a *standard* aTMX, we can easily produce a plTMX (using the default available converters or adapting them to our necessities) and, based on that, build a set of reusable tools to produce scalable rich bilingual resources.

$$
\mathit{imagine}\left\{\begin{array}{lcr}
\mathit{imagine} & : & 26.76\ \% \\
\mathit{wonder} & : & 6.57\ \% \\
\mathit{think} & : & 3.93\ \% \\
\mathit{see} & : & 1.29\ \% \\
\mathit{suppose} & : & 0.94\ \% \\
\mathit{sense} & : & 0.31\ \% \\
\mathit{*have} & : & 0.26\ \% \\
\mathit{assume} & : & 0.24\ \% \\
\mathit{watch} & : & 0.05\ \% \\
\mathit{imply} & : & 0.05\ \% \\
\mathit{*do} & : & 0.04\ \% \\
\mathit{consider} & : & 0.04\ \% \\
\mathit{look} & : & 0.02\ \% \\
\mathit{find} & : & 0.02\ \% \\
\mathit{discover} & : & 0.02\ \% \\
\mathit{figure} & : & 0.01\ \%
\end{array}\right.
\qquad
\mathit{imaginar}\left\{\begin{array}{lcr}
\mathit{imaginar} & : & 27.22\ \% \\
\mathit{pensar} & : & 6.30\ \% \\
\mathit{supor} & : & 1.13\ \% \\
\mathit{ver} & : & 0.96\ \% \\
\mathit{perguntar} & : & 0.93\ \% \\
\mathit{saber} & : & 0.35\ \% \\
\mathit{sentir} & : & 0.28\ \% \\
\mathit{*ter} & : & 0.19\ \% \\
\mathit{achar} & : & 0.14\ \% \\
\mathit{perceber} & : & 0.08\ \% \\
\mathit{descobrir} & : & 0.07\ \% \\
\mathit{*haver} & : & 0.04\ \% \\
\mathit{pressentir} & : & 0.03\ \% \\
\mathit{calcular} & : & 0.03\ \% \\
\mathit{notar} & : & 0.03\ \% \\
\mathit{*ir} & : & 0.02\ \%
\end{array}\right.
$$

**Fig. 7.** Pseudo-Probabilistic Synonymous Set.

## 6 Conclusion and Future Work

In this paper we defended the necessity of a simple yet versatile format to store multilingual annotated corpora. In order to achieve this we suggested the use of the Translation Memory Exchange format blended with the Open Corpus Workbench column-oriented format. The result allows the annotation of corpora with few overhead of syntactic sugar.

We presented a pair of algorithms using two different approaches, FreeLing-lib based and using external taggers (like Apertium-Tagger), to produce this format.

When processing corpora in annotated TMX format we were able to take advantage of the linguistic information for different objectives, like the extraction of lemmatized probabilistic translation memories. In our experience the use of plTMX proved to be very important and effective because they allow the use of word-based tools over annotated TMX. plTMX recycles annotated sentences back to sentences.

## References

1. Almeida, J.J., Simões, A.: XML::TMX — processamento de memórias de tradução de grandes dimensões. In: XATA 2007 — 5ª Conferência Nacional em XML, Aplicações e Tecnologias Associadas. pp. 83–93 (February 2007)
2. Evert, S., Hardie, A.: Twenty-first century Corpus WorkBench: Updating a query architecture for the new millennium. In: Proceedings of the Corpus Linguistics 2011 conference. University of Birmingham, UK (2011)
3. Forcada, M.: On the annotation of tmx translation memories for advanced leveraging in computer-aided translation. In: LREC'14. Reykjavik, Iceland (may 2014)

4. Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M.: Apertium: A free/open-source platform for rule-based machine translation. Machine Translation 25(2), 127–144 (Jun 2011)
5. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 177–180. ACL, Stroudsburg, PA, USA (2007)
6. de Marneffe, M.C., Potts, C.: Developing linguistic theories using annotated corpora. In: Ide, N., Pustejovsky, J. (eds.) The Handbook of Linguistic Annotation. Springer, Berlin (2014), to appear
7. Melamed, I.: Models of translational equivalence among words. Computational Linguistics 26(2), 221–49 (2000)
8. Morin, E., Prochasson, E.: Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. pp. 27–34. BUCC '11, ACL (2011)
9. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 74–81. SIGIR '99, ACM, New York, NY, USA (1999)
10. Nikoulina, V., Kovachev, B., Lagos, N., Monz, C.: Adaptation of statistical machine translation model for cross-lingual information retrieval in a service context. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 109–119. EACL '12, ACL (2012)
11. Nivre, J., Hall, J., Nilsson, J.: Maltparser: a data-driven parser-generator for dependency parsing. In: Proceedings of LREC-2006 (2006)
12. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
13. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: LREC. pp. 2473–2479. European Language Resources Association (ELRA) (2012)
14. Padró, L.: Analizadores multilingües en FreeLing. Linguamática 3(2), 13–20 (December 2011)
15. Sheikh, Z.M.A.W., Sánchez-Martínez, F.: A trigram part-of-speech tagger for the Apertium free/open-source machine translation platform. In: Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation. pp. 67–74. Universidad de Alicante, Alicante (2009)
16. Simões, A., Carvalho, N.: Desenvolvimento de aplicações em Perl com FreeLing 3. Linguamática 4(2), 87–92 (Dezembro 2012)
17. Simões, A., Fernandes, S.: XML schemas for parallel corpora. In: XATA 2011 — 9$^{a}$ Conferência Nacional em XML, Aplicações e Tecnologias Associadas. pp. 59–69. Vila do Conde, Portugal (1–2 June 2011)
18. Simões, A.M., Almeida, J.J.: NATools – a statistical word aligner workbench. Procesamiento del Lenguaje Natural 31, 217–224 (September 2003)
19. Tiedemann, J.: Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Ph.D. thesis, Uppsala University, Uppsala, Sweden (2003)

# Answering Machine Detector for Call Centers

Jon A. Gómez[1], Luis M. Montes-Novella[2], and José García[2]

[1] Universitat Politècnica de València, València, Spain
jon@dsic.upv.es, http://www.dsic.upv.es/~jon
[2] Tecsible, S.L. C/ Pepe Alba, 1, 46022 València, Spain
lmontesnovella@gmail.com

**Abstract.** This paper presents a system for detecting answering machines in call centers. The task is to distinguish if the audio from the other side of the telephone line belongs to a recording or to a person. The challenge of this system is to make a decision after processing the first three seconds of audio.
The detector system tells the call center whether hanging up or transferring the call to a human agent. Three sources of information are used for making the decision: the time to pick up the phone, the percentage of silence in the three seconds of audio and the recognized word sequence. The system was tested in a real scenario but few calls were used.

**Keywords:** automatic speech recognition, call centers, answering machine detector

## 1 Introduction

Automatic Speech Recognition (ASR) systems are used for several purposes besides the standard task of converting speech to text. Spoken Language Translation is one of the most common applications of ASR where the text is not given as a result to the user [3]. Another common use is as part of Interactive Voice Response (IVR) systems [4]. Asterisk is a software platform for call centers where ASR can be integrated in order to automatically perform usual and repetitive tasks [8, 1]. Asterisk is used as the baseline platform for several applications, namely, call centers, IVR systems, automatic call distributors, desktop dialers, auto-dialers, and more [1].

When used for making outbound calls in auto-dialing tasks such as tele-sales or political campaigns, the automated dialer uses an unfiltered list of phone numbers. So, each outbound call can be answered by a person, a fax or an answering machine reproducing a recorded message. Faxes are easy to be automatically detected, however, distinguishing between a person and an audio message recorded by a person is much more difficult. Non-detected answering machines waste the time of agents.

Several solutions for detecting answering machines exist. LumenVox [7] checks if there is the typical beep after the recorded message. The software of Plivo [10] detects recorded messages by using heuristic algorithms which are based on tone of voice, speed of spoken word, etc. The solution provided by the default answering machine detector in Asterisk [2] analyses the audio signal in order to obtain the amount of silence at the beginning and at the end of the utterance. This solution also counts the number of possible words by searching for transitions between silence and speech. Then, following some knowledge-based rules a decision is made. The company Twilio [11] proposes a solution that consists in looking for the typical pattern when a person answers the phone. A person says *Hello* and then waits for the other party to respond *Hello*. Basically, a sound followed by silence. However, the typical pattern for recorded messages is to continue speaking, this is constant sound with no silences.

In the present work, we present a system that combines several of the abovementioned approaches. Our ASR [5] is used as part of the system whose main goal is to differentiate whether an incoming audio signal via the telephone line corresponds to an answering machine or to a person who is answering the phone call. Our ASR has been integrated into an automatic dialer system based on Asterisk.

Next section describes the system proposed here, section three gives details about the implementation of the detector, section four describes the experimentation and presents the results, finally, conclusions and future works are presented in last section.

## 2 System overview

The system is composed by three modules: the detector, the ASR and the UniMRCP module for Asterisk. This module was adapted in order to provide information to the detector.

The communication between the speech recognizer and Asterisk was implemented by using the Media Resource Control Protocol (MRCP) and the Realtime Transport Protocol (RTP) [9]. RTP for transferring the audio stream from Asterisk to the ASR, and MRCP for exchanging control messages and returning the ASR outcome to Asterisk. The UniMRCP Open Source Project [12] provides a server and a client module that can be integrated in Asterisk for this purpose. Our ASR was adapted to be used as a plug-in for the UniMRCP Server. Several instances of the ASR can be used simultaneously while Asterisk is making calls in parallel. Each instance of the recognizer is used for processing one outbound call.

The detector is an independent module who receives information from Asterisk and the ASR via the modified UniMRCP module. The detector tells Asterisk whether to drop the call or to pass it to an agent by means of a simple message with two possible labels: MACHINE or HUMAN.

Figure 1 shows how the modules are interconnected and the ASR is integrated as a plug-in for the UniMRCP Server. The components of the system (Asterisk, the UniMRCP Server and the detector) run as daemons that are in connection with the auto-dialer server for performing the required tasks. The system operates as follows for every outbound call:

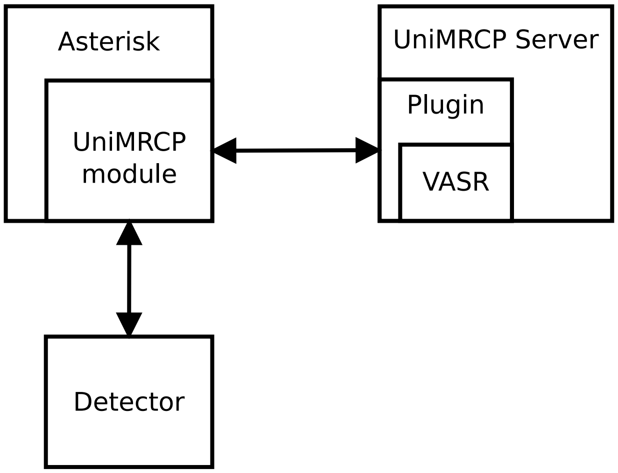


**Fig. 1.** Interconnection scheme of the modules in the AMD

1) The auto-dialer server[3] triggers Asterisk to initiate an outbound call.
2) When the call is picked up Asterisk initiates the communication with the UniMRCP Server for transmitting the audio stream. The message RECOGNIZE is sent from Asterisk to the UniMRCP Server, this message includes several parameters, the relevant one in this case is the `recognition timeout`, which is set to three seconds.
   The time to pick up the phone is then available in the Call Detail Record of Asterisk.
3) The plug-in for the UniMRCP Server stops the recognition process once the timeout of three seconds is over, then sends the message RECOGNITION COMPLETE to Asterisk. This message includes the outcome of the ASR: the recognized word sequence with its score and the percentage of silence.
4) The modified UniMRCP module for Asterisk cuts the audio stream when it receives the message RECOGNITION COMPLETE.
   If the outcome from the ASR delays more than four seconds the call is automatically transferred to an agent and a message for aborting the recognition process and releasing resources is sent to the UniMRCP Server.
   The detector doesn't perform any computation in this case.
5) If the results from the ASR are available on time, then the UniMRCP module for Asterisk sends to the detector the following data: (a) the recognized word sequence with its score, (b) the percentage of silence in the three seconds of audio signal, and (c) the elapsed time in seconds since the phone number was dialed until the call was answered, i.e., the response time.

[3] No information about the private software that uses Asterisk as call center can be provided.

6) With the three sources of information, the detector computes a probability that a human being is answering the call. Different formulas were tested that are going to be explained in the next section.

The message HUMAN is sent to Asterisk if the probability of human is higher than an empirically adjusted threshold, otherwise MACHINE is sent. Results are shown with this threshold set to 0.7, but this threshold can be manually adjusted by users of the system in order to change the behavior of the system; to avoid loosing calls answered by people or, in the opposite direction, to avoid wasting time of agents.

## 3 The Detector

This is the most important module of the system presented here. The ASR is also important but it is a general purpose software for speech recognition that was not developed on purpose for this project. Just few changes were made in the ASR for adapting it to be used in the UniMRCP Server.

As said before, the detector works with three sources of information: the response time, the percentage of silence, and the word sequence. The three data are used for independently computing the probability of human given one of them. Additionally, five different formulas were tested. Each formula combines the three independently computed probabilities in a different way.

### 3.1 Response Time

The response time is the elapsed time in seconds since the phone number was dialed until the call was answered. Asterisk estimates the value for this parameter with a resolution of milliseconds. A simply study was carried out with real data obtained from 79478 true calls labeled automatically as HUMAN or MACHINE by the AMD provided by default in Asterisk [2]. The amount of errors due to this automatic labeling procedure is unknown.

Figure 2 shows the probability that a person is answering a call as a function of the response time. The red line corresponds to the estimation from real data with a resolution of 1 second. The green line represents the approximate function used in the detector. The formula for the approximate function is:

$$Pr(human|responseTime) = 0.1 + \frac{0.86}{1 + e^{-1.25\times(x-4.2)}} - \frac{0.46}{1 + e^{-1.0\times(x-22.5)}}$$

This function was adjusted to fit the probability obtained by using real data, but returns 0.5 for a response time greater than 30 seconds. This decision was made because the number of calls decreases significantly as the response time increases. The values of the probability obtained from real data when the response time gets longer than 25 seconds cannot be considered as representative.

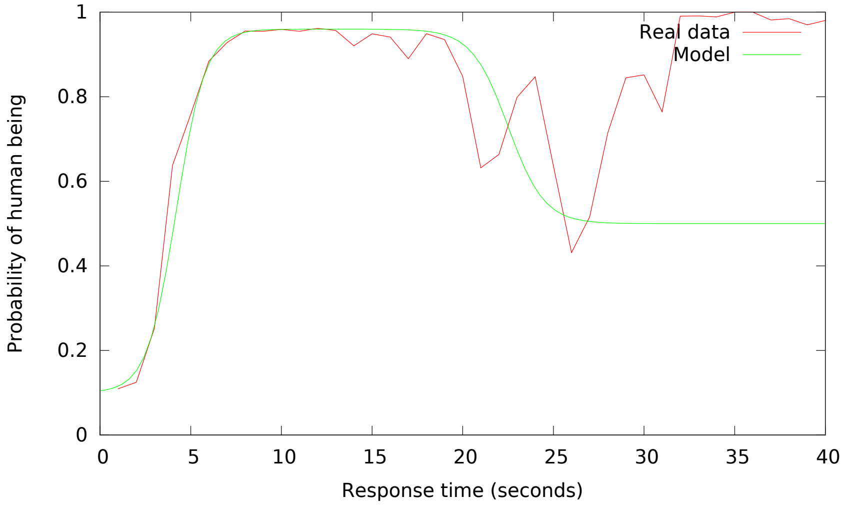


**Fig. 2.** $Pr(human|response\ time)$ from real data and as approximate function

### 3.2 Percentage of Silence

A study was carried out in order to figure out how this parameter could be relevant for determining whether an outbound call is answered by a person or an answering machine with a recorded voice message. As shown in results this is a relevant parameter, with a degree of influence depending on the combination of other parameters. The threshold is empirically adjusted and the $Pr(human|silence)$ is computed as a logistic function.

Figure 3 shows the curve for $Pr(human|silence)$ for a threshold equal to 25%. Usually a recorded message is longer than the answer given by a person, who awaits the caller identifies itself by uttering some words.

### 3.3 Word sequence

The word sequence returned by the ASR after processing the first three seconds of audio is used to distinguish whether it belongs to the beginning of a recorded message or to a person. The ASR is a general purpose LVCSR with a reduced language model, the one corresponding to all the expected sentences, a reduced lexicon just containing the words of the expected sentences, and with an ad-hoc functionality to stop the process of recognition at any moment and retrieve the sequence of words corresponding to the best current hypothesis.

The detector uses two sets of sentences, the set labeled as MACHINE contains typical beginnings of recorded messages. The set labeled as HUMAN contains typical answers of people when they pick up the phone. Each set of sentences is stored in a file on disk that is loaded during the initialization of the detector

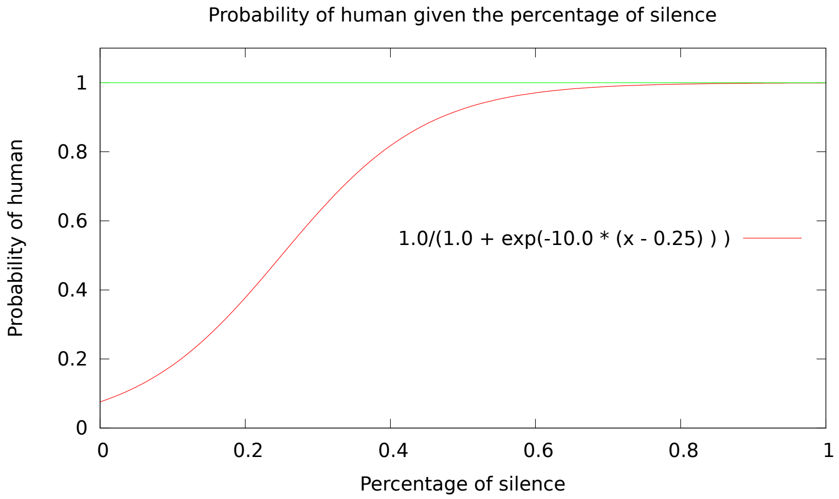


**Fig. 3.** $Pr(human|silence)$ with a threshold in 0.25

daemon. A Deterministic Finite Automaton (DFA) is built for representing each set. We denote as $dfa_H$ the DFA representing the set of sentences corresponding to people and as $dfa_M$ the other one. Both DFA are reduced to their equivalent minimal DFA by using the Hopcroft algorithm for automata minimization [6].

An algorithm was designed and implemented for computing $s(words|dfa)$, i.e., a measure of the similarity that the word sequence *words* was generated by the DFA *dfa*. This algorithm tries to find the best path *pw* through the DFA that best matches the word sequence *words*. The logarithm of the similarity is computed as:

$$\log s(words|dfa) = \frac{-cost}{1 + min(l_{words}, l_{pw})}$$

where $l_{words}$ and $l_{pw}$ are the length measured in characters of *words* and *pw*, respectively. *cost* is the sum of the Levenshtein distance between each pair of words $(w_i, w_j)$, $w_i \in words$ and $w_j \in pw$, that are matched up by the algorithm while searching the best path. Some restrictions were applied to this algorithm in order to avoid delays that can be harmful to the normal operation of the whole system.

Once the similarities are available the probability that a person is answering the call is computed as follows:

$$Pr(human|words) = \frac{s(words|dfa_H)}{s(words|dfa_H) + s(words|dfa_M)}$$

$Pr(human|words)$ is bounded by 0.9 as upper limit and by 0.1 as lower limit in order to do not reduce the effect of the other probabilities in the functions described below. As figure 2 shows, the approximate function for

$Pr(human|responseTime)$ is also bounded, but in this case because the goal is to match the curve estimated by using real data. This is not the case for $Pr(human|silence)$, in the extreme values of this probability the other ones are less relevant.

### 3.4 Probability functions

Five different ways of computing the probability function for discriminating between human and machine were tested. All the functions use a combination of the three estimated probabilities $Pr(human|responseTime)$, $Pr(human|silence)$ and $Pr(human|words)$. Two formulas are the well known arithmetic mean and geometric mean. The other three functions combine the geometric mean of two of the probabilities with the other one in a weighted sum.

$$\begin{aligned} functionSilence &= w_s \times Pr(human|silence) \\ &+ (1 - w_s)\times \sqrt{Pr(human|responseTime) \times Pr(human|words)} \end{aligned}$$

where $w_s$ is the weight for enhancing or attenuating the influence of $Pr(human|silence)$.

$$\begin{aligned} functionWords &= w_w \times Pr(human|words) \\ &+ (1 - w_w)\times \sqrt{Pr(human|responseTime) \times Pr(human|silence)} \end{aligned}$$

where $w_w$ plays the analogous role of $w_s$ in the previous equation.

$$\begin{aligned} functionResponseTime &= w_{rt} \times Pr(human|responseTime) \\ &+ (1 - w_{rt})\times \sqrt{Pr(human|words) \times Pr(human|silence)} \end{aligned}$$

where $w_{rt}$ plays the analogous role of $w_s$ and $w_w$ in the previous equations.

## 4 Results

Results were obtained with the system running in a real scenario but with few calls. Only 100 calls, 79 corresponding to people and 21 corresponding to answering machines. The phone numbers used during the test belong to collaborators who accepted participating in our experimentation.

Figure 4 shows precision versus recall for different values of $w_s$ when using $functionSilence$ for discriminating between human and machine while varying the threshold for the percentage of silence. $w_s$ is the weight for enhancing or attenuating the relevance of $Pr(human|silence)$. Figures 5 and 6 show the same representation with respect to functions $functionWords$ and $functionResponseTime$, for different values of the corresponding weights, $w_w$ and $w_{tr}$. It can be observed that for these three functions there exist several combinations of the parameters for which the system has an almost optimum performance.

Figure 7 show the evolution of precision and recall versus the threshold for the percentage of silence. Both figures reveal that this threshold should be lower than 40%. It can also be observed that these functions have a performance that is somewhat inferior to the performance observed for the other three functions.

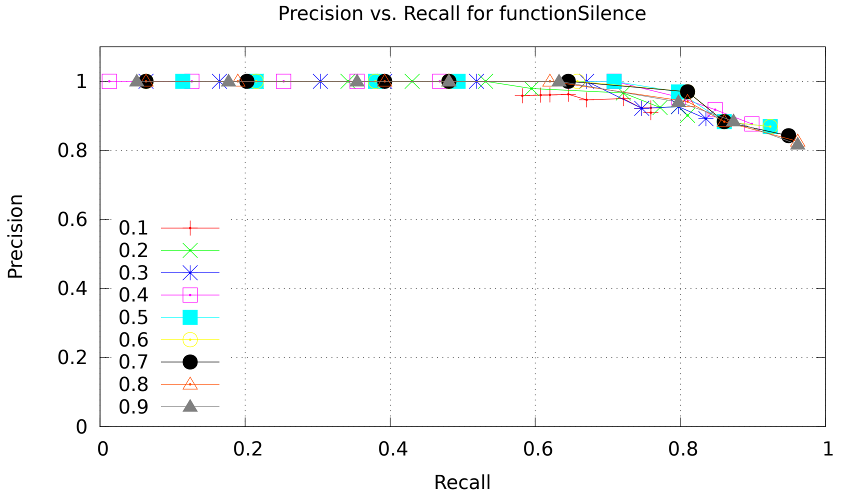


**Fig. 4.** Precision vs. Recall when the detector uses the $functionSilence$ for discriminating between human and machine

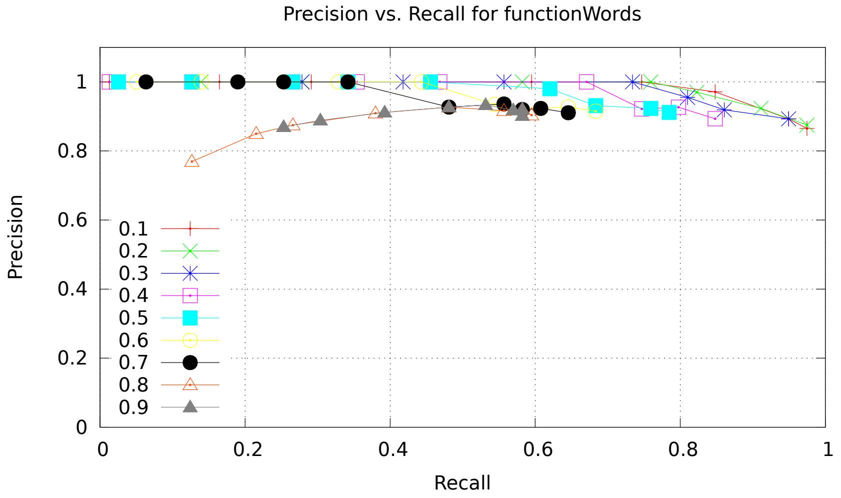


**Fig. 5.** Precision vs. Recall when the detector uses the $functionWords$ for discriminating between human and machine

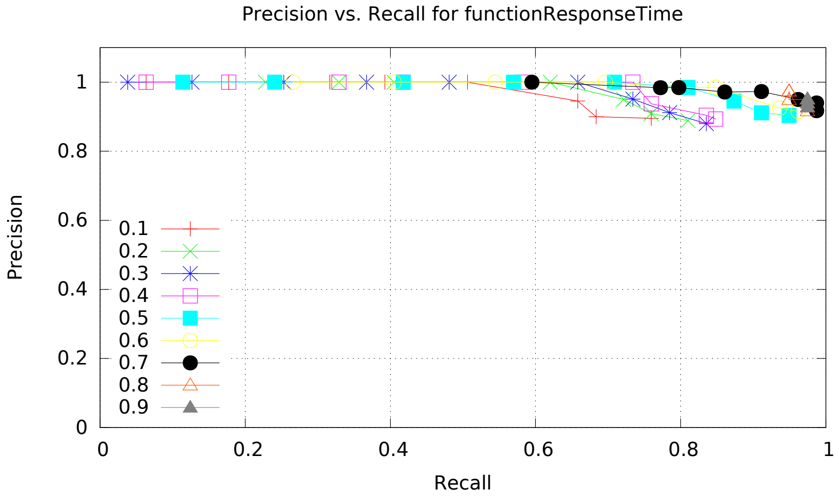


**Fig. 6.** Precision vs. Recall when the detector uses the *functionResponseTime* for discriminating between human and machine

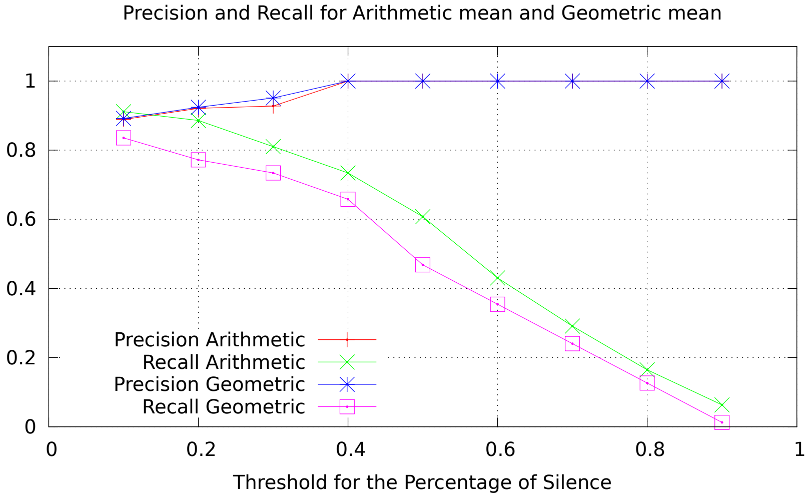


**Fig. 7.** Precision and Recall versus the threshold for the percentage of silence when the detector uses the *arithmeticMean* or the *geometricMean* function for discriminating between human and machine

## 5 Conclusions and Future Work

The results show a good performance of the system for detecting answering machines in tasks where outbound phone calls are initiated continuously. As they can be tele-sales or political campaigns.

Despite there are several combinations of the adjustable parameters where both precision and recall take high values, these results are preliminary because of the small number of calls. We must check the behavior of the answering machine detector presented here by using a larger number of outbound calls. This will allow us to better adjust all the parameters and detect possible errors. We used a rule-based approach because of the lack of enough samples for training. Any classifier will be a better solution if more samples are available.

Anyway, the development of the different parts of the system and their integration into standard software as Asterisk and UniMRCP open us multiple choices for testing our ASR in real and challenging scenarios.

## Acknowledgments

This work was partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01.

## References

1. Asterisk, http://www.asterisk.org/get-started/applications/call-center
2. Asterisk AMD, http://svn.asterisk.org/svn/asterisk/trunk/apps/app_amd.c
3. Casacuberta, F., Federico, M., Ney, H., Vidal, E.: Recent Efforts in Spoken Language Translation. IEEE Signal Processing Magazine 25(3), 80–88 (2008)
4. Corkrey, R., Parkinson, L.: Interactive voice response: Review of studies 1989-2000. Behavior Research Methods, Instruments, & Computers 34(3), 342–353 (2002), http://dx.doi.org/10.3758/BF03195462
5. Gómez, J.A., Sanchis, E.: Using Word Graphs as Intermediate Representation of Uttered Sentences. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, vol. LNCS 7441, pp. 284–291. Springer (2012)
6. Hopcroft, J.E.: An n log n algorithm for minimizing states in a finite automaton. In: Kohavi, Z., Paz, A. (eds.) Theory of Machines and Computations, pp. 189–196. Academic Press (1971)
7. LumenVox, http://www.lumenvox.com/products/speech_engine/amd.aspx
8. Megelen, J.V., Madsen, L., Smith, J.: Asterisk: The Future of Telephony. O'Reilly Media, Inc., 1005, Gravenstein Highway North, Sebastopol, CA 95472, second edn. (August 2007)
9. Media Resource Control Protocol v2, http://tools.ietf.org/html/rfc6787
10. Plivo, https://www.plivo.com/faq/voice/how-does-plivos-amd-answering-machine-detection-work
11. Twilio, http://www.twilio.com/help/faq/voice/can-twilio-tell-whether-a-call-was-answered-by-a-human-or-machine
12. Uni MRCP Open Source Project, http://www.unimrcp.org

# Obtaining parallel corpora for Multilingual Spoken Language Understanding tasks

Fernando García, Marcos Calvo, Emilio Sanchis,
Lluís-F. Hurtado, Encarna Segarra

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
{fgarcia,mcalvo,esanchis,lhurtado,esegarra}@dsic.upv.es

**Abstract.** Many Spoken Language Understanding systems are based on statistical methods like Stochastic Finite State Automata and Classification Techniques. Although many efforts have been made to develop semi-supervised and unsupervised learning techniques for semantic modeling, almost all of the statistical methods are supervised, hence, it is necessary to have a semantically labeled training corpus in order to learn the models. The multilingual approaches to Spoken Language Understanding can be grouped in two classes, train-on-target and test-on-source. In both approaches the translation of the corpus from the original language to other languages is needed. In this work, we present how to obtain a translated corpus from the original one to estimate the new Spoken Language Understanding system and how to obtain a parallel corpus to estimate a task-specific machine translation system. We also present a methodology to translate not only the sentences, but also the semantic labels and the segmentation of the corpus. Finally we present some preliminary experiments using both approaches.

**Keywords:** Multilingual Spoken Language Understanding, Parallel corpora, Machine translation.

## 1 Introduction

Nowadays, automatic language portability of different speech-based systems is an important challenge. Finding a efficient solution to this problem would imply saving a lot of time, money and effort in translating and labeling corpora, as well as adapting and re-training the systems that already work for some language. Also, having multilingual systems available, allows the study of the behavior and robustness of the speech-based systems when the language changes. This would make it possible to identify the strengths and weaknesses of the models and the systems depending on the input language, which would probably lead to some interesting language-dependent improvements. Although the goal of making speech-driven systems for many languages in a totally unsupervised manner seems to be still quite far, Machine Translation (MT) technology can help us to address this problem in a semi-supervised way.

One kind of these speech-based systems is limited-domain spoken dialog systems, which in the last few years has received many efforts from the research community. They try to provide a fluent speech-based interaction between a human and a computer, in the context of a well defined task. An important part of these systems is the Spoken Language Understanding (SLU) component. Its aim is to provide a semantic interpretation of the input sentence in terms of some semantic units (or concepts), and to identify the relevant information (or values) that are attached to each one of them. The semantic units are defined beforehand according to the nature of the task and represent both the user intention and the different types of pieces of information that are expected to be provided to the system. Many SLU systems are based on statistical methods like Stochastic Finite State Automata and Classification Techniques [7, 9, 5, 6, 3]. Although many efforts have been made to develop semi-supervised and unsupervised learning techniques for semantic modelization [10, 8], almost all the statistical methods are supervised, hence, it is necessary to have a semantically labeled training corpus in order to learn the models.

The multilingual approaches to SLU can be grouped in two classes, so-called train-on-target and test-on-source. In the train-on-target approach a new SLU model is trained in the target language, that is the language in which the user utterances are pronounced. To do this, it is necessary to translate the training corpus from the original language to this new language, and to learn the corresponding SLU models. Once we have a model in this target language the understanding process can be solved as in the monolingual SLU, because the input utterance and the models are in the same language. This hand-based translating process requires a lot of effort, which makes it very interesting to develop a methodology to perform this step automatically.

In the test-on source approach the input sentences are translated into the original language of the corpus, therefore, the understanding process is performed always in this language. This means that the SLU module should be fed by a translation module that translates the input utterances, which implies that the SLU models are not needed to be in more than one language. Thus, it is very important here to develop a good-performance task-specific MT system, as mistakes during the translation process can produce many errors in the SLU output.

In this work, we present how to obtain two parallel corpora French-Spanish and English-Spanish, from an original corpus in Spanish. We also present a methodology to translate not only the sentences, but also the semantic labels and the segmentation of the corpus. Finally, we present some experimental results that show the behavior of the techniques of translating the corpus in both cases.

## 2 The multilingual DIHANA corpus.

The multilingual DIHANA corpus that we have developed is based on the original Spanish DIHANA corpus. This is a set of 900 dialogs in Spanish in a telephone-based information service for trains. The corpus was acquired using the Wizard of Oz technique. Three scenarios were defined and posed to the speakers:

– In the first scenario the aim of the user is to obtain the timetables for a one-way trip.
– In the second scenario the users were told to obtain the price of the tickets, and optionally the timetables, of one-way trains.
– The third scenario was analogous to the second one, but considering a round trip.

The corpus has a total of 10.8 hours of speech uttered by 225 different speakers.

In order to use this corpus for SLU tasks, a semantic labeling was performed. 30 semantic labels were defined, and all the user turns were manually and completely segmented and labeled in terms of these labels. The labeling process, as well as the definition of the set of semantic labels itself, were developed in such a way that each sentence is associated to a sequence of semantic labels and a segmentation of it in terms of these labels (one semantic label per segment). For example, the sentence in Spanish ”Me podría decir los horarios para Barcelona este jueves?” (Could you tell me the timetables to go to Barcelona next Thursday?) would be segmented this way (the special symbols <> denote a question about the concept that is between the symbols):

```
me podría decir : courtesy
los horarios de trenes: <time>
para Barcelona : destination_city
este jueves: date
```

Some characteristics of the semantically labeled corpus are shown in Table 1.

**Table 1.** Characteristics of the semantically labeled corpus.

| | |
|---|---|
| Number of user turns: | 6,229 |
| Total number of words: | 47,222 |
| Vocabulary size: | 811 |
| Average number of words per user turn: | 7.6 |
| Total number of semantic segments: | 18,588 |
| Average number of words per semantic segment: | 2.5 |
| Average number of segments per user turn: | 3.0 |
| Average number of samples per semantic unit: | 599.6 |

The corpus was split into a training set of 4,887 turns and a test set of 1,340 turns.

## 3 Translating the original corpus.

We have translated this corpus into English and French following different procedures for the training and the test sets. For the training set, the process was based on the combination of the output from several web open-domain translators. This decision was made since for SLU purposes some errors in the translations can dramatically spoil the behavior of the system. For example, mistranslating any keyword strongly related to the semantic meaning, or even any polysemic word that could be translated erroneously by the translator using any other of its meanings, can severely damage the whole meaning of the sentence. If several hypotheses are generated by different translators there are more possibilities that the correct translation appear in one of the translated sentences.

Furthermore, as our aim is to work with limited-domain tasks, it would be desirable to have a task-dependent translation system that makes the minimum semantically important mistakes. Unfortunately, we are not able to train a task-dependent machine translation system, as the original corpus is just monolingual. Therefore, our option was to use web open-domain translators. However, as open-domain translators usually make many errors, our proposal is to obtain several outputs from several translators, and provide them all in our new multilingual corpus. Then, statistical SLU models using this variability can be trained, and a proper combination of these translations may improve the quality of the individual translations [1]. Hence, an improvement of the coverage and the overall quality of the multilingual SLU system could be achieved through this combination.

The training corpus translated can be used in different ways, being two of them the following ones:

- To learn a new model in the target language, from the translated training corpus and their segmentation/semantic-labeling associated.
- To learn an in-domain machine translation system from the pairs of sentences generated by the web open-domain translators.

To obtain these multilingual training corpus we have developed three kinds of corpus translations, all of them based on several open-domain web translators (4 translators were used for French, and another set of 4 web translators for English) as it is shown in Figure 1.

- First corpus translation: In this case the sentences are translated to the other language using the different translators. This way we have several hypotheses for each sentence. For example:

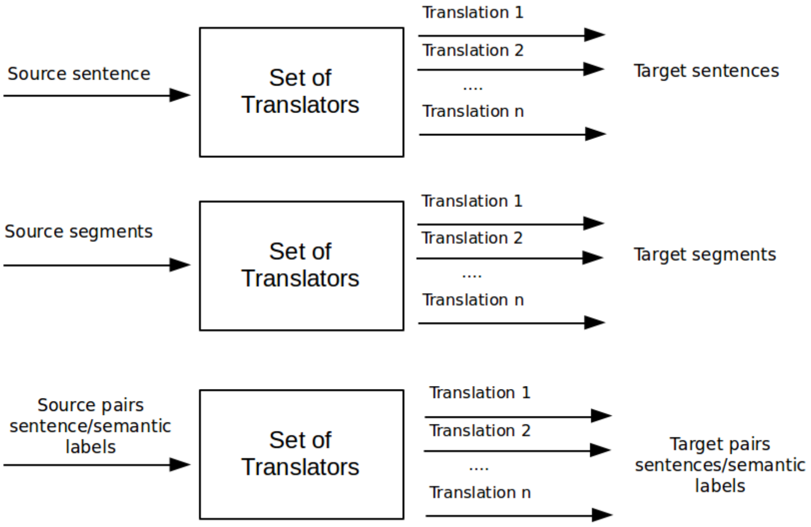


**Fig. 1.** Translations with open-domain web translators.

Spanish:
`Quisiera horarios para el próximo sábado a Barcelona por la mañana`

French:
```
Serait planifier pour le samedi matin à Barcelona
J'aimerais planifier pour samedi prochain à Barcelona le matin
Il voult horaires pour le prochain samedi à Barcelona par le matin
Il voudrait des horaires pour samedi prochain à Barcelona le matin
```

English:
```
I would like schedule for Saturday morning to Barcelona
I would like to schedule for next Saturday to Barcelona in the morning
It wanted schedules for next Saturday to Barcelona in the morning
Would want schedules for next Saturday to Barcelona in the morning
```

- Second corpus translation: Also each semantically labeled sequence of words was translated on its own. This way, we also obtained the translations of each semantically relevant piece (the lexicalization of each concept), as well as the segmentation and labeling of the corpus in the new language, without any manual effort. All this information makes it possible to learn a statistical semantic model for the task in the new language. For example: `por la tarde, à l'après-midi, in the afternoon ... o cuánto cuesta, quel est le prix, how much does it cost, ....`
- Third corpus translation: Considering that the translation of short segments can generate more errors than translations of complete sentences due to the

fact that they have not information about context, we have performed a translation of the complete sentences adapting the segmentation and semantic labeling to the translated sentences.
The adaptation of the semantic segmentation and labelling is performed in the following way: as we know the sequence of semantic units we can build a sentence as the concatenation of segments of words corresponding to the semantic units. Then, we obtain the best alignment between this artificial sentence and the sentence generated by the translator. This alignment associates a segmentation to the translated sentence. In this approach we assumed that the sequence of semantic units is the same in both languages.[1]

| | | | |
|---|---|---|---|
| **Quisiera horarios** | **para el próximo sábado** | **a Barcelona** | **por la mañana** |
| (Departure-Time) | (Date) | (Destination-City) | (Hour) |
| **Serait planifier** | **pour le samedi** | **matin** | **à Barcelona** |
| (Departure-Time) | (Date) | (Hour) | (Destination-City) |
| **I would like schedule** | **for Saturday** | **morning** | **to Barcelona** |
| (Departure-Time) | (Date) | (Hour) | (Destination-City) |

Regarding the test set, it was manually translated into French and English and uttered by native speakers. This way, we simulate a real scenario in which the native speaker interacts with the system using their own language. The test set in French is composed by 1,277 user turns, 500 of which were uttered by 4 native speakers. The test set in English, obtained in the same way, consists of 1,336 turns, and all of them were uttered by a total of 6 native speakers.

## 4 Experimental results for SLU

In order to study the quality of this approach, we have performed some preliminary experiments with a statistical SLU model using a test-on-source approach [2]. These experiments were performed considering as test set the French utterances. We trained a Statistical Machine Translator (MOSES) using the parallel French - Spanish training part of this new multilingual corpus (Figure 2). Then, the n-best translations for the recognized utterances were obtained by using this translator. Weighted graphs of words were built from these hypotheses, and a specific SLU decoding method for these structures was developed. The result in terms of Concept Error Rate was 22.40%, which is not too far from the result for the same test set considering the original Spanish utterances (17.72%).

We also studied the behavior of the systems when translating the training corpus to learn models in the new language, using a train-on-target approach. This study was done before the acquisition of this multilingual DIHANA corpus, and it was evaluated by using the French MEDIA corpus. In this previous work [4], we also explored the possibility of using both Conditional Random Fields (CRFs) and Stochastic Finite State Automata (SFSA) for the semantic modelization using a translated training corpus. The Concept Error Rate

[1] If this was not true, the forced alignment could be done between the sentence and an ergodic model of concatenation of the segments.

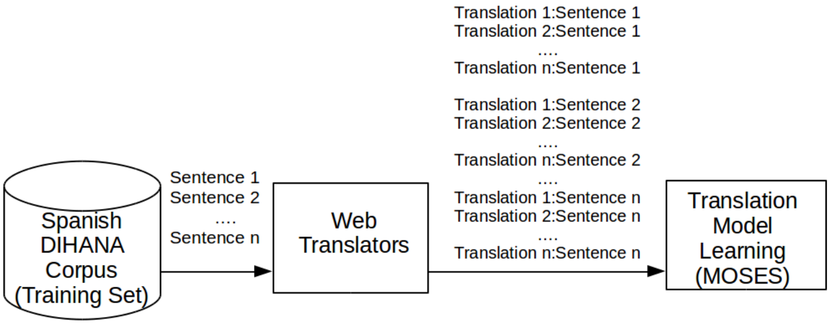


**Fig. 2.** The creation of a parallel corpus to estimate the MT system.

obtained from the experiments was 23,4% for CRFs and 25,2% for the SFSA. Results showed that it is possible to obtain accurate models by using this kind of translation process.

## 5 Conclusions

We have presented an approach to translate semantically labeled corpora to other languages, in order to build a SLU system in different languages. It includes a method to translate not only the sentences, but also the semantic labels and the segmentation of the corpus. We have also presented an approach to obtain a parallel corpus to estimate a task-specific machine translation system. Some baseline experiments are presented in order to show the capability of this approach to obtain accurate multilingual SLU systems, in both cases, in a test-on-source approach and in a train-on-target approach.

**Acknowledgements.** This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, and under FPU Grant AP2010-4193.

## References

1. Bangalore, S., Bordel, G., Riccardi, G.: Computing Consensus Translation from Multiple Machine Translation Systems. In: In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU-2001. pp. 351–354 (2001)
2. Calvo, M., Garcıa, F., Hurtado, L.F., Jiménez, S., Sanchis, E.: Exploiting multiple hypotheses for multilingual spoken language understanding. CoNLL-2013 pp. 193–201 (2013)
3. De Mori, R., Bechet, F., Hakkani-Tür, D., McTear, M., Riccardi, G., Tür, G.: Spoken language understanding: A survey. IEEE Signal Processing magazine 25(3), 50–58 (2008)

4. García, F., Hurtado, L., Segarra, E., Sanchis, E., Riccardi, G.: Combining multiple translation systems for Spoken Language Understanding portability. In: Proc. of IEEE Workshop on Spoken Language Technology (SLT 2012). pp. 282–289. Miami (2012)
5. He, Y., Young, S.: Spoken language understanding using the hidden vector state model. Speech Communication 48, 262–275 (2006)
6. Lefèvre, F.: Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. vol. 4, pp. 13–16. IEEE (2007)
7. Maynard, H.B., Lefèvre, F.: Investigating Stochastic Speech Understanding. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2001)
8. Ortega, L., Galiano, I., Hurtado, L.F., Sanchis, E., Segarra, E.: A statistical segment-based approach for spoken language understanding. In: Proc. of InterSpeech 2010. pp. 1836–1839. Makuhari, Chiba, Japan (2010)
9. Segarra, E., Sanchis, E., Galiano, M., García, F., Hurtado, L.: Extracting Semantic Information Through Automatic Learning Techniques. IJPRAI 16(3), 301–307 (2002)
10. Tür, G., Hakkani-Tür, D., Schapire, R.E.: Combining active and semi-supervised learning for spoken language understanding. In: Speech Communication. vol. 45, pp. 171–186 (2005)

# Tracking Emotional Hints in Spoken Interaction

Raquel Justo, Ohiane Horno, Manex Serras, M. Inés Torres

University of the Basque Country. Sarriena s/n 48940 Leioa. Spain
{raquel.justo,manes.torres}@ehu.es

**Abstract.** The main goal of this work is to design a preliminary approach for a speech based emotion detection system. This system should work in a call-center that provides client assistance in different fields. Thus, the challenge of dealing with real users' emotion and spontaneous Spanish speech has to be faced. Different parameters and techniques, successfully employed in other areas, are validated and compared to each other. For doing it, a set of real data recorded in the call center was considered and a qualitative analysis of the parameters was carried out. Additionally, different versions of a classification algorithm was also used to compare the different approaches quantitatively.

**Keywords:** Emotion detection, Speech, Dialogues

## 1 Introduction

Speech is the most natural way in which humans communicate with each other. Indeed, speech signal contains different kind of information related to the message, speaker, language,... When speakers interact in a natural way, their emotional state is also reflected in the manner in which they speak. Being able to detect this emotional state would be very helpful to improve Human Machine Interaction (HMI) because it would provide the systems with the capability to adapt themselves to the user status. Thus, emotion recognition has become a very popular researching area in the last years [16, 14, 1, 4].

The problem of emotion recognition can be viewed as a classical classification problem, where each class corresponds to a specific emotion. However, there are some issues that make it hard to carry out the emotion detection in speech. Firstly, when considering human-human interaction, the emotional state is shown as a combination of multimodal signals like facial expression, movements, gestures, speech,... Therefore, there is a significant loss of information associated to the emotional speech detection problem. Secondly, the decision of which emotions are going to be detected is not obvious because the human detection of emotions is subjective. Thus, it is not clear which is the best classification of the different emotions that are involved in human interactions. Additionally, there are not reference databases to learn robust models and to be used for comparison purposes between different approaches. Moreover, most of the existing corpora are built using experienced actors instead of real users [10, 5, 4]. For instance, actors usually overact while real users represent emotions in a milder way.

Moreover, an actor will always represent the same emotion in a similar way, for instance anger is usually simulated shouting, whereas humans may express anger in a wide range of ways (speaking quickly, using clipped speech,...). Additionally, these databases are often based on read speech, not spoken. In consequence, the extracted conclusions cannot be interpolated to real situations. Finally, emotion recognition systems are influenced by both speaker and language. Indeed, language, is closely related to cultural features that have a great influence in the way in which people express their emotions. For instance, Spanish speakers will probably express anger more loudly than Japanese ones.

There is a wide range of disciplines where emotion detection in speech might be useful such as health informatics, games, education and all the systems in which the adaptation to the user is relevant. In this work, we deal with the specific case of a call-center. The services offered by these companies are related to user assistance in different fields. The user experience is very important in these cases and companies invest significant effort in improving it [17, 3]. Detecting when the user is getting angry and when he/she is happy with the service could be very useful for the improvement of call-center services.

The aim of this work is to carry out a preliminary analysis of the design of an emotion recognition system in Spanish for a specific call-center. The company is a SME located in the Basque Country and it is devoted to the design and management of projects related to client assistance, customer loyalty, sales, post-sales,... In order to improve their services they wanted to detect the users' emotions and change the operators acting protocols accordingly.

The contribution of this work is related to the validation of emotion recognition parameters and classification algorithms, successfully employed with other applications and languages, in this specific case. Let us note that specific features such as, spontaneous speech in Spanish associated to real dialogues with real users' emotions have to be considered. As far as we know, there is no other work in the literature considering all of these features together. For instance, *Ahumada* database [11] consists of real dialogues in Spanish but it was designed for speaker verifications. There are also other works considering real users' emotions [7, 15] but they try to detect polarity in English. In this work, real data, obtained from different telephone calls recorded in the call-center, were used. The acquired speech signals were then segmented and labeled according to the corresponding emotions. The signals were also analyzed and different feature sets were extracted and compared to each other. Finally, different versions of a classification algorithm were considered for comparison purposes.

This paper is organized as follows: Sec. 2 describes the specific data employed in the analysis. In Sec. 3 the analysis of speech signals was reported, different features were considered and compared to each other qualitatively. Sec. 4 describes the employed classification algorithm and analyzes the differences between different versions. Finallly, Sec. 5 summarizes the extracted conclusions and further work.

| | *Happiness 1* | *Anger 1* | *Anger 2* |
|---|---|---|---|
| Classification goal | Happiness vs. Neutral | Anger vs. Neutral | Anger vs. Neutral |
| Language | Spanish | | |
| Length | 1 min. 44 sec. | 6 min. 50 sec. | 2 min. 19 sec. |
| # Annotated Segments | 43 | 137 | 42 |
| Neutral / Not-Neutral | 36/7 | 34/103 | 9/33 |

**Table 1.** Features of the different audio data files recorded for the preliminary analysis.

## 2 Data

Different telephone calls were recorded by the company for the preliminary design of the emotion recognition system. All together 3 audio files, corresponding to 3 different speakers, were considered. The features corresponding to the employed data are summarized in Table 1. In the *Happiness* call a woman asks for support to change from radio mode to television mode and she expresses her happiness when she succeeded. In the second call, a man, who is in a work travel in China, is complaining about the problems he is having with his mobile phone line. In the third call, anger is also detected. A man is having problems, once and again, with the Internet connection at home. *Anger 1* and *Anger 2* correspond to very different ways of being angry. Whereas the first one is more aggressive and ruder, the second one is more relaxed and calmer, something closer to boredom. Note that, even when only some example data files were considered, different ways of expressing the same emotion appear in real users applications.

## 3 Analyzing Different Features

Prosodic features are good indicators of high arousal emotions, as anger and happiness [12, 9, 17, 1, 6]. Based on this strong relationship and considering that they can be extracted very easily using a specific toolkit, *Praat* [2]), we decided to analyze the following features: 1) *Intensity*: the volume of the voice, 2) *Pitch*: the vibration frequency of the vocal folds and 3) *Formants*: resonances of the vocal and nasal track shown in the spectrogram of the signal. Additionally the qualitative analysis of the duration and number of pauses and speech speed will be carried out.

### 3.1 Qualitative Analysis of Features

In this section, the changes in the features were analyzed throughout the whole signal. Let us note that, given the preliminary nature of this work and the absence of different emotions associated to each speaker, our goal was limited to the speaker dependent discrimination of neutral vs. happiness/anger signals.

***Happiness1*** The overall tone of the call is neutral and it has very few segments labelled as happy (7/43) as shown in Table 1. Thus, Pitch and Intensity are more or less constant throughout all the signal. The fluctuations that appear in the

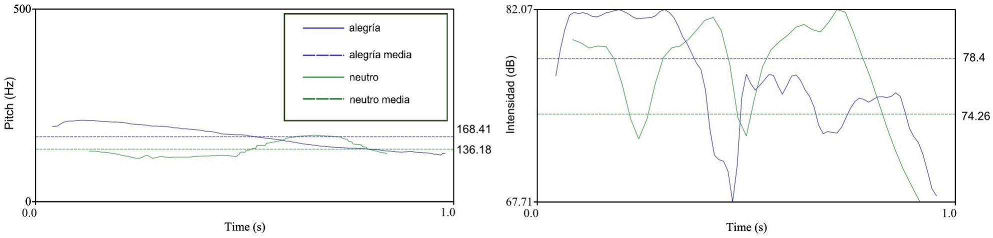


**Fig. 1.** Distribution of the pitch and the intensity in happy labeled frame ' *ole*' and neutral labeled frame '*mira*'.

Intensity are related to the small silences between two words. The speech speed is medium-fast around 17-22 phonemes per second and there are small pauses between segments that contain 6-12 words.

The difference of the features behaviour between neutral and happy segments is notorious. As predicted in the theory the values of both intensity and pitch increase in the segments labelled as happy, as can be seen in Fig. 1. The solid blue and green lines stand for the neutral and the happy states respectively and the dashed lines represent their mean values. The two segments have been selected in order to consider one word and a similar length.

Additionally, Fig. 1 shows a tendency to lengthen and separate the syllables in the segments labelled as happy. The intensity value decreases suddenly at the end of the syllables and the pitch value variation between two syllables is also notorious. This behaviour is also appreciable in the speed, it decreases to 3-8 phonemes per second. This is an interesting behaviour, since several studies showed how the speech speed increases in emotions associated to a high excitement. Nevertheless, happy labelled segments in our signal, consist of sentences with less than 3 words and there are more interjections than full sentences. This aspects might correspond to a different way of expressing happiness that is not related to an increase of the speed.

***Anger1*** The global tone of the call is angry with few neutral segments that due to the serious tone of the speaker are hard to distinguish from the moderate angry ones. The behaviour of the features in the neutral segments is quite different when comparing it with the previous call. Although the value of the Pitch is also constant, it stops abruptly at the end of the words and syllables in many cases. The value of the Intensity diminishes at the end of the syllables rather than at the end of the words. This fact, together with the pitch's tendency, makes the speaker sound serious. The speech is slower in this case (11-16 phonemes per second) and with more pauses with less duration ($\leq$ half a second).

Regarding the segments labelled as angry, there is no visible difference in the Intensity but for the slightly higher mean value. In the Pitch, the abrupt stops at the end of the words and syllables are also observed (see Fig. 2), but in this case, this behaviour is amplified. The words stops abruptly more often and the

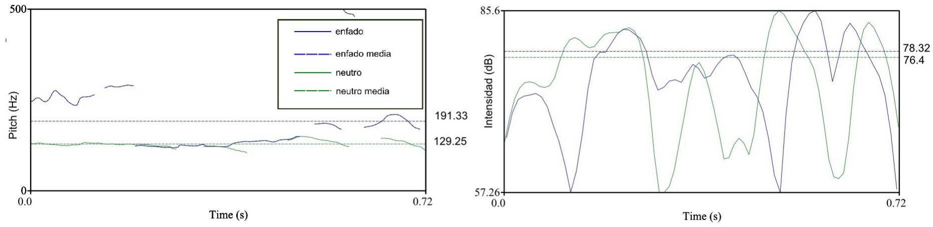


**Fig. 2.** Distribution of the pitch and the intensity in angry labeled frame '*obviamente no*' and neutral labeled frame '*decisión fue*'.

distribution is not constant, it has ups and downs. These changes are related to the interest of the speaker in emphasizing some specific words. The speed is higher ( $\sim$ 22 phon. per sec.) and the pauses number and duration is also higher.

***Anger2*** The overall tone is also angry in this call, but it is a different type of anger. The first one is more serious and harder whereas in this one boredom and tiredness are reflected.

The pitch and intensity of the segments labelled as neutral show a mixture between the two previous calls (see Fig. 3). The pitch does not show abrupt interruptions so many times, it has a more continuous behaviour. The continuous aspect of the speech is also reflected in the intensity, its value does not decrease that much at the end of the syllables, only at the end of words. It is appreciable the monotonous sound of the speech that is produced by the exceptionally flat pitch contour. The speed of the neutral segments is slightly faster than in the previous case (16-18 phon. per sec.) and the pauses matches the punctuation of the sentences.

In the case of the segments labelled as angry, both the pitch and the intensity follow the pattern of the former call. However, a more continuous way of speaking can also be noticed here (not so many interruptions, low pitch values,...). This fact makes the difference between the two angers. The speed also increases, as happened in the previous case, to 19-23 phon. per sec. The number of pauses and the duration is somewhat greater but not enough to give relevant information about the emotional state of the speaker.

## 4 Automatic Classification

Once the appropriate features are selected they can be considered as the signal representation in a classical classification problem. In this way, the emotion associated to each signal can be estimated as eq. (1) shows, being $y$ the representation of the acoustic signal uttered by the speaker and $w$ the associated emotion.

$$\hat{w} = \arg\max_{w} P(w|y) \tag{1}$$

### 4.1 SVM Classifier

The Support Vector Machines (SVM), are non parametrical binary classifiers, that were successfully employed in the emotion recognition and related fields [17]. They can be built from a set of labeled training samples and their goal is to determinate an hyperplane that would be able to divide our data representation space into two regions, each one corresponding to a specific class. Then, a new, not labeled, sample will be classified according to the region in which it is located.

When data are not linearly separable, in the feature space we are dealing with, different kernel functions can be used for the transformation or our initial data. In this work, two types of kernel function were used. 1) *Linear function:* $K_1(x, x') = \langle x, x' \rangle$. The linear dot product will keep the dimension of the input dataset, so it considers that our data are linearly separable. 2) *Radial Basis Functions (RBF):* $K_2(x, x') = exp(-\gamma|x - x'|^2)$ *where* $\gamma$ *is an adjustable coefficient.* The RBF kernel function, transforms the original input space into a higher-dimensional feature space, allowing the SVM to be non-linear and thus potentially better for separating the two classes.

### 4.2 GMM-based SVM Classifier

It seems reasonable to think that an emotional state does not correspond to an isolated and unique emotion. It will be associated, instead, to a mix of basic emotions whose influence will vary when moving from an emotional state to the other. Given this intuition we might model the emotional state of a speaker with a Gaussian Mixture Model (GMM).

A GMM is a parametric probability density function represented as a weighted sum of gaussian components. GMMs are commonly used as a parametric model of the probability distribution of continuous features in biometric systems (i.e. speaker recognition system). The model is given by the weighted sum of M gaussian density functions, each one parametrized by the mean vector $\mu_i$ and covariance matrix $\Sigma_i$ as equation (2) shows.

$$p(y|\lambda) = \sum_{i=1}^{M} \beta_i g(y|\mu_i, \Sigma_i) \tag{2}$$

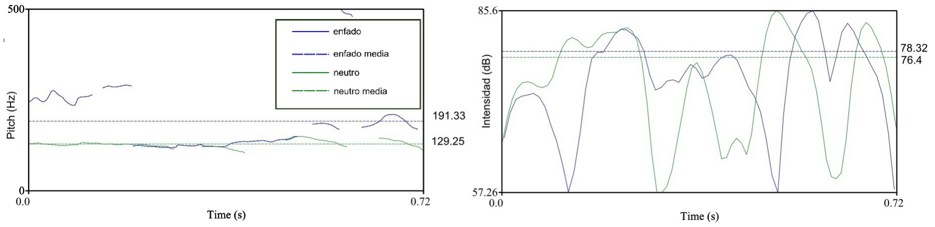


**Fig. 3.** Distribution of the pitch and the intensity in angry labeled frame '*cada dos*' and neutral labeled frame '*había una*'

where x is a D-dimensional continuous-valued data vector, $\beta_i$ for $i = 1, ..., M$ are the mixture weights and $g(y|\mu_i, \Sigma_i)$ for $i = 1, ..., M$ are the different components of the mixture. Each component is a D-variance gaussian function of the form $g(y|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp[\frac{-1}{2}(y - \mu_i)' \Sigma_i^{-1}(y - \mu_i)]$ with the mean vector $\mu_i$ and the covariance matrix $\Sigma_i$. The weights must satisfy $\sum_{i=1}^{M} \beta_i = 1$.

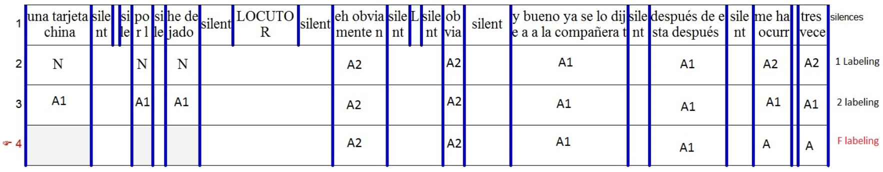
The GMM model parameters ($\lambda = \{\beta_i, \mu_i, \Sigma_i\}$ where $i = 1, ..., M$) are estimated from a training dataset using the Expectation-Maximization algorithm or Maximun A Posteriori estimation from a well trained prior model [13].

Returning to the problem of this paper, an alternative feature vector consisting of the different weights given to each component (each gaussian) might be used to represent the audio frames we are dealing with. Then, SVM classifier would divide the new feature space into the regions associated to each emotion.

### 4.3 Experimental Analysis

Before carrying out the feature extraction procedure the speech signal needed to be preprocessed. In this work, this preprocess was carried out by means of *Praat* [2] toolkit. Using it, the signal was first segmented and each segment was manually labeled according to the corresponding emotion. This procedure was accomplished by two different labelers and it was performed in two stages. The aim of the first stage was to discriminate neutral segments from angry/happy ones and in the second stage the frames labeled with angry/happy were assigned with a score representing the intensity of the emotion (score: 1-3), where 1 means less angry/happy and 3 means a maximal state of anger/happiness. Then, only the segments that were labeled in the same way by the two independent labelers were taken into account as a training set. For the preliminary experiments presented in this work only the emotion label was considered and the degree was ignored. An example of such a labeling procedure is shown in Figure 4.

In order to carry out the feature extraction procedure, smaller frames (20-40 ms), in which features remain quasi constant, were needed. Thus, each annotated segment was fragmented again and a new set of small audio frames with their corresponding labels (neutral, anger, happiness) was extracted. Then, each frame was represented by the selected features (Intensity, Pitch and Formants) that were extracted from *Praat* toolkit.



| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | una tarjeta china | sile nt | si le | po r l | si le | he de jado | silent | LOCUTO R | silent | eh obvia mente n | sile nt | L | sile nt | ob via | silent | y bueno ya se lo dij e a a la compañera t | sile nt | después de e sta después | sile nt | me ha ocurr | | tres vece | silences |
| 2 | N | | | N | | N | | | | A2 | | | | A2 | | A1 | | A1 | | A2 | | A2 | 1 Labeling |
| 3 | A1 | | | A1 | | A1 | | | | A2 | | | | A2 | | A1 | | A1 | | A1 | | A1 | 2 labeling |
| ☞ 4 | | | | | | | | | | A2 | | | | A2 | | A1 | | A1 | | A | | A | F labeling |

**Fig. 4.** Example of the labeling provided by the two independent labelers and the final labeled segments where only the coincidences are considered

| | *Happyness* | | *Anger1* | | *Anger2* | |
|---|---|---|---|---|---|---|
| | *Linear* | *RBF* | *Linear* | *RBF* | *Linear* | *RBF* |
| *I_MR* | 0.84 | 0.88 | 0.76 | 0.72 | 0.78 | 0.76 |
| *P_M* | 0.91 | 0.93 | 0.67 | 0.80 | 0.71 | 0.79 |
| *F_MSD* | 0.65 | 0.81 | 0.64 | 0.84 | 0.83 | 0.79 |

**Table 2.** Accuracy results achieved using different feature sets and different kernel functions, of a SVM classifier, for the 3 different emotion sets considered in this work.

**Classification Experiments** Once each signal frame (20-40 ms) was represented in terms of the considered feature values, a feature vector was built for each annotated signal segment made up of several frames (the number of frames in each segment depends on the length of the segment). Different feature vectors built as described below were used and compared to each other.

- *I_MR:* [Intensity Mean for all the frames in a segment, Intensity Range (difference between max/min)]
- *P_M:* [Pitch Mean for all the frames in a segment]
- *F_MSD:* [Formants Mean for all the frames in a segment, Stand. Dev.]

In this work, different series of experiments were carried out. In the first one, the feature sets described above were used to train a SVM classifier with two different kernel functions (*linear* and *RBF*). We used the implementations provided in the Scikit-learn Python package [8]. A leave-one-out cross-validation procedure was employed to evaluate the system. The obtained accuracy results for the 3 different emotions are shown in Table 2. This table shows that the achieved results are quite promising considering the early stage of the work. The best result was accomplished for *Happyness* (93%) when *P_M* was used, whereas *F_MSD* provided the highest accuracy values when considering the two kinds of *Angers*. Thus, it seems that Pitch is the best feature to represent *Happyness* but Formants are better when focusing on *Anger*. On the other hand, *RBF* kernel function provided better results in most cases.

In the second series of experiments SVM classifiers were used along with alternative feature vectors obtained from the GMM models (a total number of 15 gaussian components were selected in this work). Specifically, a GMM model was obtained for each parameter (Intensity, Pitch and Formants) and the corresponding feature vectors associated to each annotated signal segment were built as follows:

- *GMM_I:* [15 weights of the GMM model obtained for Intensity values]
- *GMM_P:* [15 weights of the GMM model obtained for Pitch values]
- *GMM_F:* [15 weights of the GMM model obtained for Formants values]

Table 3 shows the accuracy results obtained for the SVM classifier and *RBF* kernel function with the new feature sets. A leave-one-out cross-validation procedure was also employed to evaluate the system. The results show that there is no difference between the feature sets built from different information sources

(Intensity, Pitch,...) in this case. Moreover, there is no any improvement associated to this way of representing the signals when comparing it to the previous one. However, if we consider an specific method to properly choose the number of gaussian components different results might be obtained.

Finally, in a third series of experiments we wanted to test whether information coming from different features can cooperate to achieve a better system performance. Thus, different combinations of features were considered. Table 4 shows the results achieved with the new feature sets and *RBF* kernel function. These results show that the combination of the three information sources in the *S1* set improves the accuracy values achieved with the worst isolated feature set, for all the emotions, but does not reach the results obtained for the best one. That means that there is some kind of information, in each case, that instead of helping, introduces noise in the classification process. Finally, looking at the accuracy values obtained with the combination of the standard feature sets and GMM-based ones it can be concluded that they can improve the results obtained with *S1* set, for instance *S4* or *S7* results for *Happyness*, *S6* for *Anger1*. However, the reached values are indeed the same ones achieved with the best isolated feature sets (the ones presented in the first series of experiments). For example, the best result when dealing with*Happiness* is achieved with *S7* that considers the Pitch features and GMM-based features for Intensity and Formants. Although this GMM-based features do not seem to improve the results obtained with *P_M* set, it is not worse either, thus we hypothesize that with more sophisticated GMM-based feature sets these results might be improved.

## 5 Concluding Remarks and Future Work

This work was devoted to the design of a first approach for the solution of a real problem related to emotion detection in speech signals. A very challenging task that involves dialogic spontaneous speech of real users in Spanish was proposed. The main goal of the work was to adapt the employed technologies, that succeeded in similar cases, to this specific task. Several series of experiments were carried out using different feature sets and a SVM classifier. Very promising results were achieved in this preliminary stage of the work. However, a greater corpus (with much more dialogues) should be built in future work in order to be able to extract reliable conclusions. Additionally, other kind of features related to changes (derivatives) in parameters should be considered to get robust systems that could work with unknown speakers.

| | *Happyness* | *Anger1* | *Anger2* |
|---|---|---|---|
| *GMM_I* | 0.83 | 0.75 | 0.78 |
| *GMM_P* | 0.83 | 0.75 | 0.78 |
| *GMM_F* | 0.83 | 0.75 | 0.78 |

**Table 3.** Accuracy results achieved using different feature sets when SVM classifier was used along with GMM models based features sets, for the 3 different emotion sets.

| | *Happyness* | *Anger1* | *Anger2* |
|---|---|---|---|
| *S1 (I_MR + P_M + F_MSD)* | 0.83 | 0.81 | 0.78 |
| *S2 (GMM_I + P_M + F_MSD)* | 0.83 | 0.82 | 0.78 |
| *S3 (I_MR + GMM_P + F_MSD)* | 0.83 | 0.81 | 0.76 |
| *S4 (I_MR + P_M + GMM_F)* | 0.88 | 0.76 | 0.78 |
| *S5 (I_MR + GMM_P + GMM_F)* | 0.83 | 0.79 | 0.78 |
| *S6 (GMM_I + P_M + GMM_F)* | 0.81 | **0.83** | 0.78 |
| *S7 (GMM_I + GMM_P + F_MSD)* | **0.93** | 0.80 | 0.78 |
| *S8 (GMM_I + GMM_P + GMM_F)* | 0.83 | 0.75 | 0.78 |

**Table 4.** Accuracy results achieved using different features built as combinations of previous features sets and a SVM classifier, for the 3 different emotions.

## References

1. Ayadi, M.E., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44, 572–587 (2011)
2. Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)
3. Devillers, L., Vidrascu, L., Lamel, L.: Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422 (2005), 2005
4. Henríquez, P., Alonso, J.B., Ferrer, M.A., Travieso, C.M., Orozco-Arroyave, J.R.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Neurocomuting 132, 126–135 (2014)
5. Hewlett, S.: Emotion detection from speech. Tech. rep., Univ. of Stanford (2007)
6. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech: a review. International Journal of Speech Technology 15, 99–117 (2012)
7. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13(2), 293–303 (March 2005)
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
9. Pfister, T.: Emotion Detection from Speech. University of Cambridge Part II dissertation (2010)
10. Polzin, T., Waibel, A.: Detecting emotions in speech. In: Second International Conference on Cooperative Multimodal Communication (January 1998)
11. Ramos, D., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., Lucena-Molina, J.J.: Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-casework database in spanish. In: Proceedings of the INTERSPEECH (September 2008), ahumadaIII
12. Rao, K.S., Koolagudi, S.G.: Robust Emotion Recognition Using Spectral and Prosodic Features. Springer Publishing Company, Incorporated (2013)
13. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. Speech and Audio Processing, IEEE Transactions on 3(1), 72–83 (Jan 1995)
14. Schuller, B., nton Batliner, Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53, 1062–1087 (2011), 2005 Special Issue

15. Shafran, I., Mhori, M.: A comparison of classifiers for detecting emotion from speech. In: Proceedings of the Interational Conference on Acoustic, Speech and Signal Processing. Philadelphia, Pennsylvania, USA (March 2005)
16. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods. Speech Communication 48(9), 1162–1181 (September 2006)
17. Vidrascu, L., Devillers, L.: Detection of real-life emotions in call centers. In: Proceedings of the INTERSPEECH. Lisboa, Portugal (September 2005)

# Bilingual sentence selection strategies: comparative and combination in statistical machine translation systems

Mara Chinea-Rios[1], Germán Sanchis-Trilles[1] and Francisco Casacuberta[1]

Pattern Recognition and Human Language Technologies,
Universitat Politècnica de València, Valencia, Spain
{machirio,gersantr,fcn}@prhlt.upv.es

**Abstract.** Bilingual corpora constitute an indispensable resource for translation model training in statistical machine translation. However, it is not really clear if including all the training data available actually helps to improve translation quality. Bilingual sentence selection aims to select the best subset of the bilingual sentences from an available pool of sentences, with which to train a SMT system. This article studies, compares, and combines two kinds of data selection methods: the first method is based on cross-entropy difference, and the second method is based on infrequent n-gram occurrence. Experimental results report improvements compared with a system trained only with in-domain data. In addition, the results obtained with the system trained with the data selected are comparable to those obtained with a system trained with all the available data.

**Keywords:** statistical machine translation; bilingual sentence selection; infrequent n-gram; cross-entropy

## 1 Introduction

Entities such as the European Union, the United Nations, and other international organizations need to translate all documentation they generate into different languages. Those translations provide very large multilingual corpora. Statistical machine translation (SMT) [1] relies heavily on the availability of such bilingual corpora. Usually, bilingual corpora are used to estimate the parameters of the translation model. However, large corpora are difficult to process, increasing the computational requirements needed to train statistical machine translation systems robustly. For example, the corpora made available for recent machine translation evaluations are in the order of 1 billion running words [2].

Different problems that arise when using this huge pool of sentences are:

- The use of all corpora for training increases the computational requirements.
- These corpora are obtained from different domains. The domain of the corpus used to train the SMT system might be different from the domain of the text to translate, and this typically entails a significant loss in translation quality.

Despite the previous problems, the standard consists in training SMT systems with all the available data. It is assumed that the more data used to train the system, the better. This assumption is correct if all the data belongs to the same domain. However, this is not the case in the problems tackled by most of the SMT systems. In fact, most SMT systems are designed to translate specific text, such as user manuals or medical prospects. In this scenario, the question posed is whether training a system with all the data available yields better performance than training it with a wisely selected subset of bilingual sentences.

*Bilingual sentence selection* (BSS) aims to select the best subset of bilingual sentences from an available pool of sentences. This paper is focused in studying and combining two different strategies of BSS. By doing so, we pretend to improve the state of the art in terms of translation quality obtained and computational requirements, without using the complete pool sentences. We will refer to the pool of sentences available as *out-of-domain* corpus because we assume that it belongs to a different domain than the one to be translated. Similarly, we refer to the corpus of the domain of the text to translate as *in-domain* corpus.

The most recent literature [3, 4] defines the SMT problem as follows: given an input sentence $\mathbf{x}$ from a certain source language, the purpose is to find an output sentence $\mathbf{y}$ in a certain target language such that:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y}) \tag{1}$$

where $\lambda_m$ is the weight assigned to $h_m(\mathbf{x}, \mathbf{y})$ and $h_m(\mathbf{x}, \mathbf{y})$ is a score function representing an important feature for the translation of $\mathbf{x}$ into $\mathbf{y}$, as for example the language model of the target language, a reordering model, or several translation models. The weights $\lambda_m$ are normally optimised with the use of a development set. The most popular approach for adjusting $\lambda_m$ is the one proposed in [5], commonly referred to as *minimum error rate training* (MERT). This algorithm implements a coordinate-wise global optimisation.

The main contributions of this paper are:

- We compare two different BSS strategies. To the best of our knowledge, such study does not exist in the literature.
- We present a combination of both strategies and report improvements over each strategy independently.

This paper is structured as follows. Section 2 summarises the related work. Section 3 presents the two BSS techniques selected, namely, recovery of infrequent n-grams and cross entropy selection. In Section 4, experimental results are reported. Finally, conclusions and future work are presented in Section 5.

## 2 Related work

Bilingual sentence selection has been receiving an increasing amount of attention within the SMT research community. Most of the methods developed [6–8] define a ranking of the sentences in an out-of-domain corpus according to their importance when training an SMT system for an in-domain corpus.

For example, in [9, 10] several BSS techniques have been applied for training MT systems when large training corpora are available. A similar approach that gives weights to different sub-corpora was proposed in [11].

The simplest instance of this problem can be found in language modelling, where perplexity-based selection methods have been used [12]. Here, out-of-domain sentences are ranked by their perplexity score. Another perplexity-based approach is presented in [7], where cross-entropy difference is used as a ranking function rather than just perplexity, in order to account for normalization. Here, we apply this criterion for the task of selecting training data for SMT systems.

Different works use perplexity-related BSS strategies [8, 13]. In [8], the authors used three methods based in cross-entropy for extracting a pseudo in-domain corpus. This pseudo in-domain corpus is used to train small domain-adapted SMT systems. In [13] the authors describe the *XenC* open source toolkit for data selection. *XenC* uses the two strategies described in [12] and [7] for data selection. The best results were obtained using difference in cross-entropies. In our experiments with cross-entropy, we will be using *XenC*.

Two different approaches are presented in [6]: one based on approximating the probability of an in-domain corpus and another one based on infrequent n-gram occurrence. The technique approximating the probability relies on conserving the probability distribution of the task domain by wisely selecting the bilingual pairs to be used. Hence, it is mandatory to exclude sentences from the pool that distort the actual probability. The technique based in infrequent n-gram occurrence will be explained in detail in the next section.

Other works have applied information retrieval methods for BSS [14], where BSS is used in order to produce different sub-models which are then weighted. In that work, authors define the baseline as the result obtained by training only with the corpus that shares the same domain with the test. Afterwards, they claim that they are able to improve the baseline translation quality by adding new sentences retrieved with their method. However, they do not compare their technique with a model trained with all the corpora available.

## 3 Data selection methods

In this section we explain the two BSS techniques that we have selected for our work. The first strategy, proposed in [7], is based in cross-entropy. This strategy is used in many different works [8, 13, 15–17]. In these papers, the authors report good results when using the strategy presented in [7], and such strategy has become a de-facto standard in the SMT research community, and this is the reason why we selected this strategy for our work.

The second strategy we used in this work is infrequent n-grams recovery. This strategy was presented in [6], alongside with another BSS strategy. The best results were obtained with infrequent n-grams recovery, achieving an improvement of 1 BLEU point. Since this strategy has not been compared as of yet with the perplexity-based strategies, we also included it in the present paper.

Both strategies above depend on the n-grams that compose the corpus considered, be it for building a language model (cross-entropy selection) or for determining which n-grams are infrequent (infrequent n-gram recovery). In this work we will also analyse the effect of varying the order of the n-grams considered, since this will also imply that the final sentence selection will be different. Specifically, we will consider 2-grams and 5-grams.

### 3.1 Infrequent n-grams recovery

The main idea underlying the infrequent n-grams recovery strategy consists in increasing the information of the in-domain corpus by adding evidence for those n-grams that have been seldom observed in the in-domain corpus. This evidence is obtained by selecting sentences from the out-of-domain corpus. The n-grams that have never been seen or have been seen just a few times are called *infrequent n-grams*. An n-gram is considered infrequent when it appears less times than a given infrequency threshold $t$. Therefore, the strategy consists on selecting from the out-of-domain corpus the sentences which contain the most infrequent n-grams in the source sentences to be translated.

Let $X$ be the set of n-grams that appears in the sentences to be translated and $\mathbf{w}$ one of them; let be $N(\mathbf{w})$ the counts of $\mathbf{w}$ in a given source sentence $\mathbf{x}$ of the out-of-domain corpus, and $C(\mathbf{w})$ the counts of $\mathbf{w}$ in the source language in-domain corpus. Then, the infrequency score $i(\mathbf{x})$ is defined as:

$$i(\mathbf{x}) = \sum_{\mathbf{w} \in X} \min(1, N(\mathbf{w})) \max(0, t - C(\mathbf{w})) \qquad (2)$$

Then, the sentences in the out-of-domain corpus are scored using Equation 2. This being done, the sentence $\mathbf{x}^*$ with the highest score $i(\mathbf{x}^*)$ is selected in each iteration. $\mathbf{x}^*$ is added to the in-domain corpus and is removed from the out-of-domain sentences. The counts of the n-grams $C(\mathbf{w})$ are updated with the counts $N(\mathbf{w})$ within $\mathbf{x}^*$ and therefore the scores of the out-of-domain corpus are updated. Note that $t$ will determine the maximum amount of sentences that can be selected, since when all the n-grams within $X$ reach the $t$ frequency no more sentences will be extracted from the out-of-domain corpus.

### 3.2 Cross-entropy selection

As mentioned in Section 2, one established method consists in scoring the sentences in the out-of-domain corpus by their perplexity score. We follow the procedure described in [7], which uses the cross-entropy rather than perplexity. Perplexity and cross-entropy are monotonically related. The perplexity of a given sentence $\mathbf{x}$ with empirical n-gram distribution $p$ given a language model $q$ is:

$$2^{-\sum_x p(x) \log q(x)} = 2^{H(p,q)} \qquad (3)$$

where $H(p, q)$ is the cross-entropy between $p$ and $q$. The formulation proposed by [7] is: Let I be an in-domain corpus and G be an out-of-domain corpus. Let $H_I(\mathbf{x})$ be the cross-entropy, according to a language model trained on $I$, of a sentence $\mathbf{x}$ drawn from $G$. Let $H_G(\mathbf{x})$ be the cross-entropy of $\mathbf{x}$ according to a language model trained on $G$. The cross-entropy score of $\mathbf{x}$ is then defined as

$$c(\mathbf{x}) = H_I(\mathbf{x}) - H_G(\mathbf{x}) \tag{4}$$

Then, all those sentences whose cross-entropy score is less than a certain threshold $R$ are selected to be included into the in-domain corpus.

# 4 Experiments

In this section, we describe the experimental framework employed to assess the performance of the data selection techniques described above. Then, we show the results for the cross-entropy selection, followed by the results obtained with the infrequent n-grams strategy. Finally, we present results obtained by combining both techniques.

## 4.1 Experimental set-up

We evaluated empirically the BSS methods described in the previous section. For the out-of-domain corpus, we used the English-Spanish parallel text from release v7 of the Europarl corpus [18]. The Europarl corpus is composed of translations of the proceedings of the European parliament. As in-domain data, we used the News Commentary (NC) corpus[1]. The NC corpus is composed of translations of news articles. We evaluated our work on the News Commentary 2013 test set. The main figures of the corpora used are shown in Tables 1 and 2.

| | | SP | EN |
|---|---|---|---|
| Europarl | $\|S\|$ | 1.9M | |
| | $\|W\|$ | 51.5M | 49.1M |
| | $\|V\|$ | 422k | 308k |

Table 1: Europarl corpus main figures. k denotes thousands of elements, M denotes million of elements, $|S|$ stands for number of sentences, $|W|$ stands for number of words (tokens) and $|V|$ for vocabulary size (types).

All experiments were carried out using the open-source SMT toolkit Moses [19] in its standard non-monotonic configuration. The language model used was a 5-gram standard in SMT system, with modified Kneser-Ney smoothing [20], built with the SRILM toolkit [21]. The phrase table was generated by means

[1] available at `http://www.statmt.org/wmt13`

| | NC-Domain | | NC-Test | | NC-Mert | |
|---|---|---|---|---|---|---|
| | SP | EN | SP | EN | SP | EN |
| $\|S\|$ | 149k | | 3000 | | 2051 | |
| $\|P\|$ | 4.5M | 3.9M | 62.6k | 56.9k | 47.1k | 43.4k |
| $\|V\|$ | 178k | 143k | 10.5k | 8.9k | 8.7k | 7.7k |

Table 2: News Commentary main figures. NC-Domain is the in-domain corpus, NC-Test is the evaluation data and NC-Mert is development set. M denotes millions of elements and k thousands of elements, $|S|$ stands for number of sentences, $|W|$ for number of words (tokens) and $|V|$ for vocabulary size (types).

of symmetrised word alignments obtained with GIZA++ [22]. The log-lineal combination weights in Equation 1 were optimized using MERT (minimum error rate training) [5] on the NC-Mert data, which was the test set used in the 2008 WMT evaluation[2].

We compared the selection methods with two baseline systems. The first one was obtained by training the SMT system with NC-Domain data. We will refer to this setup with the name of `baseline-nc`. A second baseline experiment has been carried out with the concatenation of the Europarl corpus and News Commentary training data. We will refer to this setup with the name of `baseline-all`.

Results are shown in terms of BLEU [23], which is an accuracy metric that measures n-gram precision, with a penalty for sentences that are too short.

### 4.2 Results for infrequent n-grams recovery

Figure 1 shows the effect of adding sentences using infrequent n-grams selection, up to the point where the specific value of $t$ does not allow to select further sentences, and the two baseline systems evaluated on the News Commentary 2013 test set. We only show results for threshold values $t = \{10, 25\}$, and using either 2-grams or 5-grams.

Several conclusions can be drawn:

- The translation quality provided by the infrequent n-grams technique is significantly better than the results achieved with the system `baseline-nc`. Specifically, the improvements obtained are in the range of 2.5 BLEU points.
- Selecting sentences with the infrequent n-grams technique does not provide better results than including all the available data (`baseline-all`) but the results are very similar using less than 7% of the out-of-domain corpus.
- A slight gain in terms of translation quality is achieved when using 5-grams instead of 2-grams, although the differences are not statistically significant. However, using 5-grams allows to select more sentences.

[2] `http://www.statmt.org/wmt08`

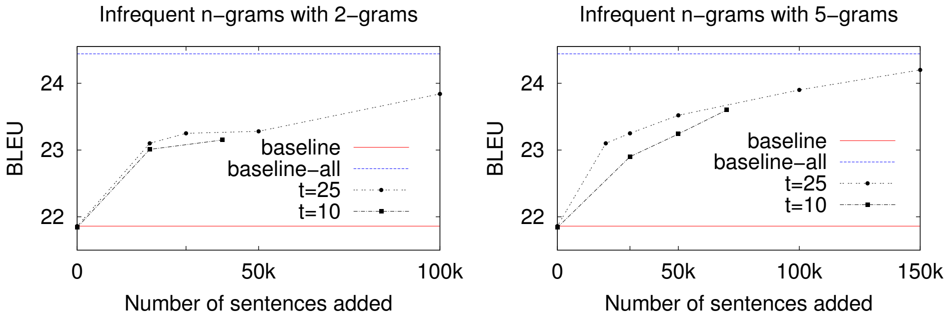


Fig. 1: Effect to adding sentences over the BLEU score using infrequent n-grams recovery (with different n-ngram value). Horizontal lines represent the score the `baseline-nc` system and `baseline-all` system

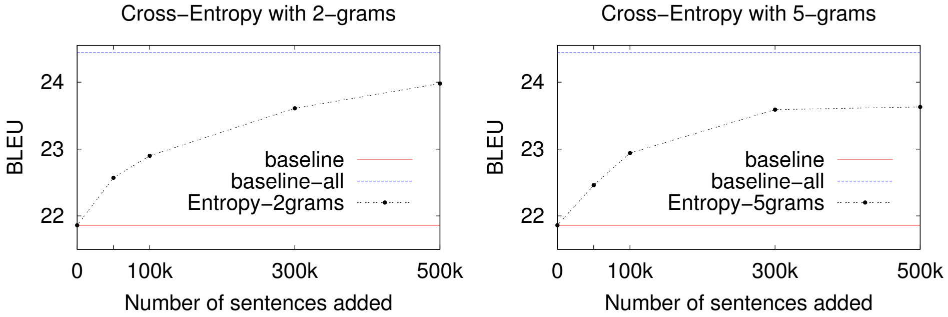


Fig. 2: Effect to adding sentences over the BLEU score using cross-entropy strategy (with different n-gram value). Horizontal lines represent the score the baseline-nc system and baseline-all system

- As expected, $t = 25$ allows to select more sentences than $t = 10$, which also leads to higher BLEU scores. In addition, the results with $t = 10$ are slightly worse than with $t = 25$, for the same amount of sentences. We understand that this is because $t = 25$ entails a better estimation of the n-grams considered infrequent.
- The best result is achieved with $t = 25$, 5-grams and adding 150k sentences.

### 4.3 Results for cross-entropy strategy

Figure 2 shows the effect of adding sentences by means of the cross-entropy strategy, alongside with the two baseline systems, on the News Commentary 2013 test set. We obtained results using both 2-grams and 5-grams.

Several conclusions can be drawn:

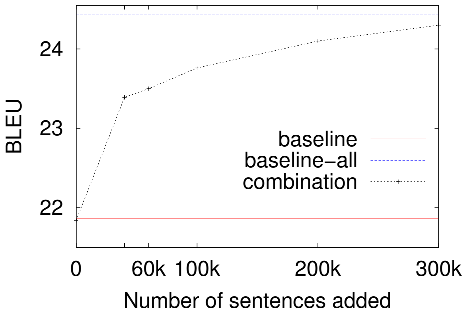


Fig. 3: Effect to adding sentences over the BLEU score using strategies combination (cross-entropy and infrequent n-grams). Horizontal lines represent the score the baseline-nc system and baseline-all system

- Adding sentences selected by means of cross-entropy improves over `baseline-nc` from the very beginning, and seems to stale at around 500k with an approximate improvement of 2 BLEU points.
- Performing sentence selection with cross-entropy does not achieve better performance than `baseline-all`.
- The order of the n-grams used does not seem to affect significantly, except for the case of 500k sentences, where using 2-grams behaves slightly better.
- The best result obtained is using 2-grams and adding 500k sentences.
- Lastly, it is also worth noting that the results obtained with the cross-entropy strategy are slightly worse than the ones obtained with infrequent n-gram recovery, even though more sentences are selected.

### 4.4 Combining sentences from infrequent n-grams and cross-entropy

In this section, we present the experimental results obtained by concatenating the best sentences selected by means of both strategies described above, namely infrequent n-grams recovery and cross-entropy selection. The best sentences are obtained from a sorted list that is generated by each strategy according to Equations 2 and 4. The best sentences for each strategy were obtained with the best parameters seen in the previous sections, i.e. $t = 25$ and 5-grams in the case of infrequent n-gram recovery, and 2-grams in the case of cross-entropy. We selected the same amount of sentences for each strategy and such sentences were added to the in-domain corpus. The objective is to add the best sentences of each strategy to the in-domain corpus.

Figure 3 shows the effects of adding the sentences obtained as described above to the in-domain corpus, alongside with the two baseline systems on the translation quality obtained for the News Commentary 2013 test set.

Several conclusions can be drawn:

– The results are very similar with respect to each selection strategy individually, although the best result achieved does improve slightly over the best result achieved by each strategy individually.
– The combination provides similar results than including all the available data `baseline-all` using less than 14% of the out-of-domain.

Table 3 shows the best results obtained with both strategies individually and with the combination and the baseline systems.

| Strategy | BLEU | Number of sentences |
|---|---|---|
| Baseline-nc | 21.8 | 138k |
| Baseline-all | 24.4 | 2.1M |
| Cross entropy | 24 | 500k |
| Infreq $t = 25$ | 24.1 | 150k |
| Infreq+Cross entropy | 24.3 | 300k |

Table 3: Summary of the best results obtained with each setup.

## 5 Conclusion and future work

Bilingual sentence selection has been receiving an increasing amount of attention within the SMT research community. There are a lot of data selection methods. In this work, we study two different BSS strategies. These strategies have been used frequently for selecting the sentences in the literature. In this work, we perform a comparison of these two techniques. The results obtained are very similar, although the best results were obtained by the infrequent n-grams. With the combination proposed we obtained better results than with each strategies independently. Even though the improvements obtained are not statistically significant, we understand that they are encouraging and indicate that more sophisticated combination strategies, such as score interpolation, could yield more significant improvements. Lastly, we achieved similar results than when using all the data available, but with only 15% of the available data.

In future work, we will carry out new experiments with bigger and more diverse data sets, as was done in [6], and with higher $t$ values. In addition, we also intend to combine the two strategies proposed in more sophisticated ways.

## References

1. W. Weaver and L. W, "Machine translation of languages.," in *MIT Press*, pp. 15–23, (1945).
2. C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. F. Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," in *Proc. of ACL*, pp. 17–53, (2010).

3. K. A. Papineni, S. Roukos, and R. T. Ward, “Maximum likelihood and discriminative training of direct translation models,” in *Proc. of ICASSP*, pp. 189–192, (1998).
4. F. J. Och and H. Ney, “Discriminative training and maximum entropy models for statistical machine translation,” in *Proc. of ACL*, pp. 295–302, (2002).
5. F. J. Och, “Minimum error rate training in statistical machine translation,” in *Proc. of ACL*, pp. 160–167, (2003).
6. G. Gascó, M.-A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta, “Does more data always yield better translations?,” in *Proc. of EACL*, pp. 152–161, (2012).
7. R. C. Moore and W. Lewis, “Intelligent selection of language model training data,” in *Proc. of ACL*, pp. 220–224, (2010).
8. A. Axelrod, X. He, and J. Gao, “Domain adaptation via pseudo in-domain data selection,” in *Proc. of EMNLP*, pp. 355–362, (2011).
9. Z. Li, C. Callison-Burch, C. Dyer, J. Ganitkevitch, A. Irvine, *et al.*, “Joshua 2.0: a toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies,” in *Proc. of ACL*, pp. 133–137, (2010).
10. G. Gascó, V. Alabau, J. Andrés-Ferrer, J. González-Rubio, M.-A. Rocha, G. Sanchis-Trilles, F. Casacuberta, J. González, and J.-A. Sánchez, “Iti-upv system description for iwslt 2010.,” in *Proc. of IWSLT*, pp. 85–92, (2010).
11. S. Matsoukas, A.-V. I. Rosti, and B. Zhang, “Discriminative corpus weight estimation for machine translation,” in *Proc. of EMNLP*, pp. 708–717, (2009).
12. J. Gao, J. Goodman, M. Li, and K.-F. Lee, “Toward a unified approach to statistical language modeling for chinese,” *ACM TALIP*, pp. 3–33, (2002).
13. A. Rousseau, “Xenc: An open-source tool for data selection in natural language processing,” *The Prague Bulletin of Mathematical Linguistics*, vol. 100, pp. 73–82, (2013).
14. Y. Lü, J. Huang, and Q. Liu, “Improving statistical machine translation performance by training data selection and optimization.,” in *Proc. of EMNLP-CoNLL*, pp. 343–350, (2007).
15. H. Schwenk, A. Rousseau, and M. Attik, “Large, pruned or continuous space language models on a gpu for statistical machine translation,” in *Proc. of the NAACL-HLT*, pp. 11–19, (2012).
16. R. Sennrich, “Perplexity minimization for translation model domain adaptation in statistical machine translation,” in *Proc. of EACL*, pp. 539–549, (2012).
17. S. Mansour, J. Wuebker, and H. Ney, “Combining translation and language model scoring for domain-specific data filtering.,” in *Proc. of IWSLT*, pp. 222–229, (2011).
18. P. Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *MT summit*, pp. 79–86, (2005).
19. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: open source toolkit for statistical machine translation,” in *Proc. of ACL*, pp. 177–180, (2007).
20. R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in *Proc. of ICASSP*, vol. 1, pp. 181–184, (1995).
21. A. Stolcke, “Srilm-an extensible language modeling toolkit,” in *Seventh International Conference on Spoken Language Processing*, (2002).
22. F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational linguistics*, vol. 29, pp. 19–51, (2003).
23. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proc. of ACL*, pp. 311–318, (2002).

# Holistic analysis based on the index of proportionality for offline handwritten biometric identification

[1]Carlos M. Travieso, [1]David Hernández-Abreu, [2]José Luis Vásquez, [1]Jesús B. Alonso

[1]Institute for Technological Development and Innovation in Communications, Signals and Communications Department, IDETIC-ULPGC, University of Las Palmas de Gran Canaria.
Las Palmas de Gran Canaria, SPAIN.
[2]Sede de Atlántico, Universidad de Costa Rica.
{carlos.travieso, jesus.alonso}@ulpgc.es,
david.hernandez121@alu.ulpgc.es, jose.vasquez@ucr.ac.cr

**Abstract.** A biometric system based on offline handwriting has been implemented by on the parameter "index of proportionality", which is manually using on graphometric forensic techniques. The study of this parameter has been developed by merging up to 15 words, for 100 users. Given the handwritten nature of the information and its great variability, a feature selection based on principal component analysis and neural network classifier was used, reaching an accuracy rate of 84%.

**Keywords:** Handwriting recognition, offline system, biometric identification.

## 1 Introduction

All studies agree that writing is primarily the execution of a gesture, and any gesture is a result of the activation of certain brain regions in which they are designed and controlled movements. Any gesture has to be prepared, programmed, produced and managed with varying degrees of consciousness, and therefore the act of writing is an individual execution on a particular form for each individual [1]. Moreover, as it is a gesture, unlike the vast majority is captured in a plane and in turn depends like any other personality, and it has great value of identification.

The identification of people based on scanned manuscripts is a technique applicable in the forensic field, and too in the analysis of historical texts. Nowadays, the technological advances and the proliferation of heavy computers in modern society are indisputable facts, and its use to analyze handwritten documents and personal style are a great goal, when writing remain of great importance. Due to the wide use of such documents, many manuscripts are subject to falsification, distortion, or copies, usually for illicit use. Therefore, to certify and judge the authenticity or falsity of handwritten documents cover the most time of experts, working in this field [2]. The applied technology can be a great opportunity to do easy this hard and strong manual tasks.

adfa, p. 1, 2011.

As stated above, the writing is an individual mark. A person can hardly supplant writing another individual; perhaps yes from the fast view, but the features, which characterize a writer, will be unintentionally incorporated in the writing process. A correct and complete analysis allow to extract a number of features that differences writing between persons, or a person vs. herself/himself at different stages of his life. The interest of the person identification through his writing has given rise to many disciplines. The first name given to the study of writing was Graphology, and had the task of making a psychic analysis of the author, but now has become a branch within the general study of writing, since the parameters are not only extracted psychic. This has led to the general concept of Graphometric as a general discipline studying writing.

In recent years, it has been done lots of work in the field of biometric identification based on a variety of handwriting parameters. In the next paragraphs, some of those work will be cited:

- In [3], a writer identification using directional ink-trace width measurements performed in 2010, it is carried out to identify the individual from stroke width and address. It is shown that the combination of both parameters is a powerful tool in the identification, as they are interrelated. It was achieved an efficiency between 63-95%.
- In the study conducted in 2013 called Texture-based descriptors for writer identification and verification [4], the issue of identification covers from descriptors textures, using a classification scheme based on the representation of dissimilarities for an verification approach. Two databases from different sources were analyzed, which a 96.7% and 99.2% respectively were obtained.
- In [5], a writer identification using an alphabet of gradient contour descriptors was published in 2013. This method for the writer identification emulated approaches taken by forensic document examiners. It combined a novel feature, which uses a gradient contour to capture the shape and curvature, with character segmentation to create a pseudo-handwritten script for a given sample. A distance classification system is defined between the elements of the alphabet created and included the similarity between two samples. This approach achieved up to 96.5% accuracy at best case.
- A novel sparse model based forensic writer identification is presented in [6] and published in 2014. This work is based on the sparse representation of structural primitives, called graphemes. The main novelty of this system compared to others using grapheme is the use of sparse coding instead of vector quantization, and then, the graphemes are represented in terms of Fourier. Depending on the number of writers, the efficiency of this system varies between 98% for 5 users, and 88% for 650 writers.
- A calculation of parameters based on forensic science for writers identification [7] was done in 2012. This study showed the effectiveness of different parameters, when are extracted from the scanned samples that make up a database, and later, showed its combination. The accuracy reached up to 97.36%.

- Another interesting work is in [8], where a large variety of classifiers were used in these and other works for handwriting identification. This work included a study based on the use of ROC curves ((Receiver Operating Characteristic), for neural networks, Support Vector Machines (SVM) and models using Generalized Gaussian Distribution (GGD) among some other [8].

The parameters used in forensic identification for handwritings are a lot and exploit different forensic features of the writing. The study of these parameters was typically done manually, without any automation, and with large and cost study from a time point of view. Finally, different experts add a certain aspect of subjectivity analysis on the development of this task depending on the behavior and this charge of work [9].

Therefore, the main motivation and innovation of this work are to develop an automated and reliable system in which it is possible to identify a individual using a single parameter, or at least, to get a higher rate of success.

In addition, the field of biometric identification based on forensic techniques has not been automated, but using manual systems. Therefore, the goal of this work is to design and implement a system of automatic recognition of people from scanned samples of handwritten words. For this, a specific classification approach will be developed for a parameter called "index of proportionality", by which characteristic or key points of some of the words, available in the database; and the calculation of Euclidean distances between the key points may define the index of proportionality. This parameter is very used in forensics, but it has not been automated, yet.

The process is carried out in different words, in particular, from 1 to 15 different words. After to get the parameter, some different modules will be included, due to great variability of the handwritten information. The first one is the feature selection, later, a data fusion blocks for adding information from some works, and finally, a score fusion block, in order to give more robustness to the classification stage. The idea is to achieve a success rate as high as possible; obtaining a reliable system for person identification, using a single parameter.

## 2 Preprocessing and feature extraction

From the data as image, the most fast and efficient way of working is with the binarized image, so that the preprocessing stage consists of a binarization block, applied based Otsu method [10]. Subsequently, parameters are calculated to extract the information of the image.

The index of proportionality, which refers to writing, is the relationship between the various elements of graphics, for example, the heights of the upper and outstanding letters with respect to the average height of letters, etc. There is a normal proportionality between all elements of the writings that comply with calligraphic precepts. This proportionality is perhaps one of the most consistent characters in spontaneous spelling, because although the size and extent of writing may vary as a result of the person and her/his activity, from the built dataset and its particularities, the proportionality stays constant for the intra-classes and varies for the inter-classes.. Moreover, it is one of the elements that easily escape for impostors, which are more attentive to altering the shape of the letters and some other outstanding element, such as size, direction, inclination, etc. [11].

Once these key or control points have been extracted, the Euclidean distance is calculated existing between each one of them, according to equation 1. Later, the distances are stored on an own matrix of each of the analyzed words, repetitions and users. This matrix will be the input for our classification approach. These extracted distances are assumed proportional between the way of writing for each user. Five points are extracted for each word; thereby obtaining ten distances, the combination of distances between 5 five key points without repetition. They will be those that characterize a given word.

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

For locating points, it has been developed the following algorithm. The initial t of the binary image, looks for the object pixels, and with that information, the following parameters are calculated according to the word. In particular, in the case of "perfect", it is been extracted as the following parameters:

- Initial point of the word.
- Last point of the word.
- Over-word strokes of “f”
- Under-word stroke of “f”
- Over-word strokes of “t”

The result of the execution of this process described is as follows (see Fig. 1):

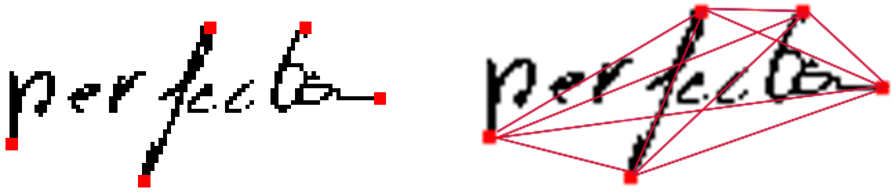


**Fig. 1.** Election of key points and its Euclidean distance between them.

## 3 Feature selection and classification system

On handwritten applications, the variability of the data is very large, so it has decided to use a feature selection based on principal components analysis (PCA) [12] and subsequently, to apply an artificial neural network [13].

PCA is a technique used to reduce the dimensions of a data set. Intuitively, the technique is useful to find the causes of the variability of a data set and sort them by importance. This analysis looks for the projection of the better data representation in terms of least squares [14].

This analysis builds a linear transformation that chooses a new coordinate system for the original data set in which the eigenvector of the biggest data variance is shown on the first axis (called First Principal Component), the second largest eigenvector is the second axis, and so on. The fundamental advantage of the analysis is the sort of the different components, since in that sort is included the information of each component [12].

A neural network with supervised learning is used for the classification system, and therefore, a part of the samples will be used for training and the rest of test data. The classifier is a *Feedforward* neural network, consisting of three layers, an input, hidden and output layer. The chosen model is the Multilayer Perceptron (MLP) with *Back-Propagation* learning algorithm [13].

In order to improve the efficiency of our classification system, a Fusion Score module is applied. The goal of this module is the random initial of the weights used in the neural networks. Using N neural networks, it is obtained N different solutions for the same input data. The idea is to fuse the N outputs in order to generalize the output to the results of neural networks that perform better.

So, this fusion block is take the random initialization of weights and different convergences reached by the classifier, to establish a generalization, and always applying the same input data [7]. The scheme of operation of this type of fusion on the neural network system is shown in Fig. 2.

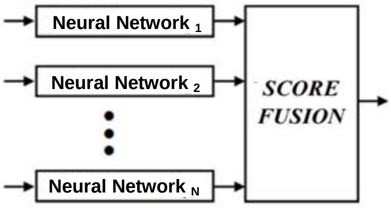


**Fig. 2.** Block diagram of the *Score Fusion*

# 4 Experiments and Results

## 4.1 Database

The databases of handwriting data have different characteristics depending on the training process and usability, through which they were created. They may be dependent or independent of pen, of limited vocabulary and open vocabulary, of paper, of on-line or off-line system. For this work, an offline dataset has been used, the same dataset used in [7] and [15].

To formalize the database, a paragraph of 15 lines in the official language Spanish was used. This paragraph has been repeated 10 times on separate A4 sheets (297 mm. × 210 mm.) per each user, composing the body of writing (offline dataset). The paragraph has been chosen because this size is considered to be sufficient for writers to show their personal characteristics and reflect his/her habit of writing (biometric).

The database consists of 100 writers and 10 repetitions per each one during a week, getting the effect of temporal invariance. The statistical data of user is very heterogeneous, between 14 and 65 year old for both genders.

The database has been digitized by a basic scanner at a resolution of 200 dpi for greyscale (8 bits) and BMP format. Then, it has been extracted 14 words of each body of writing. The choice of words is performed based on two premises: lengths of the word and words have some particular features or/and strokes in order to show the characteristics of each user. The words selected were: "*bastante*", "*Cervantes*", "*edición*", "*española*", "*Hidalgo*", "*idealista*", "*ingenioso*", "*Mancha*", "*necesidades*", "*perfecta*", "*positivista*", "*publicó*", "*Quijote*", "*tipos*" y "*universal*".

### 4.2 Experimental settings

The idea is to determine the optimal combination of the number of principal components to be used and the number of neurons in the hidden layer. For this, various tests were carried out where those previous parameters were varied. To find the optimum parameters some fast simulations were done, fixing some parameters, in particular, 200 iterations as stopping criterion and 5 neural networks for the score fusion.

The test process is performed on the data from a single word and the most effective combination of principal components and number of neurons in the hidden layer is applied to the rest of the analyzed words. This is done to simplify the complex system of simulations and combinations, and to get approximate results.

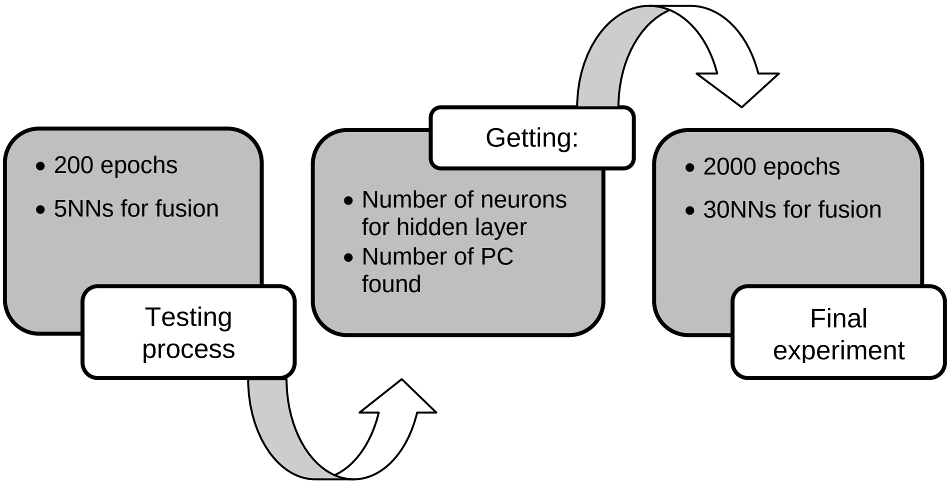


**Fig. 3.** Experimental settings

Having analyzed the results of the previous process, it is observed the combination of best accuracy, and therefore it will be the final combination: the number of iterations for this final experiment is 2000 (10 times higher than that used in the previous testing process), and 30 neural networks for the fusion process. Besides, the data fusion from 1 to 15 words will be added, too. Fig. 3 illustrates the experimental settings to get the final results.

The previous conditions are dependent of the number of words, because the number of inputs changes when is increased the data fusion (more words). And therefore, the process has to be repeated for the different combinations of word.

### 4.3 Results and Discussions

As already mentioned, the first methodological conditions were to fix the stopping criterion for the neural network training, which is 200 iterations, and straightaway, to use 5 neural networks, working in parallel, for the score fusion block. This process is performed on 1 of the 15 words ("*perfecta*"), and the rest of conditions were searching for the best accuracy.

Beside, a sweeping between 20 and 400 hidden layer neurons was established with a step of 20 neurons. Meanwhile, another sweeping is responsible for varying the number of principal components used, ranging from 1 to 10 with a step of 1. The result of this test is shown in figure 4.

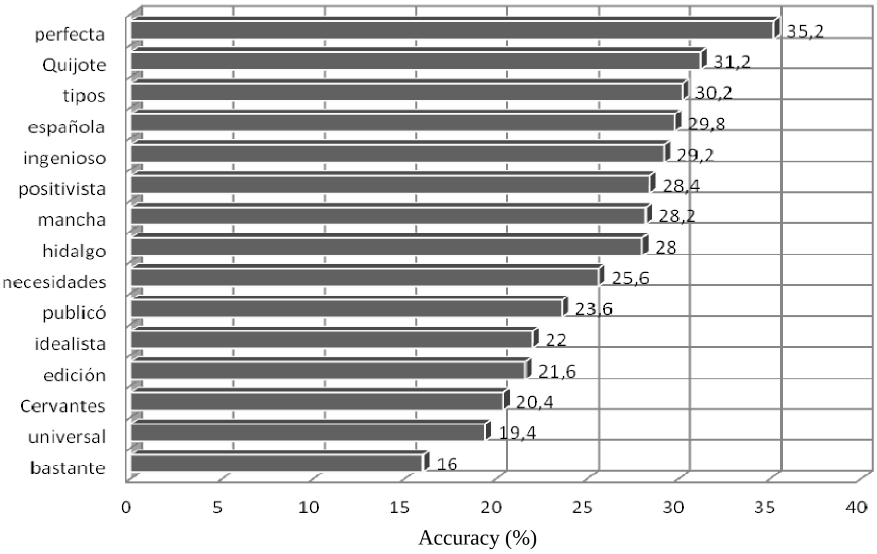


**Fig. 4.** Accuracy for each word

Afterwards, the word fusion was done (data fusion), to try to achieve greater efficiency using the merger of the distances of each of the words. It is applied PCA for the selection of the most discriminative distances, using a total of 2000 iterations and a total of 30 neural networks for the score fusion. Finally, it was varying the number of neurons in the hidden layer from 20 to 300, and varying the number of components

(PC) of the PCA for searching the best combination. The best accuracies of this data fusion are shown in the following table;

**Table 1.** Accuracy of the fusion of words

| Number of words | Neurons on hidden layer | Number of PC | Accuracy |
|---|---|---|---|
| 1 | 278 | 10 | 35,20 % |
| 2 | 191 | 20 | 47,80 % |
| 8 | 67 | 79 | 74,20 % |
| 15 | 53 | 138 | 84,20 % |

Besides, too it is shown the detail of all results for “15 word fused”, as it is observed on the following Fig. 5;

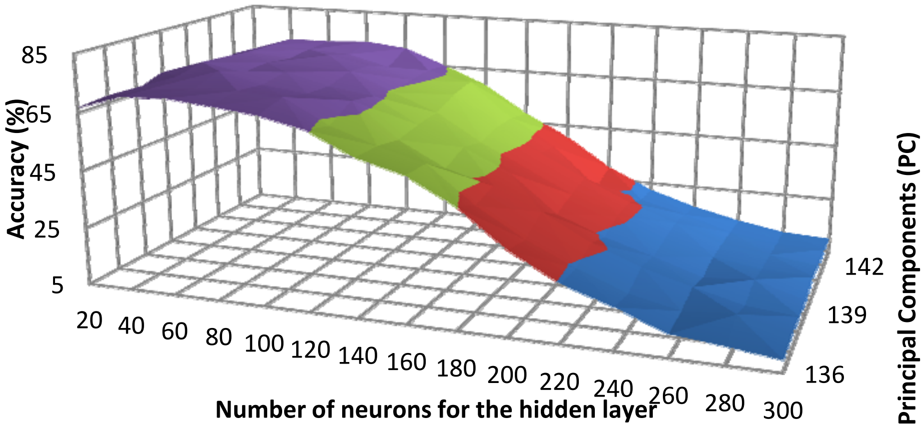


**Fig. 5.** Accuracy for the fusion of 15 words

From the above results, the parameter “index of proportionality” has proved very effective working in isolation. Without fusion, it is achieved up to 35.2% success rate for a single word, and for data fusion of 2, 8 and 15 words, it is reaching up to 47.8%, 74.2% and 84.2%, respectively.

De los parámetros mostrados en [7], el mejor obtenía un 14,4% de acierto, mientras que la proporcionalidad de la palabra es capaz de dar desde un 35,8% para una palabra concreta hasta un 84% tras la fusión de 15 de ellas.

To compare the results versus other parameters, working in isolation, they were compared with [7], because it shows the efficiency of the isolated parameters, different of our proposal; thus allowing a comparative. From parameters in [7], the best on reached up to 14.4%, while our proposal (the index of proportionality) for a word gave up to 35.8%. And up to 84%, for the fusion of 15 words.

## 5 Conclusions and future lines

It is shown that the index of proportionality is a powerful parameter in offline forensic biometric identification techniques, because experimentally, it obtained a better success rates than other parameters of the same nature, as it is achieved in [7] and [15].

It is observed that the word "*perfecta*" achgieved the highest accuracy, and according to its characteristics, this was mainly due to the number of under-words and over-words, which presents; the separation between them, and finally, because it is not a short word. This conclusion will be the rule or condition for implementing this system. The ratification is found in the following word with the second best accuracy, This, which was "*Quijote*", which also meets these characteristics, and which suggests that this should be the type of word used when making an identification based on the proportionality index of the word.

As a future line, it is raised to check more key or control points, add more words for the data fusion, and try to combine it with other parameters used on forensics, to improve the success found.

## Acknowledgment

This work is supported by funds from Spanish Government, under Grant MCINN TEC2012- 38630-C04-03.

## References

1. G. Serratrice, M. Habib. Escritura y cerebro. Ed. Masson, Barcelona. 1997.
2. C.F. Romero, C.M. Travieso, M.A. Ferrer, and J.B. Alonso, Handwriting Knowledge Based on Parameterization for Writer Identification. Proceedings of the European Computing Conference. Chapter 1. Página 3, 2009.
3. A.A. Brink, J. Smit, M.L. Bulacu, L.R.B. Schomaker, Writer identification using directional ink-trace width measurements, Pattern Recognition, Volume 45, Issue 1, DOI:10.1016/j.patcog.2011.07.005.
4. D. Bertolini, L.S. Oliveira, E. Justino, R. Sabourin, Texture-based descriptors for writer identification and verification, Expert Systems with Applications, Volume 40, Issue 6, May 2013, Pages 2069-2080, ISSN 0957-4174, DOI: 10.1016/j.eswa.2012.10.016.
5. R. Jain and D. Doermann, Writer Identification Using An Alphabet Of Contour Gradient Descriptors. Language and Multimedia Processing Laboratory University of Maryland, College, Master Degree, 2013.
6. Rajesh Kumar, Bhabatosh Chanda, J.D. Sharma, A novel sparse model based forensic writer identification, Pattern Recognition Letters, Volume 35, 1 January 2014, Pages 105-DOI: 10.1016/j.patrec.2013.07.001.
7. Z.A. García-Hernández, Cálculo de Parámetros Basados en Ciencias Forenses para la Identificación de Escritores, B.ch. Thesis, ULPGC, 2012.
8. E.C. Herrera. Identificación del autor de un texto manuscrito. Ph.D. Thesis. Mexico, 2008. http://www.saber.cic.ipn.mx/cake/SABERsvn/trunk/Repositorios/webVerArchivo/309/1, (Last visit: 26/06/2014)

9. Schomaker, L., Advances in Writer Identification and Verification, Document Analysis and Recognition (Volume: 2), pp. 701, 2007
10. N. Otsu, A threshold selection *method* from gray-level histograms, IEEE, January 1979
11. Marcelo A. Duarte. Documentología II. Licenciatura en Criminalística, Cátedra. Technical Report, 2010.
12. I.T. Jolliffe, Principal Component Analysis, second edition (Springer). 2002.
13. Kruse, Borgelt, Klawonn, Moewes, *Computational Intelligence: A Methodological Introduction* by Steinbrecher, Held, Springer. 2013.
14. A. Hyvärinen, J. Karhunen, E. Oja, *Independent component analysis*. Ed: Wiley-Interscience. 2001
15. O.Santana, C.M. Travieso, J.B. Alonso, M.A. Ferrer, "Writer Identification based on Graphology Techniques", in *IEEE Aerospace and Electronic System Magazine*, , ISSN: 0885-8985, vol. 25 no. 6, pp. 35-42, June 2010.

# ATVS-UAM System Description for the Albayzin 2014 Audio Segmentation Evaluation

Javier Franco-Pedroso, Elena Gomez Rincon, Daniel Ramos and Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group
Universidad Autonoma de Madrid (UAM). Spain
http://atvs.ii.uam.es
javier.franco@uam.es,elena.gomezr@estudiante.uam.es,daniel.ramos@uam.es,joaquin.gonzalez@uam.es

**Abstract.** This document describes the audio segmentation system developed by the ATVS – Biometric Recognition Group, at Universidad Autonoma de Madrid (UAM), for the Albayzin 2014 Audio Segmentation Evaluation (ASE). This system is based on three independent GMM-UBM acoustic-class detectors based on MFCC-SDC features. Each acoustic-class detector (‘mu’, ‘no’, ‘sp’) evaluates test recordings in a frame-by-frame manner, and the score-streams are filtered and calibrated previous to the detect-decision stage. Although the performance of the independent acoustic-class detectors is far from being perfect in terms of EER, the resulting audio segmentation systems achieves low miss (7.9%), false alarm (10.6%) and class error (3.0%) rates, given a final 21.43% SER on our development subset.

**Keywords:** audio segmentation, MFCC-SDC, GMM-UBM, calibration

## 1 Introduction

In contrast to our previous participation in Albayzin ASE campaigns (the 2010 edition [2]), this year we present a lighter but more robust system that avoids the overfitting introduced by Maximum Mutual Information discriminative training when the available data is scarce. Moreover, the system developed fits better the approach followed in this campaign by the organizers to the problem of evaluating automatic segmentation systems [3]: instead of labeling non-overlapping segments of (maybe overlapped) different acoustic classes, the presence of each acoustic class should be independently annotated in different segments (maybe overlapped with other acoustic classes). Although the problem can be solved from both perspectives (training different models for each possible acoustic-classes combination as we did in 2010 campaign), considering one independent detector for each acoustic class provides a more scalable solution and avoids the constraints regarding the available data for training the acoustic models.

The system developed consists in three independent acoustic-class detectors (speech –‘sp’-, music -‘mu’-, and noise –‘no’-) based on the classical GMM-UBM

adfa, p. 1, 2011.

framework [4]. Each detector performs a frame-by-frame scoring of the test recordings, obtaining one log-likelihood stream per acoustic class. These score-streams are smoothed through a mean filter over a sliding window in order to deal with the high variability of frame-scores. Finally, smoothed frame-scores are independently calibrated by means of a linear logistic regression trained on a subset of the development dataset.

The remainder of this paper is organized as follows. Section 2 describes the feature extraction process. Sections 3 and 4 describe, respectively, the acoustic-class modeling and the acoustic-class detection stage. Section 5 explains the experimental protocol followed, and shows the results obtained in our development subset. Finally, Section 6 summarizes the key points of our submission, exposes the computational requirements and draws some conclusions.

## 2 Feature Extraction

Shifted Delta Coefficients (SDC) [5] 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 capture the different time dependency of the music over the speech or noise. In fact, experiments carried out over a subset of the development tracks revealed that GMM-UBM detectors build from MFCC-SDC features outperform those trained on MFCC plus delta coefficients.

For both development and evaluation tracks, one feature vector was extracted every 10 ms by means of a 20 ms Hamming sliding window (50% overlap). For each window, 7 MFCC features (including C0) were computed from 25 Mel-spaced magnitude filters over the whole available spectrum (0-8000 Hz). These features have been mean-normalized, RASTA filtered and Gaussianized through a 3-second window. Finally, their SDC were computed on a 7-1-3-7 (N-D-P-K) configuration and concatenated with them in a 56-coefficient feature vector.

## 3 Acoustic-Class Modeling

Acoustic classes have been modeled adopting the classical GMM-UBM framework [4] widely used for speaker recognition. First, a 1024-component UBM was trained by means of a 1-iteration k-means initialization followed by a 5-iteration EM stage. For this purpose, one half of the development dataset provided was used (tracks 01-10). Secondly, acoustic-class models were MAP-adapted [4] from this UBM through 1 single iteration and using a relevance factor r=16. Again, tracks 01-10 were used also for this step.

For each acoustic class, training data were extracted from segments belonging to the same acoustic-class as appeared in the provided development labels. This means that, for instance speech segments may contain not only isolated speech but also any of the other acoustic classes overlapped with it. As we are aiming to develop an acoustic-class detector, our assumption is that the acoustic-class models should collect their own acoustic class in any possible condition it may appear. On the other hand, segments where each class can be found isolated are very scarce in the database provided, so robust acoustic-class models cannot be trained from such small amount of data, as we found out in our preliminary experiments.

## 4 Acoustic-Class Detection Stage

Acoustic-class detection stage is based on a frame-by-frame scoring of the test track against every acoustic-class model. Frame-by-frame log-likelihoods are highly variable over time, as it can be seen on Figure 1. For a segment with an isolated acoustic-class, it is expected that the mean log-likelihood will converge to a stable value as long as more frames are incorporated, as it has been shown for the speaker recognition task in [6]. For this reason, these score-streams were smoothed through a mean filter over a sliding window in order to have a more stable frame-score that approaches the "true" score of the acoustic class present in the surrounding frames. Figure 2 shows the result of applying this mean filtering stage for a 700-frame sliding window. The window length was independently optimized for each acoustic-class detector, looking for the length that provides the best detection performance in terms of EER. Results are shown in Figure 3 for our development subset (tracks 11-15).

Finally, the frame-by-frame log-likelihoods were calibrated by means of a linear logistic regression implemented in FoCal toolkit [1]. One different logistic regression is used for each acoustic-class detector, all of them trained on the same development subset used for the window length optimization (tracks 11-15).

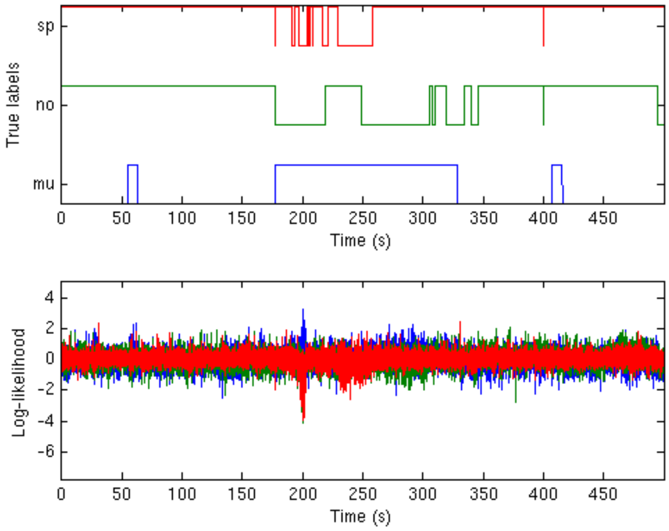


**Fig. 1.** Detail of the frame log-likelihoods for a 500-second segment of track11.

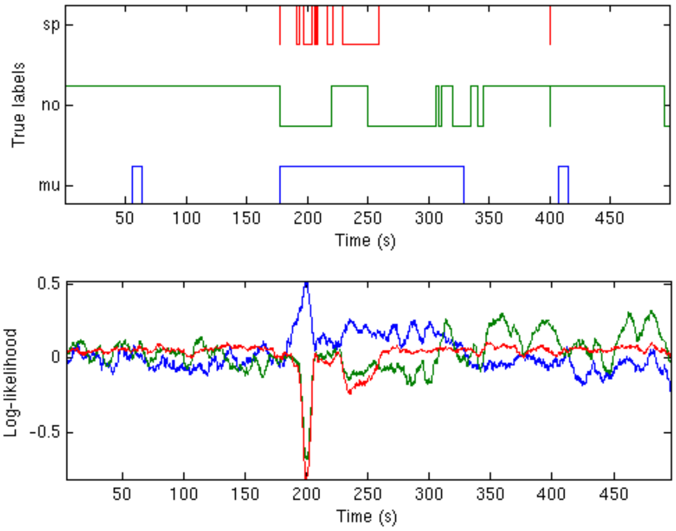


**Fig. 2.** Detail of the frame log-likelihoods for a 500-second segment of track11 after the mean filtering stage.

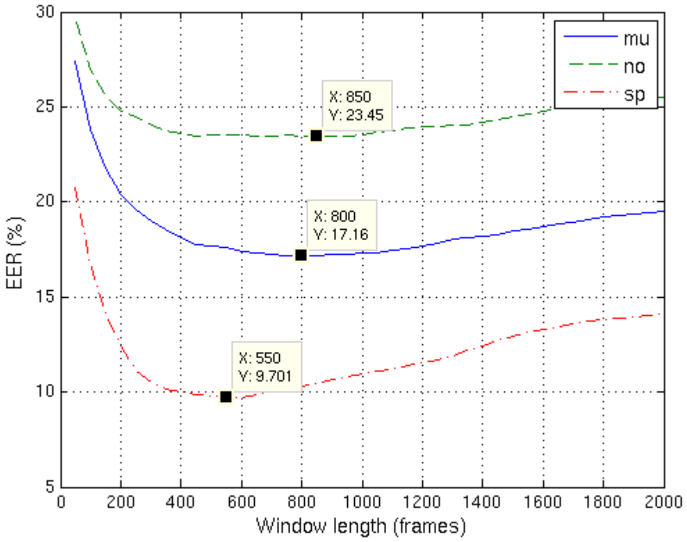


**Fig. 3.** EER as a function of the mean-filtering window-length, obtained for our development subset (tracks11-15). Best results are highlighted (X: window length, Y: EER).

## 5 Experimental setup and development results

Table 1 shows how the development data have been partitioned in order to be used for different purposes. One half of the development dataset has been devoted to train the acoustic models. From the remaining subset, one half has been used to find the optimum window length for the frame-scores mean-filtering, and the resulting frame-scores used to train the calibration rule; the final 5-track subset has been left apart in order to test the developed system.

**Table 1.** Dataset partitioning for system development.

| Purpose | Track numbers |
|---|---|
| UBM training | 01-10 |
| Acoustic-class modeling | 01-10 |
| Window length optimization | 11-15 |
| Calibration training | 11-15 |
| Audio segmentation testing | 16-20 |

Segmentation results obtained for our test subset (tracks 16-20) are shown in Table 2. As it can be seen, in spite of having acoustic-class detectors of relatively low detection performance (9.7% EER for ‘sp’, 17.2% EER for ‘mu’ and 23.4% EER for ‘no’), the whole audio segmentation system achieves good performance compared with results shown in previous Albayzin ASE campaigns.

**Table 2.** Performance of the audio segmentation system: missed class time, false alarm class time, class error time and overall segmentation error, in seconds and percentages.

| Error | Time (s) | % scored class time |
|---|---|---|
| Missed Class | 2262.51 | 7.9 |
| False Alarm Class | 3057.21 | 10.6 |
| Class error | 853.85 | 3.0 |
| **Overall Segmentation Error** | **21.43 %** | |

## 6 Summary and conclusions

ATVS – Biometric Recognition Group has developed an efficient and light audio segmentation system. This system is based on three independent GMM-UBM acoustic-class detectors that can be developed and tuned independently. For instance, detectors in submitted systems make use of a different mean-filtering window-length and independent score-calibration rules, but they could be based in different features as well. Moreover, the adopted approach of modeling broad acoustic classes (‘mu’, ‘no’, ‘sp’) instead of the specific sub-classes given by all the possible combinations (‘mu+no’, ‘sp+no’, etc.) allows to develop a more robust system and avoids overfitting when the available training data is scarce. Finally, it can be seen in Table 3 that the computational requirements in terms of CPU time are very low, allowing the testing to be run in 0.225xRT for each track. Experiments were carried out in a machine equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 16GB of RAM.

**Table 3.** Testing time per track (~60 min) for the different stages and total time as a real-time (xRT) factor.

| **Stage** | **Time** |
|---|---|
| Feature extraction | 19 secs |
| Frame-by-frame scoring | 13 min |
| Scores filtering and calibration | 5 sec |
| Total (xRT) | ~0.225 |

## Acknowledgement

This work has been supported by the Spanish Ministry of Economy and Competitiveness (project CMC-V2: Caracterizacion, Modelado y Compensacion de Variabilidad en la Señal de Voz, TEC2012-37585-C02-01).

## References

1. Niko Brummer, FoCal: toolkit for evaluation, fusion and calibration of statistical pattern recognizers (2008). Online: http://sites.google.com/site/nikobrummer/focal
2. 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 Proceedings of FALA: VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop, 2010, pp. 415–418.
3. Alfonso Ortega, Diego Castan, Antonio Miguel, Eduardo Lleida. The Albayzin 2014 Audio Segmentation Evaluation. Online: http://iberspeech2014.ulpgc.es/images/segm_eval.pdf
4. Reynolds, D., Quatier, T., Dunn, R., Speaker Verification Using Adapted Gaussian Mixture Models Digital Signal Processing, vol. 10, 19–41 (2000).
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. Robbie Vogt and Sridha Sridharan, Minimising Speaker Verification Utterance Length through Confidence Based Early Verification Decisions. Lecture Notes in Computer Science Volume 5558, 2009, pp 454-463.

# GTM-UVigo System for Albayzin 2014 Audio Segmentation Evaluation

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{plopez,ldocio,carmen}@gts.uvigo.es

**Abstract.** This paper describes the GTM-UVigo systems for Albayzin 2014 audio segmentation evaluation, which consist on segmentation followed by classification approaches with the same segmentation stage, but different classification approaches. Segmentation is performed by means of a Bayesian Information Criterion (BIC) strategy featuring a false alarm rejection strategy: the process of acoustic change-points is supposed to follow a Poisson process, and a change-point is discarded with a probability that varies in function of the expected number of occurrences in the time interval formed by the previous and candidate change-points. The classifier of the primary system represents the audio segments in a total variability space and then classifies them using logistic regression; contrastive system 1 represents the audio segments by means of Gaussian mean supervectors and classification is performed using a support vector machine; and contrastive system 2 models the different classes with Gaussian mixture models and performs maximum likelihood classification.

**Keywords:** Audio segmentation, iVectors, false alarm rejection strategy

## 1 Introduction

Audio segmentation is a task consisting on dividing an audio stream into homogeneous regions according to some criteria. Audio segmentation systems can be divided in two groups: those that perform segmentation followed by classification, and those that perform audio segmentation by classification.

Albayzin 2014 audio segmentation evaluation consisted on the detection of speech, music, noise or any combination of these three classes in a set of recordings. This paper describes the audio segmentation systems developed by the GTM-UVigo team for this evaluation, which consist on segmentation followed by classification approaches.

The segmentation stage was carried out by means of the Bayesian information criterion (BIC) approach for acoustic change detection [3]. A technique to reduce the false alarm rate of the BIC algorithm was added to this system [8]: the acoustic change-point process is supposed to follow a Poisson process and, according to this, a candidate change-point that is suspicious of being a false alarm is accepted or rejected with a probability that depends on the length of the observed interval. This observed interval is, in this case, the length of the audio segment that would be formed if the candidate change-point were accepted.

Three systems with different classification approaches were submitted. The primary system represents the audio segments in a total variability subspace, also known as iVector representation [4], and classification is performed by means of logistic regression [2]. In contrastive system 1, classification is performed using a support vector machine (SVM); to do so, each audio segment is represented by means of a Gaussian mean supervector obtained by adapting a universal background model (UBM) to the feature vectors of the audio segment [10]. In contrastive system 2, a Gaussian mixture model (GMM) is trained for each class and the likelihood of a speech segment with each model is computed, selecting the one that obtains the maximum likelihood.

The rest of this paper is organized as follows: Section 2 presents an analysis of the database used in this work; Section 3 describes the audio segmentation approach in detail; Section 4 presents some preliminary results obtained on the training data; and Section 5 depicts some conclusions and future work.

## 2 Preliminary analysis of the database

Table 1 shows an overview of the two datasets of the Albayzin 2014 audio segmentation evaluation. As there is not a development partition, we decided to perform four different experiments in order to tune the parameters of the proposed system: four partitions of five recordings each were made and, on each experiment, three partitions were used for training while the remaining one was used for testing. Once the parameters of the system were tuned, the whole training dataset was used to train the system.

**Table 1.** Summary of the datasets of Albayzin 2014 audio segmentation evaluation.

| Dataset | # recordings | Duration |
|---|---|---|
| Training | 20 | 21 h 16 min 11 s |
| Test | 15 | 15 h 37 min 50 s |

Before developing the system, an analysis of the training data was performed in order to make some design decisions. Albayzin 2014 audio segmentation evaluation consisted on detecting when the classes speech, music and noise were present, which can appear individually or simultaneously. The first design decision consisted on, instead of detecting each class individually, defining a set of seven classes: speech (s00), music (0m0), noise (00n), speech with music (sm0), speech with noise (s0n), music with noise (0mn) and speech with music and noise (smn). Figure 2 shows the percentage of time that each of this seven classes appears on the training data: it can be seen that there is almost no data of classes noise and music with noise, as they appear in less than 1% of the whole training data. Thus, as the amount of data for these classes was two little to properly train a classifier that detects them, we decided to ignore these classes and keep on with the remaining five.

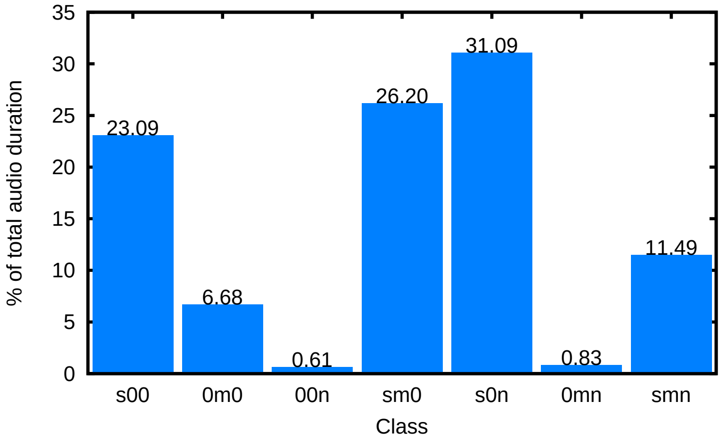


**Fig. 1.** *Distribution of the duration of the different classes on the training dataset of the database.*

## 3 System description

The audio segmentation systems presented in this paper have two main stages: segmentation and classification. First, the audio stream is segmented, and then these segments are classified using a classification approach. The specific techniques developed for this work are described in the rest of this Section.

### 3.1 Segmentation

Before performing segmentation, features were extracted from the waveform; specifically, 12 Mel-frequency cepstral coefficients (MFCCs) plus energy were obtained, leading to feature vectors of dimension $N = 13$. The features were computed using a 25 ms window and a time step of 10 ms, and cepstral mean substraction was applied, computing the mean using all the frames in each file.

The segmentation approach used in this system has three main steps:

- Coarse segmentation. A Bayesian information criterion (BIC) approach is applied in order to select candidate change-points. The BIC criterion is a hypothesis test to decide whether there is a change-point in a window of data ($\mathrm{H}_1$) or not ($\mathrm{H}_0$) by observing a value $\Delta$BIC: $\Delta\mathrm{BIC} > 0$ means that hypothesis $\mathrm{H}_1$ is stronger than hypothesis $\mathrm{H}_0$, i.e. there is a change-point in the window; $\Delta\mathrm{BIC} \leq 0$ means that there is no change-point in the window. A BIC segmentation system as described in [3] was implemented to perform audio segmentation: a window of data that slides and grows is analyzed in order to detect a candidate change-point in it by applying the BIC criterion [11]. The BIC algorithm has a tuning parameter $\lambda$, which was tuned on the training dataset.

- Change-point refinement. Anytime a candidate change-point is found, a fixed-size window is centered on this change-point and the BIC criterion is applied again in order to refine its position or to discard it. If the change-point is discarded, the system returns to the coarse segmentation stage.
- False alarm rejection strategy. A technique to reduce the number of false alarms was implemented on the BIC algorithm [8]. In this strategy, it is assumed that the change-point occurrences 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 [1].

  An homogeneous Poisson process is characterized by its rate $\gamma = \frac{1}{\tau}$, where $\tau$ is the mean time between occurrences (this rate is usually represented as $\lambda$, but in this paper we will refer to it as $\gamma$ in order to avoid confusions with the penalty of the BIC strategy). In a Poisson process, the probability of having $n$ occurrences in a time interval $t$ is:

$$p_n(t) = \frac{\gamma^n e^{-\gamma}}{n!} \tag{1}$$

  Thus, the probability of not having an occurrence in a time interval $t$ is

$$p_0(t) = e^{-\gamma t} \tag{2}$$

  and the probability of having one or more occurrences in a time interval $t$ is

$$p_{\bar{0}}(t) = 1 - p_0(t) = 1 - e^{-\gamma t} \tag{3}$$

  The specific false-alarm rejection strategy is depicted in Figure 3.1: given the $\Delta$BIC value obtained when refining the candidate change-point, we consider that the greater this value, the more likely the candidate change-point is a true change-point. Thus, if $\Delta$BIC is lower than a threshold $\Theta_{\mathrm{BIC}}$, we consider this change-point as suspicious of being a false alarm. If this happens, we will discard this change-point with probability $p_{\mathrm{discard}}$ which, in this case, is equal to $p_0(t)$, where $t$ is the time interval between the last confirmed change-point and the suspicious change-point. The value of $\tau$ was estimated as the median of the segment duration on the training dataset, and $\Theta_{\mathrm{BIC}}$ was tuned on that data.

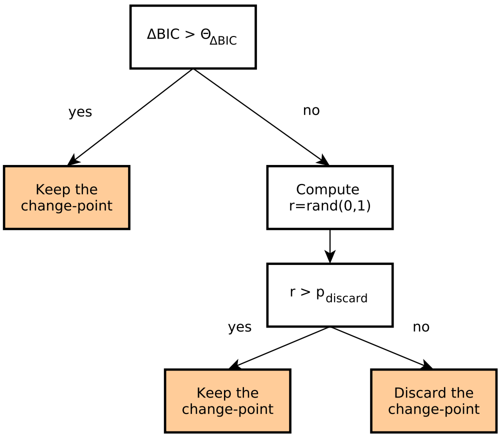


**Fig. 2.** *Block diagram of the proposed audio segmentation system.*

### 3.2 Classification

Different classifiers were developed using the segments obtained from the previous step, which are described below.

**Primary system: iVector representation with logistic regression.** Perceptual linear prediction (PLP) analysis was used to extract 13 cepstral coefficients, which were combined with two pitch features as described in [6], and augmented with their delta features. Hence, the dimension of the feature vectors was 13 PLP + 2 pitch + $\Delta$ = 30 features. After extracting the feature vectors, the segments were represented in a total variability subspace: given a Universal Background Model (UBM) with $N$ mixtures, this UBM is adapted to the feature vectors of each segment using Maximum a Posteriori (MAP) adaptation, and the means of the resulting Gaussian Mixture Model (GMM) are concatenated in order to obtain a Gaussian mean supervector for each segment. The iVector technique is applied to the Gaussian mean supervectors, which defines a low-dimensional space, named total variability space, in which the audio segments are represented by a vector of total factors, namely iVector [4]. A Gaussian mean supervector $\mathbf{M}$ is decomposed as follows:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{4}$$

where $\mathbf{m}$ is the speaker and channel independent supervector, $\mathbf{T}$ is a low-rank total variability matrix, and $\mathbf{w}$ is the iVector corresponding to the Gaussian mean supervector. In this training stage, the matrix $\mathbf{T}$ is trained as described in [7].

Once the total variability matrix $\mathbf{T}$ is obtained, this matrix and the UBM can be used to extract iVectors from the acoustic features corresponding to the different segments. In this system, the iVectors were classified using logistic regression with L-BFGS method [2]. Feature extraction, iVector representation and logistic regression were performed using the Kaldi toolkit [9]; the UBM had 512 mixtures and the dimension of the iVectors was set to 400.

**Contrastive system 1: Gaussian mean supervector representation with support vector machine classification.** The feature vectors used in this classifier were the 13 MFCCs described in Section 3.1, augmented with their delta and acceleration coefficients, leading to feature vectors of dimension 39.

Before classifying the segments obtained from the previous stage, they must be represented in a proper way. In this system, we chose to use a Gaussian mean supervector (SV) representation: a universal background model (UBM), which is a Gaussian mixture model (GMM) with $M$ mixtures, is adapted to the feature vectors of the audio segment [10], and the obtained means are concatenated, forming a supervector of dimension $M \times N$. In this system, the number of mixtures of the UBM was 256, leading to supervectors of dimension 9984.

Classification was performed employing a support vector machine (SVM), which was trained using a set of supervectors and their groundtruth class labels. Specifically, an SVM with a linear kernel and L2-regularized logistic regression was trained for this task, and its cost parameter was tuned on the training data. Once the SVM is trained, it can be used to obtain the class labels of the test audio segments. SVM training and classification were performed using the library for large linear classification LIBLINEAR [5].

**Contrastive system 2: GMM-maximum likelihood classification.** The audio was represented by means of the 13 PLP cepstral features combined with pitch features and delta coefficients used in the primary system. Classification was performed doing maximum likelihood classification using Gaussian mixture models (GMMs). A GMM of 512 mixtures was trained for each of the five classes mentioned above and, for each segment to classify, the log-likelihood between the feature vectors of the segment and each of the GMMs was computed, selecting the class that achieved the highest log-likelihood [10]. Feature extraction, training and log-likelihood computation were performed using the Kaldi toolkit [9]; the training of the GMMs was performed by doing MAP adaptation of a universal background model (UBM) with full-covariance matrix.

## 4 Preliminary results

This Section describes different experiments that were performed to make design decisions about the segmentation and classification stages. The four experiments mentioned in Section 2, which are summarized in Table 2, were performed using different audio segmentation approaches. The experimental results are presented in function of the missed class time (MCT), false alarm class time (FACT), class error time (CET) and segmentation error rate (SER).

**Table 2.** Description of the four audio segmentation experiments. Recording XX stands for file "trackXX" of the training dataset.

| Experiment | Training recordings | Test recordings |
|---|---|---|
| 1 | 01-15 | 16-20 |
| 2 | 06-20 | 01-05 |
| 3 | 01-05, 11-20 | 06-10 |
| 4 | 01-10, 16-20 | 11-15 |

Tables 3 and 4 show the audio segmentation performance achieved when using the classic BIC segmentation approach and when applying the proposed false-alarm rejection strategy, respectively. The classifier used in these experiments was the Gaussian mean supervector representation with SVM classification. Comparing the two Tables, it can be seen that the false-alarm rejection strategy obtained a reduction of the SER by 1% or more in all the experimental cases, proving the validity of the proposed technique. Table 4 also shows that, on two experiments, a SER below 14% was obtained, while a SER by 16% was obtained on the remaining two, leading to a SER of around 15% on the whole training data.

**Table 3.** Audio segmentation results on the training data using the BIC segmentation stage and Gaussian mean supervector SVM classification.

| Experiment | MCT | FACT | CET | SER |
|---|---|---|---|---|
| 1 | 6.0% | 7.3% | 3.0% | 16.30% |
| 2 | 5.9% | 6.2% | 2.7% | 14.75% |
| 3 | 6.1% | 8.1% | 4.3% | 18.54% |
| 4 | 7.1% | 5.8% | 2.5% | 15.39% |
| Total | 6.3% | 6.8% | 3.2% | 16.27% |

Tables 5 and 6 show the results achieved when performing classification using the iVector representation with logistic regression and the GMM-maximum likelihood classification, respectively. Comparing these Tables with Table 4, we can observe that the best audio segmentation results were obtained when using the iVector representation with logistic regression, which improved the results obtained with Gaussian mean supervector SVM classification by 1.5%. A general improvement of all the types of errors was performed, but it can be noted that the lowest false alarm class time was achieved with the GMM-maximum likelihood classifier.

**Table 4.** Audio segmentation results on the training data using the BIC segmentation stage with the Poisson-based false alarm rejection strategy and Gaussian mean supervector SVM classification.

| Experiment | MCT | FACT | CET | SER |
|---|---|---|---|---|
| 1 | 7.0% | 6.8% | 2.2% | 15.89% |
| 2 | 6.3% | 5.5% | 1.9% | 13.65% |
| 3 | 6.3% | 6.6% | 3.5% | 16.39% |
| 4 | 7.3% | 4.8% | 1.8% | 13.85% |
| Total | 6.4% | 6.2% | 2.3% | 14.96% |

**Table 5.** Audio segmentation results on the training data using the BIC segmentation stage with the Poisson-based false alarm rejection strategy and iVector representation with logistic regression.

| Experiment | MCT | FACT | CET | SER |
|---|---|---|---|---|
| 1 | 6.8% | 5.7% | 1.4% | 13.87% |
| 2 | 6.2% | 4.8% | 1.7% | 12.72% |
| 3 | 6.1% | 5.4% | 2.2% | 13.73% |
| 4 | 5.0% | 6.7% | 1.7% | 13.40% |
| Total | 6.0% | 5.6% | 1.8% | 13.43% |

**Table 6.** Audio segmentation results on the training data using the BIC segmentation stage with the Poisson-based false alarm rejection strategy and GMM-maximum likelihood classification.

| Experiment | MCT | FACT | CET | SER |
|---|---|---|---|---|
| 1 | 6.9% | 4.8% | 2.7% | 14.46% |
| 2 | 6.0% | 4.8% | 2.7% | 13.45% |
| 3 | 7.2% | 4.8% | 3.8% | 15.77% |
| 4 | 7.9% | 4.1% | 2.4% | 14.29% |
| Total | 7.8% | 4.4% | 3.0% | 15.16% |

## 5 Conclusions and future work

This paper described the GTM-UVigo system developed for Albayzin 2014 audio segmentation classification, which consisted on a segmentation followed by classification approach.

The segmentation stage introduced a false alarm rejection strategy which was based in the principle that the occurrence of acoustic change-points follow a Poisson process. Preliminary results on the training data showed that the use of the proposed false alarm rejection strategy led to a 1.5% reduction of the segmentation error rate with respect to the classic BIC approach.

Three different classification stages were submitted, which used different audio segment representation and classification approaches. The best results were obtained when using feature vectors with cepstral PLPs and pitch features represented in a total variability subspace, and performing classification by means of logistic regression. Further analysis must be performed in order to extract stronger conclusions about the segment representation and the classification techniques, as each classifier used a different feature representation, making it difficult to conclude whether the results depend on the representation, on the classification approach or on both of them.

The classification approach that obtained the highest segmentation error was the one that obtained the lowest false alarm class time, which leads to believe that a fusion of the different classification approach may result in a reduction of the segmentation error. Thus, we plan to perform fusion experiments using the different strategies presented in this work in order to improve the audio segmentation performance.

## References

1. Allen, A.O.: Probability, Statistics, and Queueing Theory with Computer Science Applications. Academic Press, second edn. (1990)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York, NY, USA (1995)
3. Cettolo, M., Vescovi, M.: Efficient audio segmentation algorithms based on the BIC. In: Proceedings of ICASSP. vol. VI, pp. 537–540 (2003)
4. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing (2010)
5. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification,. Journal of Machine Learning Research 9, 1871–1874 (2008)
6. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2494–2498 (2014)
7. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing 13(3), 345–354 (2005)

8. Lopez-Otero, P., Fernndez, L.D., Garca-Mateo, C.: Novel strategies for reducing the false alarm rate in a speaker segmentation system. In: ICASSP. pp. 4970–4973. IEEE (2010)
9. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
10. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)
11. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6, 461–464 (1978)

# Albayzin 2014 Evaluation: TES-UAB Audio Segmentation System

Héctor Delgado and Javier Serrano

CAIAC, Autonomous University of Barcelona, Spain
{hector.delgado,javier.serrano}@uab.cat

**Abstract.** This paper describes the audio segmentation system developed by Transmedia Catalonia / Telecommunication and Systems Engineering Department, at the Autonomous University of Barcelona (UAB), for the Albayzin 2014 Audio Segmentation Evaluation. The evaluation task consists in segmenting spoken audio documents into three different acoustic classes (speech, background noise, and music), taking into account that more than one class may be present at any given time instant. Furthermore, additional difficulty has been added by fusing and merging audio from different databases. The proposed system is based on the recently presented "Binary Key" modeling approach, originally developed for speaker recognition, but successfully applied to other pattern recognition tasks, such as speaker diarization, emotion recognition and speech activity detection. Experiments carried out on the provided development data show a Segmentation Error Rate of 22.71%.

**Keywords:** audio segmentation, binary key, binary key background model

## 1 Introduction

Audio segmentation is the task of detecting the boundaries between different acoustic sources or classes within an audio signal. Over the years, audio segmentation has become an important task as a pre-processing tool for subsequent speech related tasks, such as Automatic Speech Recognition (ASR), speaker diarization, or Spoken Document Retrieval (SDR). Accurate audio segmentation labels are required to assure success of further systems.

In the last three editions of the "Jornadas en Tecnologías del Habla", audio segmentation evaluations have been conducted in the ambit of the Albayzin Evaluations. These evaluations aim at promoting research in the field of audio and speech processing, including audio segmentation, speaker diarization, language recognition and search on speech. With regard to audio segmentation, past evaluations have shown that the challenge is still far from being completely solved.

Recently, a speaker modeling technique called "binary key" was introduced in [4]. The approach provides a compact representation of a speaker model through a binary vector (vector only containing zeros and ones) by transforming the continuous acoustic space into a discrete binary one. The technique has also been successfully applied to speaker diarization [5], emotion recognition [6], and Speech Activity Detection (SAD) [7]. This latter work is specially interesting in this ambit since it proposes a novel SAD approach achieving state-of-the-art performance. In fact, SAD can be considered as a particular case of audio segmentation, where speech and nonspeech acoustic classes are considered. Then, it may seem reasonable to think that this SAD approach may be useful for audio segmentation tasks involving more audio classes, such as speech, music, background noise, and combinations of all of them. Following these thoughts, an audio segmentation system based on binary keys has been developed to be evaluated in the Albayzin audio segmentation evaluation.

The paper is structured as follows: Section 2 gives an overview of the Albayzin 2014 audio segmentation evaluation. Section 3 describes the audio segmentation system based on binary keys. Section 4 describes the experimental setup and results. Section 5 concludes and proposes future work.

## 2 Audio segmentation evaluation

This section briefly describes the Albayzin 2014 audio segmentation evaluation (refer to [10] for an in-depth description).

As in the 2012 Audio Segmentation Evaluation, the task consists in segmenting a set of broadcast audio documents into segments according to a series of audio classes. These classes are speech, music, and noise. However, combinations of the three classes can occur in the audio being evaluated (overlapping classes). Therefore, a multiple layer labeling must be provided by the segmentation system.

For this evaluation campaign, the main change is related to the audio data to be processed. The test data consist of audio from different merged, or even overlapped, databases. This modification drastically increases the difficulty of the task and has as main aim to test the robustness of systems across different acoustic conditions.

### 2.1 Database description

The database proposed for this evaluation is a combination and fusion of three databases.

The first database is a broadcast news database from the 3/24 TV channel. The database was recorded under the Tecnoparla project [2] and contains around 87 hours of recordings.

The second dataset is the Aragón Radio database from the Corporación Aragonesa de Radio y Televisión, which provided the data for the Albayzin 2012 evaluation.

The third database is composed of environmental sounds from Freesound.org [1] and HuCorpus [9] among others. These sounds are merged with segments from the two previous databases.

All the data are supplied in PCM format, 1-channel, little endian 16 bit-per-sample, 16 KHz sampling rate.

### 2.2 Segmentation scoring

To evaluate systems, the Segmentation Error Rate (SER) is computed as the fraction of correctly attributed class time. This score is computed over the entire file to be processed, including regions containing overlapped classes. The metric is calculated as the Diarization Error Rate (DER) proposed in the NIST RT Evaluations [3].

Given a test dataset $\varOmega$, each document is divided into contiguous segments at all class change points. Then, the segmentation error time $\varXi$ is computed for each segment $n$ as

$$\varXi(n) = T(n)[\max(N_{ref}(n), N_{sys}(n)) - N_{correct}(n)] \tag{1}$$

where $T(n)$ is the duration of segment $n$, $N_{ref}(N)$ is the number of reference classes that are present in segment $n$, and $N_{Correct}(n)$ is the number of reference classes in segment $n$ correctly assigned by the segmentation system. Then, SER is calculated as

$$SER = \frac{\sum_{n \in \varOmega} \varXi(n)}{\sum_{n \in \varOmega}(T(n)N_{ref}(n))} \tag{2}$$

SER includes three types of error, namely the Class Error Time, the Missed Class Time, and the False Alarm Class Time. The Class Error Time refers to the amount of time which has been assigned to an incorrect class. The Missed Class Time is the amount of time that a class is present but not labeled by the system. And the False Alarm Class Time refers to the amount of time which has been assigned to a class that is not present in the reference.

In order to take into account possible uncertainty and reference inconsistencies due to human annotations, a forgiveness collar of 1 second is applied to all reference boundaries.

## 3 Audio segmentation system description

The proposed audio segmentation system is inspired in the SAD system developed in [7], and adapted to the needs of the audio segmentation task of this evaluation.

The binary key modeling aims at transforming the input acoustic data into a binary representation, called binary key, which contains class-specific information, and therefore it is useful for discriminating between acoustic classes. This transformation is done thanks to a UBM-like model called Binary Key Background Model (KBM). Once the binary representation of the input audio is obtained, subsequent operations are performed in the binary domain, and calculations mainly involve bit-wise operations between pairs of binary keys.

### 3.1 KBM training

In this paper, the KBM is trained as follows (figure 1): First, a GMM is trained for each acoustic class (e.g., “speech”, “noise”, “music”) using Expectation-Maximization (EM) algorithm with appropriate labeled training data. Then, the final KBM is the result of pooling all Gaussian components of the individual GMMs together. As an example, a KBM build from three classes 16-component GMMs will contain 32 Gaussian components in total.

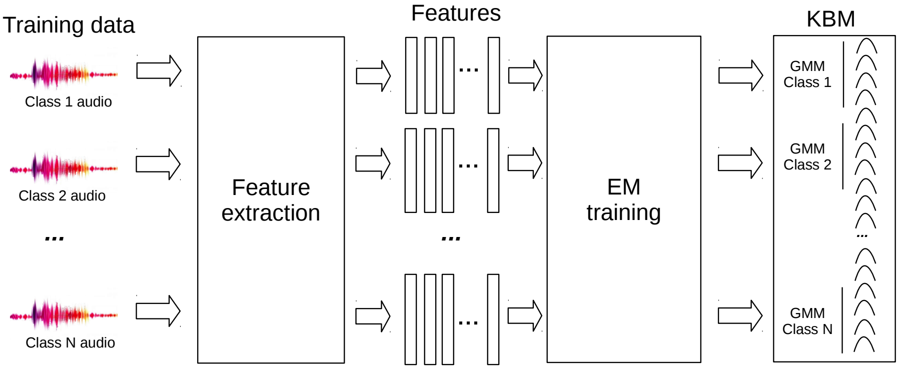


**Fig. 1.** KBM training process.

### 3.2 Binary Key computation

Once the KBM is obtained, any set or sequence of acoustic feature vectors can be converted into a Binary Key (BK). A BK $\mathrm{v}_f = \{v_f[1], ..., v_f[N]\}, v_f[i] = \{0, 1\}$ is a binary vector whose dimension $N$ is the number of components in the KBM. Setting a position $v_f[i]$ to 1 (TRUE) indicates that the $i$th Gaussian of the KBM coexists in the same area of the acoustic space as the acoustic data being modeled. The BK can be obtained in two steps. Firstly, for each feature vector, the best $N_G$ matching Gaussians in the KBM are selected (i.e., the $N_G$ Gaussians which provide highest likelihood for the given feature), and their identifiers are stored. Secondly, for each component, the count of how many times it has been selected as a top component along all the features is calculated, conforming a Cumulative Vector (CV). Then, the final BK is obtained by setting to 1 the positions of the CV corresponding to the top $M$ Gaussians at the whole feature set level, (i.e., the $M$th most selected components for the given feature set). Note that this method can be applied to any set of features, either a sequence of features from a short audio segment, or a feature set corresponding to a whole acoustic class cluster.

### 3.3 Audio segmentation process

The audio segmentation process is illustrated in figure 2. First of all, the input feature vectors must be converted to a sequence of binary keys. The input data are divided into fixed-length segments, considering some overlap and window

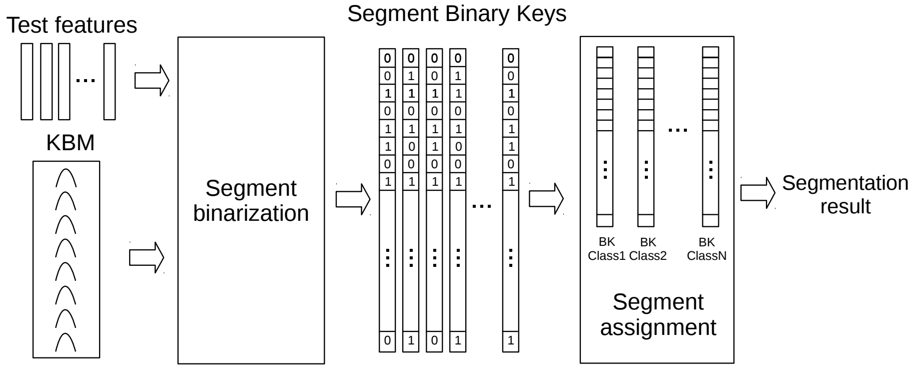


**Fig. 2.** Segmentation process

rate. Then, a BK is obtained for each segment by using the KBM following the method explained in section 3.2. From here on, all operations are performed in the binary domain. Segment assignment is done by comparing each segment BK with the $N$ BKs (previously estimated using the KBM and training data) for each of the $N$ target audio classes. Finally, the current segment is assigned to the class which maximizes the similarity between the BKs pair. The similarity between two binary keys $a$ and $b$, according to [5] is computed as

$$S(\mathrm{a},\mathrm{b}) = \frac{\sum_{i=1}^{N}(a[i] \wedge b[i])}{\sum_{i=1}^{N}(a[i] \vee b[i])} \tag{3}$$

where $\wedge$ indicates the boolean AND operator, and $\vee$ indicates the boolean OR operator. This is a very fast, bit-wise operation between two binary vectors.

In addition, alternatives to the similarity calculation involving CVs are also tested in this work. First, the Intersection and Symmetric Differences Similarity, proposed in [8], is defined as

$$S_{ISDS}(\mathrm{A},\mathrm{B}) = \frac{\sum_{i=1}^{|A\cap B|} a_i + b_i}{(\sum_{j=1}^{A-B} a_j + \sum_{j=1}^{B-A} b_j) * \sum_{i=1}^{|A\cap B|} |a_i - b_i|} \tag{4}$$

where $\{\forall a \in A, \forall b \in B | A - B \neq \emptyset, \exists a \neq b | (a,b) \in A \cap B\}$. Here, the binary vectors act as indexes for the calculations with the cumulative vectors.

Finally, a simple cosine similarity between CVs is tested as well:

$$S_{cos}(a,b) = \frac{a \cdot b}{||a||\,||b||} \tag{5}$$

where $a$ and $b$ are the CVs being compared.

## 4 Experiments and results

As in the 2012 audio segmentation evaluation, a multi-layer labeling is requested when overlapped classes are present. However, in this system all possible combinations of the three proposed classes (speech, noise, and music) are taken as separated classes as a starting point. At the end, the obtained segmentation is post-processed in order to get the final multi-layer labeling.

This year, the UAB group is submitting two different systems. They mainly differ in the way the KBM is obtained, keeping the rest of the setting unaltered for both systems. These common settings are explained next.

First, the provided database, which consists of 20 audio excerpts of around 1 hour each one, is divided into two subsets. The first one is composed of the first 14 audio files (around 70% of the corpus) and it is used for training. The rest (6 audio files conforming the remaining 30%) is used for testing.

For feature extraction, LFCCs are extracted from the audio signal using a 20ms analysis frame, a shift of 10ms, and a Hamming window. 12 static coefficients are extracted plus the energy coefficient, delta, and delta-delta coefficients, totaling 39 coefficients. The tool used for feature extraction is the SPro toolkit (`https://gforge.inria.fr/projects/spro/`).

Regarding binary key computation, the top 5 Gaussian components are taken in a frame basis. Several values of the factor of top Gaussians at segment level $M$ are tested in the experiments (0.1 and 0.15).

Finally, in the data assignment stage, binary keys are computed for each 0.3s segment, augmenting it 2.5s before and after, totaling 5.3s. This is done in order to have sufficient data to estimate the BKs, but also for avoiding very over-segmented labels. Then, the window is shifted 0.3s to calculate the next BK.

### 4.1 Primary and contrastive systems

As said above, the two systems share a common setting, but differ in the way the KBM is obtained.

In the primary system, all combinations of the three proposed acoustic classes are considered, totaling 7 combinations. Therefore, seven GMMs are trained ("sp", "no", "mu", "sp+no", "sp+mu", "sp+no+mu", "no+mu"), and the final KBM is the result of pooling all Gaussian components. However, in the contrastive system, only the three proposed classes are considered (speech, noise, and music). Therefore, in this case three GMMs are trained.

After training the KBM, in both systems BKs are estimated for the 7 combinations, resulting in 7 BKs which act as acoustic models for each class combination. Note that this is done for both systems, regardless of the number of classes used to conform the KBM. In order to clarify this, table 1 summarizes the number of components of KBM depending on the number of classes and the number of individual GMM components.

### 4.2 Experimental results and discussion

Table 2 and table 3 show the SER of the primary and contrastive systems, respectively, for different KBM sizes, different values of $M$, and the different proposed similarity measures, evaluated on the test dataset (note that this test dataset

**Table 1.** Number of components of KBM depending on the number of classes being considered.

| Primary system (7 classes) | | Constrastive system (3 classes) | |
|---|---|---|---|
| Components per class | KBM components | Components per class | KBM components |
| 128 | 896 | 128 | 384 |
| 256 | 1792 | 256 | 768 |
| 512 | 3584 | 512 | 1536 |
| 1024 | 7168 | 1024 | 3072 |

is extracted from the development files provided, as the official test ground-truth segmentation labels of the evaluation were not available at the moment of writing this paper). The best performing configuration of the primary system comprises a 3584-component KBM (i.e. 512 Gaussians per class combination), $M = 0.1$, and using the cosine similarity, providing an overall SER of 22.71%. The rest of configurations performances oscillate between 23% and 28% SER in the primary system, and between 26% and 30% in the contrastive system. It is also observed that the choice of similarity measure has more impact in performance that the value of $M$. The best performing similarity measure is the cosine similarity, followed by the ISDS similarity and the similarity given by equation 3.

It also can be seen that the primary system outperforms the contrastive one, even using a lower number of Gaussian components.

**Table 2.** SER of primary system on the test dataset, according to the number of KBM components, the factor $M$ of top Gaussians per segment, and the used similarity metric. Best results for each similarity measure are highlighted.

| SER of primary system (%) | | | | | |
|---|---|---|---|---|---|
| KBM | $S$ | | $S_{ISDS}$ | | $S_{cos}$ |
| components | $M = 0.1$ | $M = 0.15$ | $M = 0.1$ | $M = 0.15$ | - |
| 896 | 28.87 | 28.23 | 26.13 | 26.52 | 24.29 |
| 1792 | 28.73 | 28.49 | 25.58 | 26.72 | 23.76 |
| 3584 | 28.28 | **26.17** | **24.71** | 25.66 | **22.71** |

Table 4 gives individual results for each audio file with the best-performing configuration of the primary system, by breaking down SER into Miss Class Time, False Alarm Class Time, and Class Error Time. In general, miss errors become the most contributing part of the total error, with rates between 9.5% and 12.3%, and an overall rate of 10.4%. False alarm errors are lower than miss errors, but quite high for some audio files (up to 12.8%), totaling an overall rate of 6.8%. Finally, class errors are also lower than miss errors, and slightly lower than false alarm errors, with values oscillating between 2.8% and 8.5%.

**Table 3.** SER of contrastive system on the test dataset, according to the number of KBM components, the factor $M$ of top Gaussians per segment, and the used similarity metric. Best results for each similarity measure are highlighted.

| SER of contrastive system (%) | | | | | |
|---|---|---|---|---|---|
| KBM | $S$ | | $S_{ISDS}$ | | $S_{cos}$ |
| components | $M = 0.1$ | $M = 0.15$ | $M = 0.1$ | $M = 0.15$ | - |
| 768 | 29.62 | 30.08 | 27.49 | 28.5 | 26.73 |
| 1536 | 28.32 | 28.35 | **26.86** | 27.13 | **25.0** |
| 3072 | **28.04** | 28.7 | 27.2 | 27.73 | 25.64 |

**Table 4.** Most accurate system results per audio file, broken-down into error types: Miss Class Time (Miss), False Alarm Class Time (FA), Class Error Time (Class), and Segmentation Error Rate (SER).

| File ID | Miss | FA | Class | SER |
|---|---|---|---|---|
| track15 | 10.0 | 9.1 | 3.0 | 22.14 |
| track16 | 10.3 | 6.0 | 7.0 | 23.39 |
| track17 | 10.0 | 4.3 | 2.8 | 17.12 |
| track18 | 12.3 | 4.4 | 7.2 | 23.7 |
| track19 | 10.1 | 5.2 | 5.2 | 20.47 |
| track20 | 9.5 | 12.2 | 8.5 | 30.21 |
| Overall | 10.4 | 6.8 | 5.6 | 22.71 |

After analyzing results of primary and contrastive systems on development data, the best performing parameter settings are taken to be used to process the official evaluation test dataset. The setting for the primary system is 3584 KBM components and cosine distance. Regarding the contrastive system, 1536 KBM components and the cosine distance are selected.

By using the selected settings, the test dataset is then processed. The system in which audio segmentation was performed is a Debian Wheezy virtual machine with 12 assigned GB RAM, running on an Intel Xeon E5-2420 at 1.90GHz CPU. Table 5 shows execution time and real time factor (xRT) for both primary and contrastive systems. During the experiments, it has been observed that the most time consuming part of the segmentation systems is the log-likelihood computation of all the input features for each Gaussian components, needed to estimate the binary keys. This stage is speeded up significantly when the KBM size decreases. After this stage, data assignment is a very fast stage.

**Table 5.** CPU time (hh:mm:ss) and Real Time Factor (xRT) of primary and contrastive systems on the official test data (total time is 15:37:43).

| | Primary system | | Constrastive system | |
|---|---|---|---|---|
| Task | Time | xRT | Time | xRT |
| Feature extraction | 00:02:17 | 0.002 | 00:02:17 | 0.002 |
| Audio segmentation | 07:13:02 | 0.462 | 03:11:38 | 0.204 |
| Overall | 07:15:20 | 0.464 | 03:13:55 | 0.207 |

Primary system presents an overall xRT of 0.464. Although faster than real-time, this execution time could be too long for some time-critical applications. Contrastive system shows a xRT of 0.207, which is significantly lower than the primary system (more than twice faster). Although experimental results have shown weaker performance than the primary system, the contrastive system could be useful when higher speed is required, at the cost of a slight decrease of accuracy.

## 5 Conclusions

An audio segmentation system based on binary key modeling has been developed an submitted to the Albayzin 2014 audio segmentation evaluation. The system performs audio segmentation by annotating the input data according to all possible combinations of the three proposed audio classes, and finally the obtained labels are post-processed in order to get the final, multi-layer labeling. The proposed approach is based on the Binary Key modeling, and has been tested with a primary system and a contrastive system. Those systems only differ in the way the KBM is trained by considering only the three classes or all possible combinations when training GMMs to conform the KBM.

Experiments on the provided development data show that the primary system provides better performance than the best-performing system of the 2012 Albayzin evaluation (22.71% SER versus 26.34% SER in last evaluation).

**Acknowledgements.** This work is part of the project "Linguistic and sensorial accessibility: technologies for voiceover and audio description", funded by the Spanish Ministerio de Economía y Competitividad (FFI2012-31023). This article is supported by the Catalan Government Grant Agency Ref. 2014SGR027.

## References

1. Freesound.org, `https://www.freesound.org/`
2. Tecnoparla project, `http://tecnoparla.talp.cat/`
3. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, `http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf`
4. Anguera, X., Bonastre, J.F.: A novel speaker binary key derived from anchor models. In: INTERSPEECH. pp. 2118–2121 (2010)
5. Anguera, X., Bonastre, J.F.: Fast speaker diarization based on binary keys. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. pp. 4428–4431 (May 2011)
6. Anguera, X., Movellan, E., Ferrarons, M.: Emotions recognition using binary fingerprints. In: IberSPEECH (2012)
7. Delgado, H., Fredouille, C., Serrano, J.: Towards a complete binary key system for the speaker diarization task. In: INTERSPEECH (2014)

8. Hernández-Sierra, G., Bonastre, J.F., Calvo de Lara, J.: Speaker recognition using a binary representation and specificities models. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science, vol. 7441, pp. 732–739. Springer Berlin Heidelberg (2012)
9. Hu, G.: 100 non-speech environmental sounds, `http://www.cse.ohio-state.edu/dwang/pnl/corpus/HuCorpus.html/`
10. Ortega, A., Castan, D., Miguel, A., Lleida, E.: The Albayzin 2014 Audio Segmentation Evaluation. In: IberSPEECH (2014)

# Aholab Audio Segmentation System for Albayzin 2014 Evaluation Campaign

David Tavarez[1], Eva Navas[1], Agustin Alonso[1], Daniel Erro[1,2], Ibon Saratxaga[1], and Inma Hernaez[1]

[1]Aholab (UPV/EHU), ETSI Bilbao, Alda. Urquijo s/n, Bilbao, Spain
[2]IKERBASQUE, Alda. Urquijo, 36-5, Bilbao, Spain
{david,eva,agustin,derro,ibon,inma}@aholab.ehu.es

**Abstract.** *This paper describes the system developed by Aholab Signal Processing Laboratory for the Albayzin 2014 audio segmentation evaluation campaign. Two different strategies have been fused to build a robust system: a state-of-the-art HMM scheme with 8 separate HMMs which achieves a segmentation error of 21.99% in the development recordings, and a different approach based on GMM presegmentation and speech label refinement by means of i-vector classification with an error of 16.33%. A label level fusion algorithm has been applied with 18% improvement in the results for the music class, and 8% Total Class Error Time reduction.*

**Keywords:** Automatic Audio Segmentation, Albayzin Evaluation Campaigns, Broadcast Speech

## 1 Introduction

Automatic audio segmentation is the process of identifying the boundaries between different audio events and classifying each segment according to its content. The main objective of the automatic segmentation process changes according to the application: separating speech from noise and music [1], separating male voice from female voice [2], separating the segments corresponding to different speakers [3], etc. Automatic audio segmentation has many applications and is usually applied as a first pre-processing step to improve the performance of other systems like automatic speech recognition [4], speaker identification [5], content-based audio indexing and information retrieval [6], etc.

The Spanish Thematic Network for Speech Technologies (Red Temática en Tecnologías del Habla[1]) organizes each two years evaluation campaigns to establish a common framework for comparing and contrasting different systems developed for several speech technologies. In the last editions, one of these evaluations deals with automatic audio segmentation. This year's audio segmentation evaluation aims at improving the robustness of the participating systems against different acoustic conditions, by using different databases that have been merged. In this way the difficulty level has been increased compared to the last

[1] http://www.rthabla.es/

editions. This paper presents the system proposed by the Aholab Signal Processing Laboratory for this evaluation campaign.

The rest of the paper is organized as follows: section 2 describes the database used for training and testing the system. Section 3 presents the description of the system proposed by Aholab. The results obtained by the system are detailed in section 4. Finally, section 5 summarizes the conclusions of the work.

## 2 Database

The database provided by the organization for the audio segmentation evaluation campaign includes audio from three different sources:

- The Catalan broadcast news database [7] from the 3/24 TV channel used in the 2010 [8] and 2012 Albayzin Audio Segmentation Evaluations. It was recorded by the TALP Research Center from the UPC in 2009 and consists of around 87 hours of recordings with the following distribution: speech in the 92% of the segments, music in the 20% and noise in the 40%.
- The Aragón Radio database from the Corporación Aragonesa de Radio y Televisión used in the 2012 Albayzin Audio Segmentation Evaluation. It contains around 20 hours of recordings with the following class distribution: 22% clean speech, 9% music, 31% speech with music, 26% speech with noise and 12% other audio events.
- Environmental sounds from Freesound.org [10] and HuCorpus [11] among others to be merged with the audio from 3/24 TV and Aragon Radio databases.

As the result of the merging, 35 recordings have been provided by the Albayzin 2014 organization, divided into two sets: a subset of 15 recordings (around 15 hours) has been selected for testing and the rest (around 21 hours distributed in 20 recordings) has been supplied along with the corresponding reference labeling to train the systems.

In order to optimize different settings of the proposed segmentation system (see more details in section 3), we have further split the training dataset into two different subsets: The first 15 recordings have been used to train the models and the last 5 recordings to tune several parameters of the segmentation system. From now on, we will refer to these subsets as training and development datasets respectively.

Figure 1 shows the distribution of audio classes in the considered training dataset, i. e. in the first 15 recordings provided. According to it, classes containing speech are dominant in the database and represent the 91% of the total time.

## 3 Description of the Proposed System

Figure 2 shows a detailed diagram of the proposed solution. The final output of the proposed audio segmentation system is obtained by label level fusion of the outputs of two different subsystems:

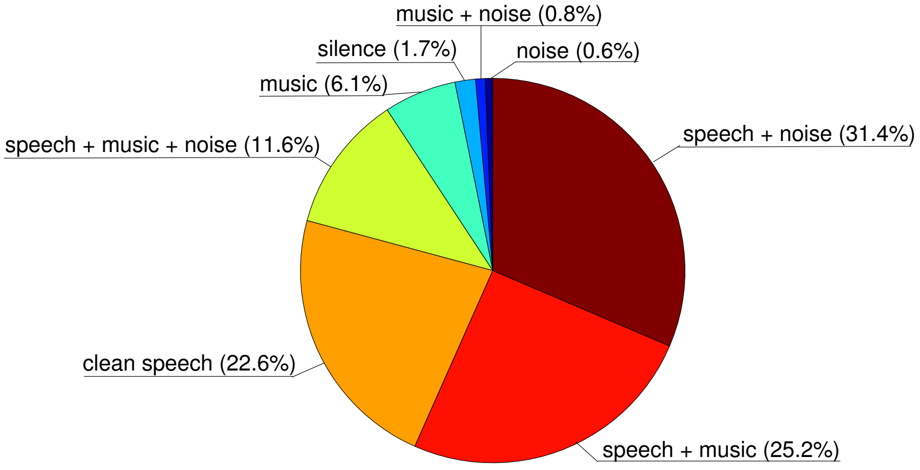


**Fig. 1.** Distribution of the different audio classes in the considered training set of Albayzin 2014 database

- A state-of-the-art hidden Markov model (HMM) scheme with 8 separate HMMs for each non overlapping class: silence, speech, music, noise, speech with music, speech with noise, music with noise and speech with music and with noise.
- A different approach based on GMM presegmentation and speech label refinement by means of i-vector classification via Multilayer Perceptron (MLP).

The outputs of both subsystems are post-processed to discard too short segments. Finally, a label level fusion algorithm is applied to combine the results of both subsystems and maximize the precision of the final labels. Next, each module of the system will be described in detail.

### 3.1 HMM segmentation

Using the training set of the database, a separate 3-state HMM with 512 mixtures has been trained for silence, clean speech, music, noise, speech with music, speech with noise, music with noise and speech with music and with noise. These eight models are used in a Viterbi segmentation to detect the boundaries of the audio segments containing the different acoustic events. 13 MFCC with first and second derivatives are used for the classification. Experiments made over the development dataset showed that the use of 512 mixtures provided the lowest Segmentation Error Rate (SER), as shown in Table 1. The HTK toolkit [12] has been used to train the models and to perform the audio segmentation.

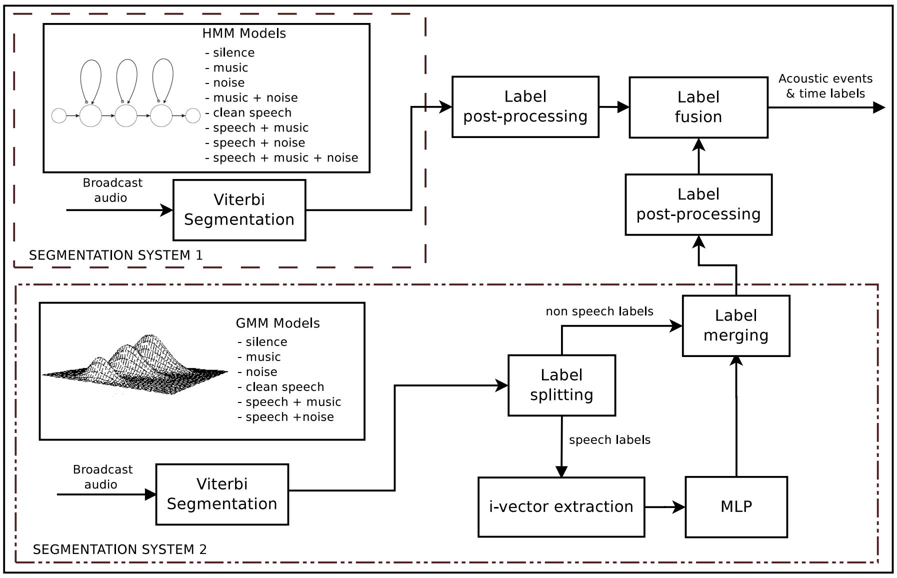


**Fig. 2.** Structure of the proposed Audio Segmentation System

| HMM mixtures | Segmentation Error Rate (SER) |
|---|---|
| 32 | 28.28% |
| 64 | 28.19% |
| 256 | 27.79% |
| 512 | **26.50%** |
| 1024 | 26.95% |

**Table 1.** Segmentation Error Rate of the HMM system for different number of Gaussians in the development dataset

### 3.2 I-vector segmentation

As classes including speech are dominant in the database, it is important to segment them accurately. In this system a first segmentation is made using GMM models and then the segments containing speech are relabeled by means of a system based on i-vectors.

First, 6 GMMs with 32 mixtures for silence, music, noise, clean speech, speech with noise, and speech with music are used in a Viterbi segmentation. The music with noise model is not taken into account in this system due to the limited presence of this event in the database and the increase of the SER when applying it to the segmentation process. The first five recordings of the training dataset were used to train the silence, noise and music models. The audio for the speech models was extracted just from the first recording as it has enough data to obtain robust models (see Figure 1 which shows the audio distribution of the dataset).

12 MFCC with first and second derivatives are used for the classification (energy related coefficient is not used in this case).

Once the speech segments are identified, the i-vector extraction process is carried out. A sliding window is used to extract the i-vectors corresponding to each speech segment. The window length is set to 5 seconds in order to obtain a reliable i-vector and 1 second is taken as frame rate as this is the resolution applied in the Albayzin 2014 evaluation for the label boundaries. Then, a MLP is used to classify each i-vector as clean speech, speech with noise, speech with music or speech with music and with noise. All the segments in the training dataset were used to train the MLP model.

To select the most appropriate dimension for the i-vectors, different MLP classifiers were built using the training dataset and 10 fold cross-validation. Table 2 shows the results, evaluated with different measures: percentage of correct decisions, unbalanced average precision (UAP), unbalanced average recall (UAR) and F score. All of them indicate that dimension 100 is the most suitable for the classification. Both MLP training and classification have been performed with WEKA software [13].

| Dimension | Correct | UAP | UAR | F-score |
|---|---|---|---|---|
| 150 | 79.91% | 0.7887 | 0.7901 | 0.7894 |
| 125 | 79.83% | 0.7886 | 0.7897 | 0.7891 |
| 100 | **80.61%** | **0.7985** | **0.8009** | **0.7997** |
| 75 | 80.27% | 0.7935 | 0.7949 | 0.7942 |
| 50 | 79.18% | 0.7816 | 0.7830 | 0.7823 |
| 25 | 74.23% | 0.7278 | 0.7328 | 0.7303 |

**Table 2.** MLP performance vs i-vector dimension

Finally, the speech labels from the i-vector classification along with the non-speech labels from the Viterbi segmentation are provided to the post-processing step to refine the boundaries of the audio segments.

### 3.3 Label post-processing

The aim of this step (equivalent in both subsystems) is twofold: on the one hand to discard short duration labels, corresponding to short silences, pauses between words and short acoustic event occurrences and on the other hand, to adapt them to the fusion step. In the first place, labels shorter than 950 ms are removed. Then, consecutive segments with the same label are unified. This joining process does not improve the results of the system but makes the fusion step faster by reducing the number of audio segments and so the number of decisions to be made.

### 3.4 Label fusion

The algorithm used for the fusion of the two systems is described in [14]. It operates at the label level and uses the confusion matrices of the systems involved in the fusion. The only information required to be applied are the labels assigned by each of the systems and the reference labels in the training/development part of the database. Experiments made over the development dataset showed that the use of the training dataset when computing the confusion matrices provided better SER, therefore, just the 15 recordings of the training dataset have been used to perform this task.

## 4 Results

This section presents the results obtained by the system in the training and development recordings of the Albayzin 2014 database. Table 3 shows the SER after the label post-processing step (both subsystems) and after the label fusion. All the results have been obtained with the evaluation script provided by the Albayzin 2014 organization.

| Track | HMM system | i-vec system | Label Fusion |
|---|---|---|---|
| 01 | 17.27% | 9.55% | 8.7% |
| 02 | 20.82% | 10.84% | 9.81% |
| 03 | 16.07% | 10.29% | 8.9% |
| 04 | 21.43% | 10.33% | 9.58% |
| 05 | 17.88% | 9.39% | 8.52% |
| 06 | 22.94% | 14.6% | 13.14% |
| 07 | 28.18% | 10.27% | 9.64% |
| 08 | 13.97% | 10.92% | 9.7% |
| 09 | 19.3% | 10.35% | 9.33% |
| 10 | 17.24% | 12.55% | 10.36% |
| 11 | 20.77% | 10.36% | 9.09% |
| 12 | 17.02% | 8.11% | 6.69% |
| 13 | 17.48% | 10.33% | 8.81% |
| 14 | 19.38% | 12.94% | 11.71% |
| 15 | 19.75% | 8.87% | 7.81% |
| ALL 1-15 | 19.38% | 10.67% | 9.49% |
| 16 | 18.79% | 15.72% | 14.55% |
| 17 | 14.17% | 12.87% | 11.23% |
| 18 | 25.74% | 16.35% | 14.98% |
| 19 | 19.68% | 15.66% | 14.85% |
| 20 | 33.09% | 21.73% | 20.55% |
| ALL 16-20 | 21.99% | 16.33% | 15.1% |

**Table 3.** Segmentation Error Rate of each step of the proposed system for the training and development sessions

As displayed in Table 3, the i-vector system performs better than the HMM based system with a 16.33% of SER in the development part of the database. It also shows less variability in the obtained values, in both training and development sets. This is possibly due to its higher class discrimination. By applying the label fusion step the segmentation error is reduced by a relative 8% in the development set, which clearly proves the validity of the proposed algorithm, even when the HMM system presents a considerably higher error rate than the i-vector system.

The following tables show the results obtained by the system for each of the target acoustic events individually (speech, music and noise) over the development set. The evaluation script provided by the Albayzin 2014 organization has been used to get these values. The error of the HMM and i-vector subsystems after the label post-processing are analyzed in the first place. Then, the error of system after the label fusion is presented.

Table 4 shows the results of the HMM system over the development set. The main source of the segmentation error resides on the noise detection (with 28.9% Missed Class Time and 41.2% False Alarm Class Time). The system also presents a considerably high error rate for the music class that comes mainly from the Missed Class Time contribution (23.2%) while the speech class labeling obtains good results (5.3% Total Class Error Time).

| Error | Speech | Music | Noise |
|---|---|---|---|
| Missed Class Time | 2.0% | 23.2% | 28.9% |
| False Alarm Class Time | 3.3% | 9.4% | 41.2% |
| Total Class Error Time | 5.3% | 32.6% | 70.11% |

**Table 4.** Results of the HMM system for speech, music and noise classes independently

The results of the i-vector system over the development set can be seen in Table 5. As it occurs in the HHM system, the main source of the segmentation error resides on the noise and music detection, although this system gets a considerably better performance compared to the HMM system in both cases (with 18.4% Total Class Error Time for music and 46.5% for noise). This music and noise error reduction is not consequence of a better “only music” and “only noise” classification, but of a better “speech with music”, “speech with noise” and “speech with music and with noise” discrimination performed by the MLP i-vector classifier. The speech labeling obtains slightly lower results compared to the HMM subsystem (7.6% Total Class Error Time) due to the Missed Class Time contribution, as a result of a less accurate speech and non-speech presegmentation performed by the GMM classifier.

Finally, Table 6 shows the final results of the proposed system (after the label fusion step) over the development set. The improvement obtained by the fusion step is mainly due to the more accurate labeling of the music class. The fusion step decreases considerably the music Total Class Error Time, reducing

| Error | Speech | Music | Noise |
|---|---|---|---|
| Missed Class Time | 4.4% | 11.9% | 18.1% |
| False Alarm Class Time | 3.2% | 6.5% | 28.4% |
| Total Class Error Time | 7.6% | 18.4% | 46.5% |

**Table 5.** Results of the i-vector system for speech, music and noise classes independently

by a relative 18% the segmentation error obtained by the i-vector system. On the other hand, it barely affects the speech and noise labeling, with similar values to those provided by the best subsystem in each case.

| Error | Speech | Music | Noise |
|---|---|---|---|
| Missed Class Time | 3.0% | 8.6% | 17.9% |
| False Alarm Class Time | 3.1% | 6.4% | 28.6% |
| Total Class Error Time | 6.1% | 15.0% | 46.5% |

**Table 6.** Results after label fusion for speech, music and noise classes independently

Table 7 shows the CPU time required in order to process all the recordings of the test dataset. The time required for the label fusion step is also shown. These measures were made on a octa-core Intel Xeon 2.27 GHz computer with 64 GB memory.

| Database | HMM system | i-vec system | Label Fusion |
|---|---|---|---|
| 15h 37m 50s | 1h 39m 17s | 8h 9m 17s | 5s |

**Table 7.** CPU time required in order to process the test part of the database

It is worth mentioning that the i-vector system implementation is far from being optimal. File management is required between the audio parametrization and the i-vector extraction steps which takes 80% of the required process time. Integration of these two steps would decrease considerably the CPU time shown in Table 7 for the i-vector subsystem.

## 5 Conclusions

This paper presents the audio segmentation system developed by Aholab Signal Processing Laboratory for the Albayzin 2014 Evaluation Campaign. The output of the proposed system has been obtained by fusion of the labels of two different subsystems: A state-of-the-art HMM scheme with 8 separate HMMs and a different approach based on GMM speech/non speech segmentation and speech label refinement by means of i-vector classification. These subsystems obtain a Total Class Error Time of 21.99% and 16.33% respectively. respectively in the development set. A label level fusion algorithm has been applied with an improvement of 18% in the results for the music class, which translates in an 8% reduction of Total Class Error Time.

A robust audio segmentation system has been built by taking advantage of the capabilities of the two subsystems: non-speech segmentation performed by the HMM system and the speech segments classification performed by the MLP i-vector classifier.

## 6 Acknowledgments

This work has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03) and the Basque Government (Ber2tek project, IE12-333).

## References

1. Lu, L., Zhang, H., Jiang, H.: Content analysis for audio classification and segmentation. IEEE Transactions on Speech and Audio Processing, 10(7), pp. 504 - 516. (2002)
2. Ore, B.M., Slyh, R.E., Hansen, E.G.: Speaker Segmentation and Clustering using Gender Information. In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, pp. 1 – 8. San Juan, Puerto Rico. (2006)
3. Moattar, M. H., Homayounpour, M. M.: A review on speaker diarization systems and approaches. Speech Communication, 54(10), pp. 1065 - 1103. (2012)
4. Rybach, D., Gollan, C.: Audio segmentation for speech recognition using segment features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 4197 - 4200. Taipei, Taiwan (2009)
5. Reynolds, D. A., Torres-Carrasquillo, P.: Approaches and applications of audio diarization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), pp. 953 - 956. Philadelphia, USA. (2005)
6. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP), vol. 2, pp. 5 - 8. Hong-Kong, China. (2003)
7. Butko, T., Nadeu, C.: Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. EURASIP Journal on Audio, Speech, and Music Processing, 2011(1), pp. 1 - 10. (2011)
8. Butko, T., Nadeu, C., Schulz, H.: Albayzin-2010 Audio Segmentation Evaluation Evaluation Setup and Results. In: FALA 2010, pp. 305 - 308, Vigo (Spain). (2010)
9. Ortega A., Castan D., Miguel A., Lleida E.: The Albayzin 2012 Audio Segmentation Evaluation. (2012) Available online: http://dihana.cps.unizar.es/~dcastan/wp-content/papercite-data/pdf/ortega2012.pdf
10. F. Font, G. Roma, X. Serra: Freesound technical demo. In: Proceedings of the 21st ACM international conference on Multimedia (MM'13), pp. 411-412. (2013) Available online: https://www.freesound.org/

11. G.Hu. 100 non-speech environmental sounds. Available online: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html
12. S.J. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P.C. Woodland The HTK Book (for HTK Version 3.4) University of Cambridge (2006).
13. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten: The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1. (2009)
14. D. Tavarez, E. Navas, D. Erro, I. Saratxaga, I. Hernaez: Nueva técnica de fusión de clasificadores aplicada a la mejora de la segmentación de audio. (in Spanish), Procesamiento del Lenguaje Natural, vol. 51, pp. 161-168. (2013)
15. Tavarez, D., Navas, E., Erro,D., Saratxaga,I.: Strategies to Improve a Speaker Diarisation Tool. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 4117-4121, Istambul (2012)

# The Albayzin 2014 Audio Segmentation Evaluation

Alfonso Ortega, Diego Castan, Antonio Miguel, Eduardo Lleida

Vivolab, Aragon Institute for Engineering Resarch (I3A)
University of Zaragoza, Spain
{ortega,dcastan,amiguel,lleida}@unizar.es
http://www.vivolab.es

**Abstract.** This document describes the 2014 Albayzin Audio Segmentation Evaluation that will be conducted as part of the Iberspeech 2014 conference. Audio Segmentation is a very important task for some speech technologies applications like Automatic Speech Recognition or Spoken Document Retrieval. This evaluation consists of segmenting and labeling broadcast audio documents to indicate which segments contain speech, music and/or noise. Unlike previous editions, this evaluation aims at providing an experimental framework for segmentation systems across different databases that can be merged or even overlapped increasing the difficulty from last editions. Therefore, the main goal is to test the robustness of the participating systems against different acoustic contexts. The Segmentation Error Rate will be used as scoring metric as Diarization Error Rate is used in the Diarization evaluations organized by NIST as part of the RT evaluations.

## 1 Introduction

In some applications of speech technologies like Automatic Speech Recognition systems for Broadcast shows or Spoken Document Retrieval in very large multimedia repositories, Audio Segmentation is considered a very important task. Speech is usually found along with music or environmental noise, and the presence of each one of these acoustic classes must be accurately labeled, since the accuracy of these labels is critical for the subsequent systems to be successful. Thus, the development of accurate Audio Segmentation Systems is essential to allow applications like ASR or SDR to perform adequately in real-world environments.

However, the audio documents are increasing day by day due to the audio-sharing websites or the audio on demand systems. Users around the world can upload and share their content and, therefore, the variability of the sound is very high. This evaluation aims at measuring the quality of the segmentation systems for different databases and different contexts.

## 2 Description of the Evaluation

The proposed evaluation consists of segmenting a broadcast audio document and assign labels for each segment indicating the presence of speech, music and/or noise. That is, two or more classes can be found simultaneously in audio segments and the goal is to indicate if one, two or the three aforementioned classes are present for a given time instant. For example, music can be overlapped with speech or noise can be in the background if someone is speaking. In this evaluation, we consider that Speech is present every time that a person is speaking but not in the background. Music is understood in a general sense and Noise is considered every time some acoustic content is present different than speech and music (including speech in the background).

### 2.1 Changes from the 2010 and 2012 Albayzin Audio Segmentation Evaluation

As in the 2010 and 2012 Albayzin Audio Segmentation Evaluation, the goal is segmenting and labeling audio documents indicating where speech, music and/or noise are present. Following the 2012 evaluation criteria, no prior classes are defined (*speech, music, speech with noise in background, speech with music in background, other*) and a multiple layer labeling is proposed. Nevertheless, the goal this year is to segment the incoming audio into the three possibly overlapped acoustic classes (Speech, Music and Noise) where the audio is composed of different databases that can be merged or even overlapped increasing the difficulty from last editions dramatically.

## 3 Database Description

The database for this evaluation is a combination and fusion of three databases defined below:

The first set is the Catalan broadcast news database from the 3/24 TV channel proposed for the 2010 Albayzin Audio Segmentation Evaluation [1, 2]. This database was recorded by the TALP Research Center from the UPC in 2009 under the Tecnoparla project [3] funded by the Generalitat the Catalunya. The Corporació Catalana de Mitjans Audiovisuals (CCMA), owner of the multimedia content, allows its use for technology research and development. The database consists of around 87 hours of recordings in which speech can be found in a 92% of the segments, music is present a 20% of the time and noise in the background a 40%. Another class called *others* was defined which can be found a 3% of the time. Regarding the overlapped classes, 40% of the time speech can be found along with noise and 15% of the time speech along with music.

The second set is the Aragón Radio database from the Corporación Aragonesa de Radio y Televisión (CARTV) which was used for the 2012 Albayzin Audio Segmentation Evaluation [4]. As the owner of the audio content, Aragón Radio and the Corporación Aragonesa de Radio y Televisión allow the use of these data for research purposes.

The last set will be composed of environmental sounds from Freesound.org [5] and HuCorpus [6] among others. This sounds will be merged with segments from the 3/24 TV and Aragón Radio databases.

All the data that will be supplied in PCM format, mono, little endian 16 bit resolution, and 16 kHz sampling frequency.

## 4 Segmentation Scoring

As in the NIST RT Diarization evaluations [7], to measure the performance of the proposed systems, the segmentation error score (SER) will be computed as the fraction of class time that is not correctly attributed to that specific class (speech, noise or music). This score will be computed over the entire file to be processed; including regions where more than one class is present (overlap regions).

This score will be defined as the ratio of the overall segmentation error time to the sum of the durations of the segments that are assigned to each class in the file.

Given the dataset to evaluate $\Omega$, each document is divided into contiguous segments at all class change points[1] and the segmentation error time for each segment $n$ is defined as

$$\Xi(n) = T(n) \left[\max\left(N_{ref}(n), N_{sys}(n)\right) - N_{Correct}(n)\right] \tag{1}$$

where $T(n)$ is the duration of segment $n$, $N_{ref}(n)$ is the number of reference classes that are present in segment $n$, $N_{sys}(n)$ is the number of system classes that are present in segment $n$ and $N_{Correct}(n)$ is the number of reference classes in segment $n$ correctly assigned by the segmentation system.

$$SER = \frac{\sum\limits_{n \in \Omega} \Xi(n)}{\sum\limits_{n \in \Omega} \left(T(n) N_{ref}(n)\right)} \tag{2}$$

The segmentation error time includes the time that is assigned to the wrong class, missed class time and false alarm class time:

- **Class Error Time**: The Class Error Time is the amount of time that has been assigned to an incorrect class. This error can occur in segments where the number of system classes is greater than the number of reference classes, but also in segments where the number of system classes is lower than the number of reference classes whenever the number of system classes and the number of reference classes are greater than zero.

[1] A "class change point" occurs each time any reference class or system class stars or ends. Thus, the set of active reference classes and/or system classes does not change during any segment

- **Missed Class Time**: The Missed Class Time refers to the amount of time that a class is present but not labeled by the segmentation system in segments where the number of system classes is lower than the number of reference classes.
- **False Alarm Class Time**: The False Alarm Class Time is the amount of time that a class has been labeled by the segmentation system but is not present in segments where the number of system classes is greater than the number of reference classes.

A forgiveness collar of one second, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a class begins or ends.

### 4.1 Segmentation Scoring Tool and Audio Segmentation Systems Output Files

The tool used for evaluating the segmentation system is the one developed for the RT Diarization evaluations by NIST "md-eval-v21.pl", available in the web site of the NIST RT evaluations and directly accessible by clicking `http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl`here.

The format's definition for the submission of the Audio Segmentation results has been fixed according to the operation of the NIST's tool. Specifically the Rich Transcription Time Marked (RTTM) format will be used for audio segmentation system output and reference files. RTTM files are space-separated text files that contain meta-data 'Objects' that annotate elements of each recording and a detailed description of the format can be found in Appendix A of the 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan [7]. Nevertheless, since in this evaluation, acoustic classes will be considered as if they were speakers in Diarization evaluations, the "SPEAKER" object will be used. Thus, the required information for each segment will be:
SPEAKER File Channel Beg_Time Dur <NA> <NA> Class_Name <NA> <NA>
Where:

- **SPEAKER**: Is a tag indicating that the segments contains information about the beginning, duration, identity, etc. of a segment that belongs to a certain speaker. In our case, instead of a speaker, an acoustic class will be considered (speech, music or noise).
- **File**: Is the name of the considered file.
- **Channel**: Refers to the channel. Since we are dealing with mono recordings this value will always be 1.
- **Beginning_Time**: The beginning time of the segment, in seconds, measured from the start time of the file.
- **Duration**: Indicates the duration of the segment, in seconds.
- **Class_Name**: Refers to the name of the class that is present in the considered segment (sp for speech, mu for music, no for noise).

The tag <NA> indicates that the rest of the fields are not used. The numerical representation must be in seconds and hundredth of a second. The decimal delimiter must be '.'.

As an example, let us consider that we are dealing with a recording contained in a file named session08.wav. Thus, the RTTM file name session08.rttm will contain the required information about that specific file. If the first three rows of the file are:

```
SPEAKER session08 1 2.67 17.91 <NA> <NA> mu <NA> <NA>
SPEAKER session08 1 11.98 13.30 <NA> <NA> sp <NA> <NA>
SPEAKER session08 1 25.28 76.20 <NA> <NA> no <NA> <NA>
```

This means that there is one segment containing music that starts at 2.67 sec. with a duration of 7.91 sec. Then in second 11.98 a speech segments starts with a duration of 13.30 sec. and finally, a noise segment starts at 25.28 sec. with a duration of 76.20 sec.

The Albayzin 2014 Audio Segmentation evaluation will use the md-eval version 21 software and the command line will be:

```
md-eval-v21-force-mapping.pl -c 1.0 -r <SPKR-REFERENCE>.rttm
-s <SYSTEM>.rttm
```

## 5 General Evaluation Conditions

The organizers encourage the participation of all researchers interested in audio segmentation. All teams willing to participate in this evaluation must send an e-mail to

- ortega@unizar.es
- dcastan@unizar.es

Indicating the following Information:

- RESEARCH GROUP:
- INSTITUTION:
- CONTACT PERSON:
- E-MAIL:

with CC to the Chairs of the IberSpeech 2014 Evaluations::

- iberspeech2014@ulpgc.es

before July 15, 2014.

All participant teams must submit at least a primary system but they can also submit up to two contrastive systems. Each and every submitted system must be applied to the whole test database. The ranking of the evaluation will be done according to results of the primary systems but the analysis of the results of the contrastive systems will be also processed and presented during the evaluation session at Iberspeech. All participant sites must agree to make their submissions (system output, system description, ...) available for experimental use by the rest of the participants or the organizing team.

Each participant team must provide also the total time required to run the set of tests for each submitted system (specifying the computational resources used). No manual intervention is allowed for each developed system to generate its output, thus, all developed systems must be fully automatic. Listening to the evaluation data, or any other human interaction with the evaluation data, is not allowed before all results have been submitted. The evaluated systems must use only audio signals. Any publicly available data can be used for training together with the data provided by the organization team to train the audio segmentation system. In case of using additional material, the participant will notify it and provide the references of this material.

### 5.1 Results Submission Guidelines

The evaluation results must be presented in just one RTTM file per submitted system. The file output file must be identified by the following code:

EXP-ID::=<SITE>_<SYSID> where,

- <**SITE**>: Refers to a three letter acronym identifying the participant team (UPM, UPC, UVI, ...)
- <**SYSID**>: Is an alphanumeric string identifying the submitted system. For the primary system the SYSID string must begin with p-, c1- for contrastive system 1 and c2- for contrastive system 2.

Each participant site must send an e-mail with the corresponding RTTM result files alogn with a technical description of the submitted systems to

- ortega@unizar.es
- dcastan@unizar.es

before September 30, 2014.

### 5.2 System Descriptions

Participants must send, along with the result files, a PDF file with the description of each submitted system. The format of the submitted documents must fulfil the requirements given in the IberSpeech 2014 call for papers. You can use the templates provided for the Iberspeech conference (WORD or LaTeX). Please, include in your descriptions all the essential information to allow readers to understand which are the key aspects of your systems.

### 5.3 Schedule

- June 23, 2014: Release of the training and development data.
- July 15, 2014: Registration deadline.
- September 3, 2014: Release of the evaluation data.
- September 30, 2014: Deadline for submission of results and system descriptions.
- October 15, 2014: Results distribute to the participants.
- Iberspeech 2014 workshop: Official public publication of the results.

## 6 Acknowledgments

The Albayzin 2014 Audio Segmentation Evaluation organizing team would like to thank the Corporación Aragonesa de Radio y Televisión and Aragón Radio for providing the data for the evaluation. Thanks also to Taras Butko and Climent Nadeu who organized the 2010 Albayzin Audio Segmentation Evaluation for their help, support and for providing the training material for this evaluation. And also to the organizing committee of Iberspeech 2014 for their help and support.

## References

[1] Butko T., Albayzin Evaluations 2010: Audio Segmentation. Online: http://fala2010.uvigo.es/images/stories/pdfs/albayzinproposalaudiosegmentation_v1.pdf, accessed on 15 May 2012.

[2] Zelenak M., Albayzin Evaluations 2010: Audio Segmentation. Online: http://fala2010.uvigo.es/images/stories/pdfs/speakerdiarizationevaluationplanfala2010_v2.pdf, , accessed on 15 May 2012.

[3] Tecnoparla Project. Online: http://www.talp.upc.edu/tecnoparla, , accessed on 15 May 2012.

[4] Ortega A., Castan D., Miguel A., Lleida E., The Albayzin 2012 Audio Segmentation Evaluation. Online: http://dihana.cps.unizar.es/ dcastan/wp-content/papercite-data/pdf/ortega2012.pdf

[5] Freesound.org. Online: https://www.freesound.org/

[6] G.Hu. 100 non-speech environmental sounds Online: http://www.cse.ohio-state.edu/ dwang/pnl/corpus/HuCorpus.html

[7] The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan. Online: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, accessed on 15 May 2012.

# The ATVS-GEINTRA STD System for ALBAYZIN 2014 Search-on-Speech Evaluation

Junchen Xu[1], Doroteo T. Toledano[1], and Javier Tejedor [2]

[1] ATVS-UAM, Escuela Politécnica Superior, Universidad Autónoma de Madrid,
Calle Francisco Tomás y Valiente, 11; 28049 Madrid, Spain
junchen.xu@estudiante.uam.es; doroteo.torre@uam.es

[2] GEINTRA, Universidad de Alcalá, Madrid, Spain
javier.tejedor@depeca.uah.es

**Abstract.** This paper describes the system developed in a joint effort by ATVS-UAM and GEINTRA-UAH for the ALBAYZIN 2014 Search-on-Speech Evaluation. Among the four different modalities of the evaluation, we have decided to participate only in the Spoken Term Detection (STD) Evaluation. Our system employs an Automatic Speech Recognition (ASR) subsystem to produce word lattices and a Spoken Term Detection (STD) subsystem to retrieve potential occurrences. Kaldi toolkit has been used both for building the ASR subsystem and the STD subsystem. The Fisher Spanish Corpus has been used for training the ASR subsystem. In order to adapt both the acoustic and the language models to the task, the development data provided by the organizers have been added to the Fisher Spanish corpus. Our best ASR result on Fisher Spanish corpus is about 50% Word Error Rate (WER), and about 75% WER on a small part of the development data provided by the organizers. Our best STD result on this part of the development data is an ATWV of 0.1863.

**Keywords:** Spoken Term Detection, Keyword Spotting, Search on Speech, Automatic Speech Recognition.

## 1 Introduction

The increasing volume of speech information stored in audio and video repositories motivates the development of automatic audio indexing and spoken document retrieval systems. Spoken Term Detection (STD), defined by NIST as 'searching vast, heterogeneous audio archives for occurrences of spoken terms' [9] is a fundamental block of those systems, and significant research has been conducted on this task [1, 5, 6, 10, 13, 14, 15, 16].

This paper presents the ATVS-GEINTRA STD system submitted to the ALBAYZIN 2014 Search-on-Speech Spoken Term Detection (STD) Evaluation. It is a collaborative work of the ATVS research group from Universidad Autónoma de Madrid and GEINTRA research group from Universidad de Alcalá. Most of the work

adfa, p. 1, 2011.

was conducted by a student (Junchen Xu) under the supervision of the other authors as part of his end of studies project of Telecommunications Engineering.

The submission involves an automatic speech recognition (ASR) subsystem, and an STD subsystem. The ASR subsystem converts input speech signals into word lattices, and the STD subsystem integrates a term detector which searches for putative occurrences of query terms, and a decision maker which decides whether detections are reliable enough to be considered as hits or should be rejected as false alarms.

The ASR subsystem is based on Gaussian mixture models (GMM) and was built using the Kaldi toolkit [12]. The training process largely followed the Switchboard s5 recipe, adapted to use the Fisher Spanish corpus [3] and the training/development materials provided by the organizers. The same tool was used to conduct decoding and produce word lattices.

In previous works [14, 15] we used a proprietary STD subsystem employing an n-gram reverse indexing approach [7] to achieve fast term search. This approach indexed word/phone n-grams retrieved from lattices, and term search was implemented as retrieving n-gram fragments of a query term. Then, the confidence score of a hypothesized detection was computed as the averaged lattice-based score of the n-grams of the detection.

For this evaluation, our goal was to compare our proprietary STD subsystem to the STD subsystem recently provided as part of the Kaldi toolkit [12]. However, due to insufficient time and resources during the evaluation, we were only able to produce scores/detections with the STD subsystem of Kaldi, leaving the interesting comparison for future, post-evaluation work.

We have finally submitted two systems, ATVS-GEINTRA_STD_pri and ATVS-GEINTRA_STD_con1. Both are almost the same system with the only difference that the primary system is designed to optimize the Actual Term Weighted Value (ATWV) while the second system is designed to optimize the Word Error Rate (WER) in the ASR subsystem.

The rest of the paper is organized as follows: Section 2 presents the details of our primary system, including the system description and the detailed description of the database used. Section 3 highlights the differences between the primary and the contrastive systems. Finally, Section 4 provides conclusions and future research directions.

## 2 Primary System: ATVS-GEINTRA_STD_pri

Our submission involves an ASR subsystem and an STD subsystem, both based on Kaldi. Figure 1 shows our system architecture. Training was conducted using the Fisher Spanish corpus [3] and the training/development data provided by the organizers. The primary system was designed to optimize the ATWV on the training/development data provided by the organizers.

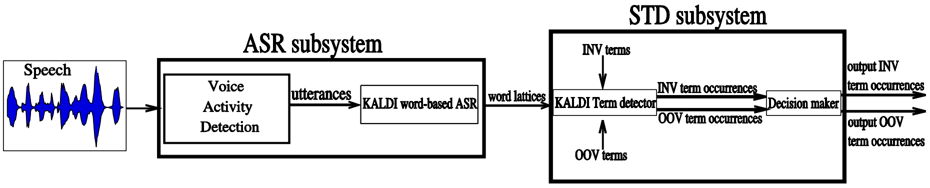


**Fig. 1.** STD system architecture.

## 2.1 System description

This section will describe the ASR and the STD subsystems in sequence. Instead of resorting to a hybrid approach using a word-based system to deal with in-vocabulary (INV) terms and a phone-based system to treat out-of-vocabulary (OOV) terms, as in previous works [14,15], we only used the method implemented in Kaldi to deal with OOV words. This method is based on proxy words and consists of substituting the OOV term to search by acoustically similar INV words (proxy words) and searching for these proxy words instead. This method allows dealing with OOV words without having to build two different ASR modules (word-based and subword-based) and correspondingly two different sets of lattices and indices. Details can be found in [12]. Our primary goal with this system was to compare it to our previous hybrid approach [14,15], but we did not have time to do that comparison by the evaluation deadline.

### Automatic Speech Recognition Subsystem

The Kaldi toolkit [12] was used to build the ASR subsystem, and we largely follow the Switchboard s5 recipe, except some minor changes in the configurations. Specifically, the acoustic features are 13-dimensional Mel-frequency cepstral coefficients (MFCCs), with cepstral mean and variance normalization (CMVN) applied to mitigate channel effects. We build two context-dependent phonetic acoustic models working directly on MFCCs, corresponding to two training iterations (we refer to these models as MFCC_1 and MFCC_2). The normalized MFCC features then pass a splicer which augments each frame by its left and right 4 neighboring frames. A linear discriminant analysis (LDA) is then employed to reduce the feature dimension to 40, and a maximum likelihood linear transform (MLLT) is applied to match the diagonal assumption in the GMM acoustic modeling. The model trained on these new features is denoted as +LDA+MLLT in the rest of the paper. After this model, the maximum likelihood linear regression (MLLR) and the speaker adaptive training (SAT) techniques are applied to improve model robustness. This model will be referred as +MLLR+SAT. Then a discriminative training approach based on the maximum mutual information (MMI) criterion is adopted to produce better models. Finally, the feature-space based maximum mutual information (fMMI) technique is applied to build the final models.

Based on the acoustic models, a word-based ASR system was built for searching INV terms. OOV terms were searched in the word lattices using the proxy words method implemented in Kaldi. The system uses a 3-gram word-based LM.

An energy-based voice activity detection (VAD) implemented in SoX is used to segment speech signals into utterances. Some heuristics of utterance and silence duration are employed to constraint the VAD process, where the heuristic parameters were optimized on a development set. The segmented utterances are then fed into the decoder implemented in Kaldi, which produces word lattices.

### Spoken Term Detection Subsystem

The Spoken Term Detection subsystem uses the keyword search tools provided by Kaldi. A brief description of the process, slightly modified from the one available in the Kaldi webpage is included here for completeness.

Lattices generated by the above ASR subsystem are processed using the lattice indexing technique described in [2]. The lattices of all the utterances in the search collection (speech data) are converted from individual weighted finite state transducers (WFST) to a single generalized factor transducer structure in which the start-time, end-time, and lattice posterior probability of each word token is stored as a 3-dimensional cost. This structure represents an inverted index of all word sequences seen in the lattices.

Given a query term, a simple finite state machine is created that accepts the term and composes with the factor transducer to obtain all occurrences of the term in the search collection, along with the utterance ID, start-time, end-time, and lattice posterior probability of each occurrence.

Finally, the decision maker simply sorts all these occurrences according to their posterior probabilities and a YES/NO decision is assigned to each occurrence.

OOV words are dealt with a method called proxy words, fully described in [4]. It essentially consists of substituting the OOV word to search with INV proxy words that are acoustically similar. The advantage of this method is that it does not require the use of a hybrid approach (word and sub-word models and lattices) as in our previous methods [14,15], being able to deal with OOV words using only a word ASR subsystem and a word-based lattice index. Our goal was to compare both methods in a different task but we could not perform that comparison for the evaluation and it remains as future work.

## 2.2 Train and development data

The evaluation task involves searching for some terms from speech data in the MAVIR corpus [8] that mainly contains speech in Spanish recorded during the MAVIR conferences. Since we did not have a large collection of comparable data, we decided to use a large database in Spanish to train the ASR module. We chose Fisher Spanish corpus [3] which amounts to 163 hours of conversational telephone speech (CTS) recordings (two sides) from 136 speakers. We used the same data for training the acoustic and language models. Since the data in the corpus (CTS) were very different from the data in the evaluation (mainly speech in conferences), we used the training and development data provided by the organizers, along with their transcriptions available in [8] to adapt the acoustic and language models.

The Fisher Spanish corpus was separated into mono recordings and divided into the three parts described in Table 1. The time in hours includes all the recorded silences.

**Table 1.** Partition of the Fisher Spanish corpus.

| Part | # Recordings | # Speakers | # Hours |
|---|---|---|---|
| Train | 1348 | 112 | ~268 |
| Dev | 146 | 12 | ~28 |
| Test | 144 | 12 | ~28 |

For training the ASR subsystem (acoustic and language models), we used the Train part and we evaluated the ASR subsystem using the Test part. The Dev part was used to tune parameters and to evaluate language model perplexity, which was 196.76.

When we started to process the corpus, we used our own rule-based grapheme-to-phoneme conversion module in Spanish to derive the phoneme transcriptions of the words in the lexicon. However, we soon realized that Fisher Spanish corpus had plenty of words in English (it is a Spanish corpus recorded mainly in the U.S.A.), so we had to perform a deeper analysis of the corpus and do *something* with the English words. Table 2 describes the types and amounts of words found in the Fisher Spanish corpus.

**Table 2.** Type and amount of words found in the Fisher Spanish corpus.

| Type | # words |
|---|---|
| Spanish words | 25400 |
| English words | over 3000 |
| Partial words | ~ 2200 |
| Interjections | 94 |
| Spanish Acronyms | 85 |
| English Acronyms | 92 |

For the English words, we decided to use the CMU Dictionary to obtain an English phoneme transcription and define translation rules from English to Spanish phonemes to build the phoneme transcription of the English words using Spanish phonemes. Interjections and acronyms were transcribed manually. In the end, we had a dictionary of about 30,000 terms, fully transcribed with a set of 24 Spanish phonemes.

Besides the phoneme models, we included models for the different types of noise present in the corpus. Table 3 summarizes these types of noise and their absolute frequency in the Fisher Spanish Corpus.

**Table 3.** Non-speech events present in the Fisher Spanish Corpus that were modeled with independent acoustic models.

| Type | # occurrences |
|---|---|
| <background> | 8480 |
| <laugh> | 10957 |
| <breath> | 4728 |
| <cough> | 618 |
| <sneeze> | 16 |
| <lipsmack> | 203 |

We also used the MAVIR data provided by the organizers for training/development. The organizers of the evaluation provided 7 recordings that amount at about 5 hours in total. From these 7 files, we decided to use only 5 for training. These 5 files have been added to the Fisher Spanish corpus to adapt a bit the acoustic and language models to the target speech type. The 2 remaining files, which amount at about 2 hours of audio, were reserved to conduct development spoken term detection experiments.

For the final systems submitted we added all the training/development data (7 files) to train the final acoustic and language models used to process the evaluation materials and produce the final results.

### 2.3 Optimization and results on development data

We conducted initial experiments using only the Fisher Spanish corpus to evaluate our ASR subsystem, and then used the MAVIR training/development data for STD experiments. Here, we report these development and optimization results.

Table 4 summarizes the ASR results obtained on the Test partition of Fisher Spanish corpus (see Table 1) in terms or Word Error Rate (WER) for the different training stages.

**Table 4.** WER obtained at the different training stages of the ASR subsystem on Fisher Spanish corpus.

| Training stage | WER (%) |
|---|---|
| MFCC_1 | 61.30 |
| MFCC_2 | 61.18 |
| +LDA + MLLT | 56.98 |
| +MLLR + SAT | 52.07 |
| +MMI | 50.27 |
| +fMMI | 49.88 |

Our ASR results are not still state-of-the-art. For instance, in [11] a WER of 36.5% is reported on Fisher Spanish corpus, although the partition used for test is different to ours.

After testing our ASR subsystem on Fisher data, we tested it on the 2 files from the MAVIR data provided as training/development that we reserved for STD experiments. Results are presented on Table 5. Initial results (with only Fisher data for training) were very poor. After adapting the language (based on a new dictionary with ~1000 additional words) and acoustic models, results improved, but still reached a WER of about 75%, significantly worse than for Fisher Spanish data. This was something expected due to the mismatch between Fisher Spanish and MAVIR data. We used the language and acoustic models adapted using MAVIR data for the rest of the evaluation. For the final results, we even used all the 7 training/development files to improve adaptation to MAVIR data.

**Table 5.** WER obtained at the different training stages of the ASR subsystem on MAVIR data for Initial models (trained only on Fisher Spanish data), with LM adapted to MAVIR data and with language and acoustic models adapted to MAVIR data.

| Training stage | WER (%)<br>Initial | WER (%)<br>LM adapted | WER (%)<br>LM & acoustic models adapted |
|---|---|---|---|
| MFCC_1 | 99.21 | 82.02 | 79.49 |
| MFCC_2 | 99.04 | 81.97 | 78.94 |
| +LDA + MLLT | 98.69 | 80.69 | 75.68 |
| +MLLR + SAT | 98.97 | 81.23 | 75.82 |
| +MMI | 98.65 | 80.61 | 74.59 |
| +fMMI | 99.09 | 83.15 | 76.69 |

We also conducted STD experiments on the 2 files reserved from the training/development data, and obtained the results presented in Table 6 in terms of Maximum Term Weighted Value (MTWV).

**Table 6.** MTWV obtained at the different training stages of the ASR subsystem on the 2 files of the MAVIR data reserved for STD experiments. As in Table 5, language and acoustic models adapted to MAVIR data used the remainding 5 MAVIR files.

| Training stage | MTWV |
|---|---|
| MFCC_1 | 0.0971 |
| MFCC_2 | 0.1160 |
| +LDA + MLLT | 0.1863 |
| +MLLR + SAT | 0.1603 |
| +MMI | 0.1845 |
| +fMMI | 0.1664 |

The system submitted as primary system was the one that optimized the MTWV, which is the LDA+MLLT system using language and acoustic models adapted to MAVIR data. YES/NO decision threshold was set to make ATWV reach the MTWV.

This system was further improved by including the 2 files from the MAVIR training/development data that we initially reserved for STD experiments in the training of the acoustic and language models. We did a final experiment on all the training/development data (7 MAVIR files) using this system. The results we obtained were MTWV=0.6287 and ATWV=0.6233. These results are much better than those obtained before because in this experiment a considerable amount of the material used for test (5 out of 7 files) was also used for training.

## 3 Contrastive System: ATVS-GEINTRA_STD_con1

The contrastive system we submitted is essentially the same, but optimized for WER instead of for MTWV. This is an MMI system optimized in threshold to make the ATWV meet the MTWV. As with the primary system, we improved the system by including in training the 2 files from the MAVIR training/development data initially reserved for STD experiments. The final experiment using all the 7 MAVIR training/development data achieved an MTWV=0.8327 and an ATWV=0.8155. Again, these results are highly unrealistic due to the re-use of data in training and test.

## 4 Conclusions and future work

This paper presents the ATVS-GEINTRA systems submitted to the ALBAYZIN 2014 Search on Speech Spoken Term Detection evaluation. Two systems were built. Both involve an ASR subsystem to produce word lattices and an STD subsystem for occurrence detection. Kaldi toolkit has been used to construct both subsystems. The systems were basically the same. The only difference relies on the ASR subsystem configuration chosen. One system employed a WER optimization-based tuning, while the other simply tuned the whole system towards the STD metric (ATWV). The best system achieved an ATWV of 0.1863 on a subset of the development data.

Future work will focus on the ASR subsystem, whose performance relates to that of the entire STD system in a large extent. For that, a Deep Neural Network-based ASR system should improve the final STD performance, as has been shown in ASR research during the last few years.

## 5 Acknowledgements

This work has been partly supported by project CMC-V2 (TEC2012-37585-C02-01) from the Spanish Ministry of Economy and Competitiveness.

## 6 References

1. Abad, A., Rodríguez-Fuentes, L.J., Peñagarikano, M., Varona, A., Bordel, G.: On the calibration and fusion of heterogeneous spoken term detection systems. Proc. of Interspeech. pp. 20-24 (2013).
2. Can, D., Saraclar, M.: Lattice Indexing for Spoken Term Detection. IEEE Trans. On Audio, Speech, and Language Processing, 19(8), pp. 2338-2347 (2011).
3. Fisher Spanish Corpus, Available at Linguistic Data Consortium Catalogue with reference LDC2010S01 (speech) and LDC2010T04 (transcripts), https://catalog.ldc.upenn.edu.
4. Guoguo C., Yilmaz, O., Trmal, J., Povey, D., Khudanpur, S.: Using proxies for OOV keywords in the keyword search task. Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 416-421, (2013).
5. Katsurada, K., Miura, S., Seng, K., Iribe, Y., Nitta, T.: Acceleration of spoken term detection using a subarray by assigning optimal threshold values to subkeywords. Proc. of Interspeech. pp. 11-14 (2013).
6. Li, H., Han, J., Zheng, T., Zheng, G.: A novel confidence measure based on context consistency for spoken term detection. Proc. of Interspeech. pp. 2429-2430 (2012).
7. Liu, C., Wang, D., Tejedor, J.: N-gram FST indexing for spoken term detection. Proc. of Interspeech. pp. 2093-2096 (2012).
8. MAVIR Corpus. Available at: http://www.lllf.uam.es/ESP/CorpusMavir.html.
9. NIST: The spoken term detection (STD) 2006 evaluation plan. National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, 10 edn. (September 2006), http://www.nist.gov/speech/tests/std.
10. Norouzian, A., Rose, R.: An approach for efficient open vocabulary spoken term detection. Speech Communication 57, 50-62 (2014).
11. Post, M., Kumar, G., López, A., Karakos, D., Callison-Burch, C., Khudanpur S.: Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus. Proc. of International Workshop on Spoken Language Translation, (2013).
12. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The KALDI speech recognition toolkit. Proc. of ASRU (2011).
13. Szoke, I.: Hybrid word-subword spoken term detection. Ph.D. thesis, Brno University of Technology (June 2010).
14. Tejedor, J., Toledano, D.T., Wang D., Colás, J.: Feature Analysis for Discriminative Confidence Estimation in Spoken Term Detection. Computer Speech and Language, 28(5), pp. 1083-1114 (2014).
15. Tejedor, J., Toledano, D.T., Wang D.: ATVS-CSLT-HCTLab System for NIST 2013 Open Keyword Search Evaluation. LNCS/LNAI Proceedings of IberSPEECH 2014 (to appear).
16. Wang, D.: Out-of-vocabulary Spoken Term Detection. Ph.D. thesis, University of Edinburgh (December 2009).

# GTH-UPM System for Search on Speech Evaluation

Julian Echeverry-Correa, Alejandro Coucheiro-Limeres, and Javier Ferreiros-López

Speech Technology Group, Universidad Politécnica de Madrid, Spain
{jdec,a.coucheiro,jfl}@die.upm.es

**Abstract.** This paper describes the GTH-UPM system for the Albayzin 2014 Search on Speech Evaluation. The evaluation task consists of searching a list of terms/queries in audio files. The GTH-UPM system we are presenting is based on a LVCSR (*Large Vocabulary Continuous Speech Recognition*) system. We have used MAVIR corpus and the Spanish partition of the EPPS (*European Parliament Plenary Sessions*) database for training both acoustic and language models. The main effort has been focused on lexicon preparation and text selection for the language model construction. The system makes use of different lexicon and language models depending on the task that is performed. For the best configuration of the system on the development set, we have obtained a FOM of 75.27 for the keyword spotting task.

**Keywords:** keyword spotting, spoken term detection, query by example, automatic speech recognition

## 1 Introduction

The search of information on speech has found many applications in the field of automatic speech recognition (ASR) in recent years. For applications such as dialog managers, conversational agents or spoken information retrieval systems, spotting significant keywords could be more important than recognizing the whole content of an utterance.

The tasks proposed in the 2014 Albayzin Search on Speech Evaluation entail several difficulties that must be taken into account in order to develop an optimal system. Besides the specific conditions and requirements of each task, there are some common features inside the MAVIR corpus (the one used for the evaluation) which demand to be studied at the early stages of the system design. These are mainly related to the acoustic conditions of the audio files. Due to the diversity of the recording conditions, the complexity of the task is increased and the robustness of the system must be optimized. On the one hand, the audios in the MAVIR corpus have several Spanish speakers, including both men and women. This means the acoustic models need to be trained covering this variety. And on the other hand, the quality of the audios changes, in terms of noise and different conditions between recording sessions. This characteristic is crucial in order to look for supplementary material corpus, whose audios should have similar acoustic conditions.

There are four tasks proposed in the Search on Speech Evaluation: Keyword Spotting (KWS), Spoken Term Detection (STD), Query-by-Example Spoken Term Detection (QbE STD) and Query-by-Example Spoken Document Retrieval (QbE SDR). They are very similar between them, since for all we have a list of terms (written or spoken) that we must search on the input speech, outputting the timestamps and a score of trust. We briefly describe their particularities. For KWS, the list of written terms (keywords) is known before processing the audios, so we can prevent the system to listen carefully for them. For STD, we pursue the same goal as for KWS except that the list of terms is known after processing the audios, so no prevention can be made. For QbE STD, we have the same conditions as for STD except that the search list is made by spoken terms and then an initial stage of recognition of these terms has to be made before searching. And for QbE SDR, as for QbE STD, the list is provided by spoken terms (with the possibility of more instances per term, all in Basque language), but now the output is a score of confidence for a spoken term appearing in a spoken document.

There are several approaches to each task in the state-of-art of Automatic Speech Recognition. As a first gross division, specially for the KWS and STD detection tasks, we can distinguish between systems based on Large Vocabulary Continuous Speech Recognition (LVCSR) and systems based only on keywords and non-keywords models. LVCSR systems allow a simple word-level search, but they need a complete training of the models in order to make possible the recognition of such a large vocabulary. Besides, they also need a proper language model for the correct connection between words in a continuous speech recognition system. Only terms in the vocabulary may be recognized, so any term out-of-vocabulary will never be recognized. For open vocabulary systems where no information about the set of keywords is provided while training the models, it may be required a large amount of training data in order to increase the probability of modeling the probable keywords [1]. And even using an extremely large corpus, we can never accurately model all possible strings of words. In this sense, the most common probabilistic approach for building language models in ASR applications is based on N-grams. This approach models the probability of finding ordered sequences of N words. Nevertheless, in order to face the data sparsity when modeling language, regarding to its variety and complexity, we can employ a smooth variation of N-gram, that is skip-grams. Skip-grams allow us to form new N-grams by skipping one or more words in a word sequence, so the context can be obtained widely around a word. This may overcome the data sparsity problem and may reduce the need of larger corpus. In [2], skip-grams are proven to outperform the standard N-grams for different test documents by using less amount of training data.

Within non-LVCRS systems we find variety depending on the purpose. For the KWS task, it is extended the use of systems based on filler models (also called garbage models). These systems make a phonetic decoding and look for the phonetic sequence that best fits the phonetic transcription of each keyword, making use of a confident measure based on word segments or on the proportion of correct phonemes. In order to minimize the number of false alarms, these systems do not only model the keywords, but also the non-keyword parts of speech. This background model is referred as filler model, and it is also based in phonetic models. One advantage of these systems over LVCSR systems is the higher speed due to their simplicity. Besides, phonetic-based systems do not depend on a large vocabulary like LVCSR, so the problem with out-of-vocabulary terms is avoided, and can be used as well for the STD [3] and QbESTD [4] tasks. Some systems, as in [5], are hybrid systems of LVCSR and phonetic engines. LVCSR is reserved for in-vocabulary terms due to its robustness and phonetic search and alignment is employed with out-of-vocabulary terms, so no query is uncovered by the system.

The systems mentioned above often make use of Hidden Markov Models. However, other approaches have been developed based on neural networks, on discriminative learning procedures, or on graphical models (GM). As an example, GM makes use of the graph theory in order to describe the time evolution of speech statistically. In [1], GM was used to perform a KWS task with a non-LVCSR system, with the particularity of being vocabulary independent and without require the training of an explicit filler model.

In the next section, we will describe the system submitted by our group for this Search on Speech Evaluation. We have attempted to perform tasks 1 to 2, with a LVCSR system as described below.

## 2 System description

As we previously said, the system developed for this evaluation consists of a LVCSR system. The feature vectors we used for the acoustic model training consisted of the first 13 PLP coefficients, as well as their first and second order time derivatives. The phoneme models were composed of three hidden states each. We used cross-word triphone models in order to account for contextual information and we consider up to 16 Gaussians per state during training.

We used the transcriptions of the training/development data set, which are available in the MAVIR web page [1], for training the models and for testing the performance of the ASR. These transcriptions are composed of 2878 sentences and a vocabulary size of 5309 words. We also used the transcriptions of the Spanish Parliament partition of the EPPS database (this database is described in section 2.1) to compose the training corpora for the language models. This database is composed of 16514 sentences and a 17.5k vocabulary.

To enrich the vocabulary and the robustness of the language models, we performed a manual data search based on the topics found in the training dataset of the MAVIR corpus. For instance, we searched for data related to *language technologies* and from the obtained results we selected texts on various topics, like

[1] http://cartago.lllf.uam.es/mavir/index.pl?m=videos

sentiment analysis, data crawling, etc. We also guided our data search through the websites of the companies that are mentioned in the audio files (for instance: daedalus, bitext, isoco, etc.). We collected nearly 2000 sentences, composed of a 7.2k vocabulary. These complementary data have been used in the training of language models.

As a first step for the recognition stage, we used a voice activity detector (VAD) to segment the speech signal and perform ASR on the segments of detected speech. The VAD that we used is included in the Voicebox toolbox [6]. We tuned the VAD for splitting the audio in segments with a length under 30 seconds.

For the KWS task and in order to boost the probability of keywords, we repeated twice the sentences in the LM training corpora that contained any keywords and we also repeated the keywords that were missing in the initial vocabulary from the training corpora. Also for this task we added to the initial vocabulary the pre-specified keyword terms so that there were no OOV keywords during ASR search. Multi-term keywords were added as separate words (each of these keywords is treated as a set of single words during recognition).

Regarding the implementation issues, the HTK Toolkit [7] was used for training acoustic models and for the ASR decoding stage. The SRILM Toolkit [8] was employed for creating the language models that the system uses. We use trigram models.

### 2.1 Databases description

We have used two databases:

- MAVIR corpus is a collection of audio and video recordings, with their corresponding orthographic transcriptions. The audio recordings come from lectures and talks held by the MAVIR consortium. The corpus is made up of 13 recordings in Spanish and English language (nevertheless for this evaluation, only the Spanish partition is available for training, development and evaluation purposes). Data were collected during the I, II, and III MAVIR Conference held in Madrid in 2006, 2007 and 2008 respectively. The details of this database are shown in Table 1. We used all the training audio files (except for MAVIR 2 and MAVIR 9) for training the acoustic models. We decided to remove MAVIR 2 and MAVIR 9 files because of the poor

**Table 1.** Details of the MAVIR database

| **Partition** | **Files** | **Length** |
|---|---|---|
| Training | MAVIR 2, 3, 6, 8, 9 and 12 | 4h56m |
| Development | MAVIR 7 | 0h21m |
| Evaluation | MAVIR 4, 11 and 13 | 2h0m |

acoustic conditions in which they were recorded. Nevertheless, we use the transcriptions of the all training files in order to train the language models.

– EPPS (*European Parliament Plenary Sessions*) is a database developed by the project TC-STAR (Technology and Corpora for Speech to Speech Translation) [9]. It consists of 61 hours of audio recordings with their corresponding orthographic transcriptions. These recordings were collected between 2004 and 2007. Most of the speakers are interpreters, nevertheless there are also native Spanish speakers. This database also includes 38 hours of audio recordings of the Spanish Parliament (PARL) collected between 2004 and 2006. All the speakers in this group are native Spanish speakers. We selected this database because its acoustic conditions can be similar to those encountered in MAVIR corpus. We use the audio files of both EPPS and PARL partitions to train acoustic models and we use the texts provided by the PARL partition to enrich the vocabulary of the system and the robustness of the language models.

## 3 Evaluation metrics

For the keyword spotting task, the Figure-of-Merit (FOM), as defined in [7], will be the primary metric for the evaluation. The FOM is defined as the detection rate averaged over the range of 0 to 10 false alarms per hour, and in its calculation it is assumed that the total duration of the test speech is $T$ hours. For each keyword, all of the spots must be ranked in score order. The percentage of true hits $p_i$ found before the $i'th$ false alarm is then calculated for $i = 1 \dots N + 1$ where $N$ is the first integer $\geq 10T - 0.5$. The FOM is defined as

$$FOM = \frac{1}{10T} \left( p_1 + p_2 + \dots + p_N + ap_{N+1} \right)$$

where $a = 10T - N$ interpolates to 10 false alarms per hour. Table 2 show the results obtained in the development set for the keyword spotting task. The results are presented in terms of the `Hits`, `FA` (false alarms) and `FOM`.

## 4 Final results

For the final evaluation we present the results obtained on the development and training sets. These results are shown in Tables 2 and 3. The only difference between the principal (PRI) and the contrastive system (CON1) is that the contrastive system employs a language model trained by using the transcriptions of the training dataset of the MAVIR corpus combined with data from the PARL partition of the EPPS database, and in contrast, the principal system does not use the resources from the PARL partition. This will allow a wider coverage for the keyword spotting in the CON1 system but also may introduce a higher

**Table 2.** Final results with the development set (mavir 07)

| System | **Task** | Hits | FA | Act. | FOM |
|---|---|---|---|---|---|
| PRI | KWS | 241 | 73 | 296 | 75.27 |
| PRI | STD | 227 | 38 | 296 | 72.78 |
| CON1 | KWS | 231 | 52 | 296 | 72.45 |

number of false alarms to the system. Next, we show the results obtained in the training set of the database. We are aware that these results do not reflect the performance of the system, since they are obtained over the same dataset for which the system was trained. Nevertheless, these results may offer an oracle approximation of the performance of the system.

**Table 3.** Results with the training set

| File | System | **Task** | Hits | FA | Act. | FOM |
|---|---|---|---|---|---|---|
| mavir 02 | PRI | KWS | 599 | 281 | 1016 | 55.31 |
| mavir 03 | PRI | KWS | 596 | 52 | 653 | 87.71 |
| mavir 06 | PRI | KWS | 427 | 20 | 446 | 94.12 |
| mavir 08 | PRI | KWS | 197 | 10 | 200 | 93.66 |
| mavir 09 | PRI | KWS | 106 | 186 | 910 | 11.26 |
| mavir 12 | PRI | KWS | 637 | 41 | 671 | 92.70 |
| mavir 02 | PRI | STD | 426 | 203 | 1016 | 40.06 |
| mavir 03 | PRI | STD | 551 | 38 | 653 | 82.47 |
| mavir 06 | PRI | STD | 412 | 19 | 446 | 91.36 |
| mavir 08 | PRI | STD | 189 | 10 | 200 | 89.99 |
| mavir 09 | PRI | STD | 39 | 94 | 910 | 4.16 |
| mavir 12 | PRI | STD | 609 | 38 | 671 | 88.65 |
| mavir 02 | CON1 | KWS | 577 | 227 | 1016 | 54.22 |
| mavir 03 | CON1 | KWS | 598 | 38 | 653 | 89.05 |
| mavir 06 | CON1 | KWS | 433 | 20 | 446 | 95.88 |
| mavir 08 | CON1 | KWS | 196 | 9 | 200 | 93.65 |
| mavir 09 | CON1 | KWS | 94 | 147 | 910 | 10.08 |
| mavir 12 | CON1 | KWS | 636 | 34 | 671 | 92.78 |

## 5 Conclusions

In this paper we have presented the description of the system submitted for the 2014 Albayzin Search on Speech Evaluation. The proposed system is based on a LVCSR system. We have used not only MAVIR corpus but also EPPS database to train both acoustic and language models. From the experiments conducted on the development dataset we can conclude that including complementary texts for the training of the language models may improve the keyword spotting but may also introduce a higher number of false alarms.

From this evaluation it is clear that for developing a proper system for a concrete task it is necessary to study the corpus under study so we can collect the adequate training data that best fits that corpus. This applies not only for the acoustic conditions and variety of speakers but also for the topics discussed in the audio recordings.

**Acknowledgements.** The work leading to these results has received funding from TIMPANO (TIN2011-28169-C05-03), and MA2VICMR (Comunidad Autonóma de Madrid, S2009/TIC-1542) projects. It has also been supported by the European Union under grant agreement number 287678. Julian Echeverry-Correa also acknowledges the support from Colciencias and from the Universidad Tecnológica de Pereira, Colombia.

## References

[1] Wöllmer, M. et al. "Robust vocabulary independent keyword spotting with graphical models". Automatic Speech Recognition & Understanding (ASRU 2009. IEEE Workshop on, 2009, 349-353), 2009.

[2] Guthrie, D. et al. "A closer look at skip-gram modelling". In Proc. of the 5th international Conference on Language Resources and Evaluation (LREC-2006), 1-4, 2006.

[3] Wallace, R. et al. "A phonetic search approach to the 2006 NIST spoken term detection evaluation". International Speech Communication Association (ISCA), 2007

[4] Shen, W. et al. "A comparison of query-by-example methods for spoken term detection". DTIC Document, 2009

[5] Mamou, J. et al. "Vocabulary independent spoken term detection". In Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 615-622, 2007.

[6] Brookes, M. "Voicebox: A Speech Processing Toolbox for Matlab". Department of Electrical and Electronic Engineering, Imperial College, London.

[7] Young, S. et al. "The HTK Book". Engineering Department of Cambridge University, 2006.

[8] Stolcke, A. "SRILM-An extensible language modeling toolkit". In 3rd International Conference on Speech and Language Technology (INTERSPEECH02), 2002.

[9] Mostefa, D. and Hamon, O. and Moreau, N. and Choukri, K. "Evaluation Report for the Technology and Corpora for Speech to Speech Translation (TC-STAR Project). Deliverable N. 30"

# GTM-UVigo systems for Albayzin 2014 Search on Speech Evaluation

Marta Martinez, Paula Lopez-Otero, Rocio Varela, Antonio Cardenal-Lopez,
Laura Docio-Fernandez, Carmen Garcia-Mateo

Multimedia Technologies Group (GTM), AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario de Vigo S/N
36310, Vigo, Spain
{mmartinez,plopez,rvarela,cardenal,ldocio,carmen}@gts.uvigo.es

**Abstract.** This paper describes the systems developed by the GTM-UVigo team for the Albayzin 2014 Search on Speech evaluation. The primary system for the spoken term detection task consisted on the fusion of two different large vocabulary continuous speech recognition systems that differed in almost all their components: front-end, acoustic modelling, decoder and keyword search approach. An isolate word recognition system was fused with the two aforementioned speech recognition systems for the keyword spotting task. For the query by example spoken term detection task, a fusion of three systems was presented: one of them followed one of the aforementioned continuous speech recognition approaches, with the difference that in this case it was necessary to obtain a transcription of the queries; the other two systems performed a dynamic time warping search, being the use of fingerprints as feature vectors the main novelty of the presented approach.

**Keywords:** Keyword spotting, spoken term detection, query by example spoken term detection

## 1 Introduction

In this paper, the systems developed by the GTM-UVigo team for the Albayzin 2014 Search on Speech evaluation are described. Specifically, systems for the tasks keyword spotting (KWS), spoken term detection (STD) and query by example spoken term detection (QbESTD) are presented.

In the KWS task, a fusion of three systems was submitted: two of them rely on large vocabulary continuous speech recognition (LVCSR) systems, while the other one is an isolate word recognition system. One of these LVCSR systems was built using the Kaldi tools [11] to train a set of acoustic models, to generate the output lattices and to perform lattice indexing and keyword search [5]. The second system used the UVigo LVCSR [9] to extract a list of N-Best hypothesis, which were converted to word meshes using the SRILM tool [15]. The first of the aforementioned systems was also submitted as a contrastive system. The other contrastive system submitted consisted on the fusion of the two LVCSR systems.

The systems presented for the STD task were equal to the LVCSR systems submitted for the KWS task, being the only difference the language model: in KWS, the keyword terms were included in the language model, while they were not included in the case of the STD task, as the terms are supposed to be unknown beforehand.

For the QbESTD task, the proposed primary system consisted on a fusion of three different QbESTD systems. The first system was the lattice-search system used in the KWS and STD tasks, with the difference that, in this case, a transcription of the queries had to be performed. The other two systems were based in dynamic time warping search, and they differed in the feature representation of the audio documents and queries: in one of them, a fingerprinting approach was used to obtain a binary representation of the audio [8], and in the other one the audio was represented by means of phoneme posteriorgrams obtained using an English phoneme recognizer based on long temporal context [14].

The rest of this paper is organized as follows: Sections 2, 3 and 4 describe the systems for the KWS, STD and QbESTD tasks, respectively; Section 5 presents the preliminary results obtained for the different tasks on the development data; and Section 6 presents some conclusions extracted from the experimental validation of the different systems.

## 2 Systems for keyword spotting

The primary system presented for the KWS task consisted on the fusion of three different systems: two different large vocabulary continuous speech recognition (LVCSR) based systems, which are described below, and an isolate word recognition system that includes all the search words in its grammar. One of the LVCSR systems was built using Kaldi while the other one was based on the UVigo LVCSR system. The Kaldi-based system was submitted as a contrastive system. A second contrastive system was submitted, which consisted on the fusion of the two LVCSR systems. The fusion strategy used in this system is also described in this Section.

### 2.1 Kaldi-based LVCSR System Description

A large vocabulary continuous speech recognition (LVCSR) system was built using the Kaldi open-source toolkit [11]. This system uses standard perceptual linear prediction (PLP) analysis to extract 13 dimensional acoustic features, and follows a state-of-the-art maximum likelihood (ML) acoustic training recipe, which begins with a flat-start initialization of context-independent phonetic Hidden Markov Models (HMMs), and ends with a speaker adaptive training (SAT) of state-clustered triphone HMMs with Gaussian mixture model (GMM) output densities. The ML stage is followed by the training of a Universal background model (UBM) from speaker-transformed training data, which is then used to train a subspace GMM (SGMM) that will be used in the decoding stage.

The Kaldi LVCSR decoder generates word lattices [12] using the above SGMM models. These lattices are processed using the lattice indexing technique described in [5] so that the lattices of all the utterances in the search collection are converted from individual weighted finite state transducers (WFST) to a single generalized factor transducer structure in which the start-time, end-time and lattice posterior probability of each word token is stored as a 3-dimensional cost. This factor transducer is actually an inverted index of all word sequences seen in the lattices. Thus, given a list of keywords or phrases, we then create a simple finite state machine that accepts the keywords/phrases and composes it with the factor transducer to obtain all occurrences of the keywords/phrases in the search collection.

The data used to train the acoustic models of this Kaldi-based LVCSR system was extracted from the Spanish material used in the 2006 TC-STAR automatic speech recognition evaluation campaign [1]. Specifically, the training data from the European Parliamentary plenary sessions and the Spanish Parliament sessions, which was manually transcribed, was used for this purpose [7]. It must be noted that all the non-speech parts as well as the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences and short speech utterances were discarded, so in the end the training material consisted of 2 hours and 36 minutes.

The language model (LM) was trained using a text database of 160 MWords composed of material from several sources (transcriptions of European and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, online courses and the transcriptions of the Mavir sessions included in the development set[2] [13]. For the development dataset, a different LM was created for each Mavir session, using the transcription of the session to obtain the optimum mixture of the partial LMs. For the evaluation set, the LM was generated using a normalized average of the weights obtained with the development sessions. In this task, the keywords were added to the language model. Note that the vocabulary was selected at the last stage of the LM training, once the partial LMs and its weights were computed. We used a trigram-based LM with a vocabulary of 60K words and a Kesser-Ney discount strategy.

### 2.2 UVigo-based LVCSR System Description

In this system, we used the UVigo LVCSR described in [9] and the lattice tool provided by the SRILM toolkit [15]. The recognition was performed in three stages. First, an energy-based voice activity detector was used to segment the audio stream in manageable chunks. In the second stage, an acoustic model was selected for each segment. Finally, the UVigo decoder was applied to extract the N-Best hypothesis.

The employed LMs were the ones described in previous Section. For the acoustic modelling we used two state demiphones, with 12 Mel frequency cepstrum coefficients (MFCCs) plus energy and their delta and acceleration values.

[1] http://www.tc-star.org

[2] http://cartago.lllf.uam.es/mavir/index.pl?m=descargas

We used acoustic models trained with the TC-STAR training database described in section 2.1, but a set of new additional models was adapted using the Mavir database material. The acoustic model selection was performed using a simple phonetic recognizer, selecting the model which provided the best acoustic scores.

Finally, the N-Best lists were post-processed using the SRI-LM toolkit [15] and converted to word meshes with posterior probabilities. The posterior probabilities were used as a confidence measure for the detected keyword.

### 2.3 Isolated word recognition system

This system consists on a decoder with a simple grammar composed of a set of N keywords interconnected in parallel. A free phoneme network is employed to obtain a hypothesis test. The acoustic modelling was the same as described in Section 2.2.

### 2.4 Fusion

Discriminative calibration and fusion was applied in order to combine the outputs of different KWS systems, aiming at taking advantage of the strengths of the individual KWS approaches [2]. First, a global minimum zero-mean and unit-variance normalization was applied, in order to prevent the scores of the individual systems to be in different ranges and also to obtain keyword-independent scores. The global minimum score produced by the system for all queries was used to hypothesize the missing scores. After normalization, calibration and fusion parameters were estimated by logistic regression on a development dataset in order to obtain improved discriminative and well-calibrated scores [3]. This calibration and fusion training was performed using the Bosaris toolkit [4].

## 3 Systems for spoken term detection

In this task we applied the LVCSR systems used for keyword spotting, which were described in Sections 2.1 and 2.2, i.e. the Kaldi-based and the UVigo-based LVCSR systems. Both systems were combined using the same techniques described in Section 2.4. The only differences between the strategy used in this task and in the keyword spotting task were that, for obvious reasons, the search terms were not included in the vocabulary nor in the LM, and the isolated word recognizer was not used here. Note that, apart from this fact, the LM training was the same, including the partial LM weights employed to compose the final model.

### 3.1 System fusion

In this task the fusion strategy described above was used. The difference was that, before applying the fusion step, the optimal operating point calculated in the development set was applied to each of the individual systems.

## 4 Systems for query by example spoken term detection

The primary system for the QbESTD consists on a fusion of three systems:

- MFCC-Fingerprint: a dynamic time warping (DTW) based system that uses audio fingerprints as feature vectors.
- Phoneme posteriorgrams: a DTW based system that uses phoneme posteriorgrams as feature vectors.
- Kaldi-LVCSR: the Kaldi-based LVCSR system described in Section 2.1 was used for QbESTD. To do so, first a transcription of the query was obtained, and then the aforementioned strategy was employed to find all the occurrences of the query.

A contrastive system was also presented, which consisted on the fusion of the MFCC-Fingerprint and the Kaldi-LVCSR systems.

A detailed description of the DTW systems mentioned above is presented in the rest of this Section, as well as a brief description of the fusion strategy.

### 4.1 Dynamic time warping systems

Two language-independent systems were developed for this task, which perform search on the audio by means of DTW. The search procedure is the same for both of them, but they differ on the feature vectors that are used. We developed an approach inspired by [1], which follows three main steps:

- Feature extraction. Acoustic features are extracted from the waveforms, both from the audio and from the queries.
  - MFCC-fingerprint. We used a fingerprint representation of the acoustic features, motivated by the fact that this representation removes the perceptual irrelevancies of the audio: we are not looking of exact matches, so the information about the speaker or the channel is negligible [8]. The fingerprints corresponding to the acoustic features of each frame were obtained as described in [10]. A convolution mask was used to binarize the acoustic features, specifically a mask for finding negative slopes on the spectrogram in two consecutive frames was applied. Given a set of acoustic features $S \in \Re^{I \times J}$ where $S_{i,j}$ is the feature corresponding to energy band $i$ and frame $j$, the value $F_{i,j}$ of the frame-level fingerprint corresponding to frame $j$ obtained after applying the convolution mask is

$$F_{i,j} = \begin{cases} 1 \text{ if } S_{i,j} - S_{i,j+1} + S_{i-1,j} - S_{i-1,j+1} > 0 \\ 0 \text{ if } S_{i,j} - S_{i,j+1} + S_{i-1,j} - S_{i-1,j+1} \leq 0 \end{cases} \quad (1)$$

After running several tests with different features, we finally chose to use MFCCs with delta, acceleration and $C_0$ coefficients.

  - Phoneme Posteriorgrams. Phoneme posteriorgrams [6] were extracted using a phoneme recognizer based on long temporal context [14] developed at the Brno University of Technology; specifically, the English system of the ones provided by them was used, as it was the one that achieved the best performance on the development data.
- Coarse search. We first perform a coarse search for candidate matches for each query and audio file by following the approach described in [1]. The Euclidean distance matrix between all the vectors of the query and the match audio is computed and the minimum distance per audio vector is selected. Then, the average of these minima in a window of the size of the query is used as an approximation of DTW. We also used a sliding window with 50% overlap.
- Fine search. After selecting those candidates that obtained the smallest distances (the number of candidates is the length of the audio divided by 100, with a minimum of 100 candidates), DTW is computed for all of them. Those candidates whose DTW distance is less than a threshold are confirmed and considered as matches, while the rest of them are discarded.

### 4.2 System fusion

The strategy used to fuse the different QbESTD systems was the one described in Section 2.4 for KWS but, in this case, a per-query zero-mean and unit-variance normalization (q-norm) was applied. In this task, the discriminative calibration and fusion was trained using all the training and development data.

## 5 Preliminary results

Table 1 shows the performance obtained with the individual and the fused systems on the development dataset for the KWS task, measured in terms of the Figure of Merit (FOM). As mentioned in Section 2, the primary system consists on the fusion of the Kaldi-LVCSR, UVigo-LVCSR and UVigo-IWR systems, while the constrastive2 system consists on the fusion of the Kaldi-LVCSR and the UVigo-LVCSR systems. This preliminary results show that the best performance was achieved by the Kaldi-LVCSR system (also submitted as contrastive system) but, as the difference between this system and the fusion of the three systems is negligible, we decided to submit the fusion as the primary system, because we rely that the combination of different systems will result in a better performance on the evaluation dataset.

Table 2 shows the performance obtained with the individual and the fused systems on the development dataset for the STD task in terms of the actual term weighted value (ATWV), the false alarm probability $P_{\mathrm{fa}}$ and the miss probability $P_{\mathrm{miss}}$. In this case, the difference in performance between the Kaldi-LVCSR and the primary system is more noticeable than in the KWS task, but we decided to keep on with the same criterion, so we submitted the fusion as the primary system.

**Table 1.** KWS systems: results on the development data

| System | FOM |
|---|---|
| Kaldi-LVCSR (Constrastive1) | 84.08% |
| UVigo-LVCSR | 46.75% |
| UVigo-IWR | 44.70% |
| Primary | 83.95% |
| Contrastive2 | 83.65% |

**Table 2.** STD systems: results on the development data

| System | ATWV | $P_{fa}$ | $P_{miss}$ |
|---|---|---|---|
| Kaldi-LVCSR (Constrastive1) | 0.581 | 0.00008 | 0.341 |
| UVigo-LVCSR | 0.215 | 0.00017 | 0.620 |
| Primary | 0.568 | 0.00007 | 0.363 |

Table 3 shows the performance obtained with the individual and the fused systems on the development dataset for the QbESTD task, in terms of ATWV, $P_{fa}$ and $P_{miss}$. It can be seen that the fusion of different systems clearly enhanced their individual performance, achieving an ATWV of 0.3026 when fusing the three proposed systems.

**Table 3.** QbESTD systems: results on the development data

| System | ATWV | $P_{fa}$ | $P_{miss}$ |
|---|---|---|---|
| MFCC-Fingerprint | 0.1787 | 0.00002 | 0.801 |
| Phoneme posteriorgrams | 0.1580 | 0.00001 | 0.834 |
| Kaldi-LVCSR | 0.1819 | 0.00006 | 0.758 |
| Primary | 0.3026 | 0.00009 | 0.607 |
| Contrastive1 | 0.2995 | 0.00009 | 0.611 |

## 6 Conclusions and future work

This paper presented different systems used to perform keyword spotting, spoken term detection and query by example spoken term detection in the framework of Albayzin 2014 Search on Speech evaluation. The preliminary results obtained for the two first task on the development data are encouraging, as a good performance was achieved in spite of the quality of some recordings where background noise is present, there are different speakers per recording and there is a big amount of pronunciation errors, which makes this scenario challenging for speech recognition based approaches. In future work, a strategy to deal with the out-of-vocabulary issue will be incorporated to the continuous speech recognition systems, as in the spoken term detection task, the out-of-vocabulary terms are completely ignored. Specifically, we intend to implement a strategy that, whenever an out-of-vocabulary word appears, similar words are used as search terms: if these similar words are spotted in the audio document we will consider that our out-of-vocabulary word is present in this document.

With respect to the query by example spoken term detection task, the presented preliminary results outperformed those obtained in the Albayzin 2012 Search on Speech evaluation. The novelty presented in this task consisted on the use of audio fingerprints as feature vectors, motivated by the idea that this representation removes perceptual irrelevancies from the audio; further experiments will be run in order to ensure the validity of this representation for query by example spoken term detection.

## Acknowledgements

This work has been supported by the European Regional Development Fund, the Galician Regional Government (CN2011/019, 'Consolidation of Research Units: AtlantTIC Project' CN2012/160), the Spanish Government (FPI grant BES-2010-033358 and 'SpeechTech4All Project' TEC2012-38939-C03-01) and 'TecAnDAli Research Network' of the Consellería de Educación e Ordenación Universitaria, Xunta de Galicia.

## References

1. Abad, A., Astudillo, R.F., Trancoso, I.: The L2F Spoken Web Search system for Mediaeval 2013. In: MediaEval'13 (2013)
2. Abad, A., Rodríguez-Fuentes, L.J., Peagarikano, M., Varona, A., Bordel, G.: On the calibration and fusion of heterogeneous spoken term detection systems. In: Proceedings of Interspeech. pp. 20–24 (2013)
3. Brümmer, N., van Leeuwen, D.: On calibration of language recognition scores. In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop. pp. 1–8 (2006)
4. Brümmer, N., de Villiers, E.: The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing. Tech. rep. (2011), `https://sites.google.com/site/nikobrummer`
5. Can, D., Saraclar, M.: Lattice indexing for spoken term detection. IEEE Transactions on Audio, Speech and Language Processing 19(8), 2338–2347 (2011)
6. Hazen, T.J., Shen, W., White, C.M.: Query-by-example spoken term detection using phonetic posteriorgram templates. In: 2009 IEEE Workshop on Automatic

Speech Recognition & Understanding, ASRU 2009, Merano/Meran, Italy, December 13-17, 2009. pp. 421–426 (2009)

7. Laura Docio-Fernandez, A.C.L., Garcia-Mateo, C.: Tc-star 2006 automatic speech recognition evaluation: The uvigo system. In: TC-STAR Workshop on Speech-to-Speech Translation (2006)
8. Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C.: Introducing a framework for the evaluation of music detection tools. In: 9th Language Resources and Evaluation Conference. pp. 568–572 (2014)
9. Martinez, M., Cardenal, A.: Experiments on keyword spotting over the transcrigal database. In: Iberspeech 2014: VIII Jornadas en Tecnologa del Habla and IV SLTech Workshop (2012)
10. Neves, C., Veiga, A., Sá, L., ao, F.P.: Audio fingerprinting system for broadcast streams. In: Proceedings of Conference on Telecommunications - ConfTele. vol. 1, pp. 481–484. Santa Maria da Feira, Portugal (2009)
11. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
12. Povey, D., Hannemann, M., Boulianne, G., Burget, L., Ghoshal, A., Janda, M., Karafit, M., Kombrink, S., Motlcek, P., Qian, Y., Riedhammer, K., Vesel, K., Vu, N.T.: Generating exact lattices in the wfst framework. In: ICASSP. pp. 4213–4216. IEEE (2012)
13. Sandoval, A.M., Llanos, L.C.: MAVIR: a corpus of spontaneous formal speech in Spanish and English. In: Iberspeech 2012: VII Jornadas en Tecnologa del Habla and III SLTech Workshop (2012)
14. Schwarz, P.: Phoneme Recognition based on long temporal context. Ph.D. thesis, Brno University of Technology (2009)
15. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of International Conference on Spoken Language Processing. pp. 901–904 (2002)

# Albayzin 2014 search on speech @ ViVoLab UZ

Antonio Miguel, Julia Olcoz, Jesús Villalba, Alfonso Ortega and Eduardo Lleida

Voice Input Voice Output Laboratory (ViVoLab)
Aragon Institute for Engineering Research (I3A)
University of Zaragoza, Spain
{amiguel,jolcoz,villalba,ortega,lleida}@unizar.es

**Abstract.** This paper describes the system developed for the Albayzin 2014 spoken term detection and keyword spotting Evaluation. The system output is the fusion of the scores provided by an acoustic-phonetic recognizer, which uses continuous density HMMs with triphone contextual units and word begin and end markers. Two scores are obtained from the recognizer. The first one is the edit distance using the phoneme output sequence and the second one is the sum of the aligned confidences obtained from the recognition lattice. Finally, scores are calibrated with Bosaris toolkit.

**Keywords:** Spoken term detection, HMM, Fusion, Edit distance

## 1 Introduction

The system proposed by the Vivolab-UZ team from I3A at University of Zaragoza is the fusion of two metrics obtained from an acoustic-phonetic recognizer. The recognizer is composed of standard triphone contextual HMMs with mixtures of diagonal covariance Gaussians as observation density functions in the states.

This paper is organized as follows: in Section 2, the system is presented and each of the different subsystems discussed. The fusion system and the training and development data are also explained in Section 2. Conclusions are shown in Section 3.

## 2 System Vivolab-I3A-UZ

### 2.1 System description

In Figure 1 it is displayed the block diagram of the whole system. We can see how the different subsystems are used to produce the final scores. The process consists mainly in four steps.

The first step is the recognition of the sequence of phonemes and generation of lattices, which will be used by the other parts of the system for the search process. For this purpose we have used the tool HDecode from the HTK well

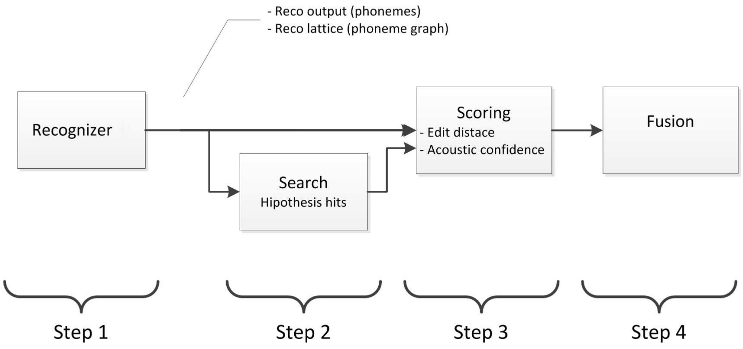


**Fig. 1.** Modules and processing steps of the search on speech system

known software [1] for all the systems. The particular details of the proposed recognizer are explained later.

The second step is the generation of hit hypothesis. The idea in the step is to use the recognizer outputs offline at search time to produce the hit hypothesis, which we will refer to as trials. The threshold for generating these hypotheses is quite low in order to miss a small number of targets even there are many non-target trials.

The third step is the scoring. We use the previously generated search hits with each recognizer to provide the scores to the fusion system. We obtain two scores for each recognizer: the edit distance and an acoustic confidence measure obtained from the lattice.

The final step is to merge the previous scores and fuse them to obtain a single score, which will be used to classify as target and non-target. For the fusion we have used a linear model trained with the Bosaris toolkit [2].

**Step 1, Recognizer** The recognizer is a standard phonetic decoder, with a trigram language model, that considers phonemes as words, and contextual triphone three state HMM units with mixtures of Gaussians in the states (diagonal covariance matrices). The features used in this recognizer are the ETSI MFCCs with derivatives and cepstral mean compensation, with a total dimension of 39. To increase the robustness of the system, histogram equalization was applied to the features.

**Language model**

In the case of a phonetic decoder, the output of the recognizer is the sequence of phonemes, then the phoneme is acting as a word. The trigram language model provides the likelihood of sequences of three phonemes $p(ph1|ph2, ph3)$, in the same way as if they were the words of a large vocabulary decoder. We have trained the phoneme trigram language model using the phonetic transcription

[1] http://htk.eng.cam.ac.uk/

[2] https://sites.google.com/site/bosaristoolkit/

of the training set, which is described later in this section. The language model has been estimated using the SRI Language model toolkit [3] with the default configuration.

**Acoustic model**

The acoustic model, as described before, is composed of standard continuous density contextual triphone HMMs trained using the audio described later in this section. The number of components in the mixtures is 32 per state and we use three states in the contextual units. Nevertheless, we included a modification in the standard procedure of HTK to train this kind of models. The acoustic units in our case were more specific and we defined special phoneme units for beginning and ending phonemes of each word, in a similar idea of head and tail models found in the literature [1], but in our case is used to obtain more specific phoneme models, not digit-digit word models. We provide an example of how the dictionary for training these units is written for some random words:

```
creido =k r e j D o=
llegaban =L e G a B a n=
enfrenten =e m f r e n t e n=
milagro =m i l a G r o=
organizamos =o r G a n i T a m o s=
flagrante =f l a G r a n t e=
aseguraba =a s e G u r a B a=
...
```

(Example of dictionary using specific phonemes)

We use the character '=' to specialize these phonemes being head or tail models. After this definition is done, we proceed to train the contextual units and to tie the more similar using the same procedure as the standard contextual training [2], so that the final number of units depends on the seen examples and the unseen are tied to those seen using a phonetic clustering tree. With this method the number of units grows with respect to a standard model but it still can be managed thanks to the clustering.

**Step 2, Generation of hypothesis hits** For the generation of hypothesis we used the output phonemes decoded by the previous recognizers as the source text for an edit distance search. Each possible target will be any substring which has an edit distance with the words we are searching of less than 50% of the length of the word in phonemes. This means that we need to correctly decode half of the phonemes of a word to be a hypothesis hit. Then we take all the hits provided by the different recognizers and we merge them when there is overlap using the best one. This will provide a list of trials, which will be scored in the next module. A trial will include the following information: the starting time, the ending time and the word. In the case of training and development data the trial has also information about the true labels whether is target or not target. This information will be used in the fusion step to train the system.

[3] http://www.speech.sri.com/projects/srilm/

**Step 3, Scoring** Once the list of trials has been created, we take each of them and we evaluate two scores for each recognizer, the edit distance using the recognition output and a confidence measure based on the lattice.

**Edit distance**

The edit distance is a well known algorithm where we measure the cost in number of substitutions, deletions and insertions to transform one string into another. For each trial we take the times stored in the previous step and measure the edit distance of the recognized phoneme sequence which lays between the limits and the word phoneme sequence. In order to use this as score, we simply use the negative of the edit distance divided by the length of the word as score for each trial.

**Confidence measure**

To obtain the acoustic confidence measure we follow these steps:

- First, we determinize the lattice so that a smaller and more useful graph is used in the next step. For this task we use the HTK command HLRescore.
- Second, we obtain an acoustic mesh graph of the decoded phoneme lattice using the lattice-tool from the SRILM toolkit.
- Third, The confidence calculated in the acoustic mesh graph is used in a modified edit distance algorithm where instead of costs all equal to one, we only sum the confidence of matching phonemes with the search word.

Then the score of a trial is the sum of the confidences through the acoustic mesh of the searched word between the time limits defined by the trial. This score is also normalized by the length of the word.

**Step 4, Fusion** We take all the scores for each trial and compose a matrix with the information of target and non target. This information is used in the Bosaris toolkit to learn a linear fusion. For the costs of miss and false alarm we have used the values: 1 and 0.1. For the probability of target prior we have taken: 10−2, which is estimated from our target rate in the hypotesis generator. Nevertheless, we have to say that this method is not providing the true error of the task since we have missed many true hypothesis hits in the previous step, which are not accounted as errors now. Then the error measure has to be calculated on the final result of the system.

## 2.2 Train and development data

**Training acoustic models** For training the acoustic models the following corpora have been used:

- Albayzin [3]. This is a phonetic balanced corpus which was recorded in a noise free environment. We have used 13600 files with a total length of 12.80h.
- Speech-Dat-Car [4]. This is a corpus recorded in a car in different driving conditions, where we have only taken the close-talk microphone. The number of files that we have used is 25378 and the total length 18.85h.

- Domolab [5]. This corpus was recorded in a domotic environment. We have used the close talk microphone and one of the lapel microphones. A total of 11998 files and 9.33h length.
- TCSTAR [6]. The TCSTAR corpus consists on a series of transcriptions of Spanish parliament sessions. We have used in the training a total of 43357 audio sections with a length of 111.89h.

A total length of 152.87h of training for the acoustic models.

**Training and development of fusion** For training the fusion we have used the previously explained procedure to obtain a list of trials. To obtain this list we have used all the Mavir [4] files provided for the training.

## 3 Conclusions

We have presented the fusion of two metrics for the spoken term search and keyword spotting evaluation. The methodology has been oriented to obtain a series of hypothesis hits using the edit distance with a loose threshold. These hits are fed to the scoring and fusion system in order to rank them according to the knowledge provided by more sources than a single system. The final scores have been calibrated using the training set, so that in test time the found words can be written to the output files when the fused score is greater than the calibrated threshold.

**Acknowledgments.** This work has been supported by the Spanish Government and the European Union (FEDER) through projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000".

## References

1. Juang, B.h., Chou, W., Lee, C.h.: Minimum Classification Error Rate Methods for Speech Recognition. IEEE Transactions on Speech and Audio Processing **5**(3) (1997) 257–265
2. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modelling. Proceedings of the workshop on Human Language Technology - HLT '94 (1994) 307
3. Casacuberta, F., García, R., Llisterri, J., Nadeu, C., Pardo, J.M., Rubio, A.: Development of Spanish Corpora for Speech Research (Albayzín). In: Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods. (1991)
4. Moreno, A., Lindberg, B., Draxler, C.: Speechdat-car: A large speech database for automotive environments. LREC 2000 (April) (2000)

[4] MAVIR corpus. http://www.lllf.uam.es/ESP/CorpusMavir.html

5. Justo, R., Saz, O., Guijarrubia, V., Miguel, A., Torres, M.I., Lleida, E.: Improving dialogue systems in a home automation environment. In: 1st International ICST Conference on Ambient Media and Systems. (February 2010)
6. Heuvel, H.V.D., Choukri, K., Gollan, C.: TC-STAR: New language resources for ASR and SLT purposes. In: In Proceedings of the LREC 2006. Number March (2006) 2570–2573

# Speech Recognition on the *Percepción* Project

Emilio Granell[1], Carlos-D. Martínez-Hinarejos[1], Guillermo Amat[2], Javier Fernández[3], Álvaro Arranz[3], Ángel Ramos[3], José Miguel Benedí[1], and Alberto Sanchis[4]

[1] PRHLT, Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain,
[2] FHC25, Roger de Lauria 19 5-B, Valencia, Spain,
[3] FHC25, Calle Rozabella, 4, Las Rozas, Madrid, Spain,
[4] DSIC, Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain,
{egranell, cmartine}@dsic.upv.es, {guillermo.amat, javier.fernandez, alvaro.arranz, angel.ramos}@bitmonlab.com,{jbenedi,josanna}@dsic.upv.es

**Abstract.** This paper presents the description of the implementation of a natural speaking environment in a Smart City scenario. Our system is implemented in a distributed architecture, including a speech recogniser with a speaker/device adaptation module and a dialogue manager to choose the most appropriate language model for each state of the interaction process. This system was developed under the scope of the *Percepción* project and it was successfully implemented in the *Universitat Jaume I* of *Castelló*, Spain, and in the *Centro Universitario de Tecnología y Artes Digitales* of *Madrid*, Spain.

**Keywords:** Speech recognition, speaker adaptation, distributed architecture, smart city

## 1 Introduction

The main objective of the *Percepción* project is to technologically assist people, giving to them a "sixth sense" to interact with the environment. This is achieved by using the increasing functionalities of current mobile devices with respect to processing environment information. Thus, by analysing input data and using interaction, the system would enhance human perception in a natural and transparent way.

Deploying the *Percepción* project resulted in an application for assistance and geolocation in a university environment called Smart Campus. The main implementation of this application was carried out on the campus of the UJI (*Universitat Jaume I* of *Castelló*), and another successful implementation was made in the campus of the U-TAD (*Centro Universitario de Tecnología y Artes Digitales* of *Madrid*). The Smart Campus application includes three tasks: indoor/outdoor positioning and navigation, augmented reality, and speech recognition.

In indoor/outdoor positioning, geolocation techniques alternative to the GPS are used, such as fingerprint based techniques (GSM, WIFI, and magnetic field), and mark based techniques (images, and QR-codes). The main advantage of these techniques is a lower energy consumption compared with the GPS energy consumption. All the interior/exterior spaces of the university campus were mapped in order to allow the indoor/outdoor navigation.

Augmented reality is used to show the information to users, such as data of relevant spaces contained on QR-code tags, and to show the route and the environmental information on indoor/outdoor navigation.

Speech recognition permits users to interact with the system in natural language speaking to find relevant spaces in the all university campus, such as buildings, parking spaces, waste containers, services (restaurants, sport areas, ...), and interior spaces (classrooms, offices, bathrooms, ...). In order to deal with the huge amount of different spaces, and to improve the quality of the recognition, it includes a dialogue manager, and a speaker/device adaptation module.

This project was managed by FHC25 (Factory Holding Company 25, S.L.), which is the head of a set of entities such as Ilion Animation Studios, U-TAD, and Pyro Mobile. At the same time, this project involved the INIT (*Instituto de Nuevas Tecnologías de la Imagen*) research group from UJI for the tasks of augmented reality and indoor/outdoor navigation. The PRHLT (Pattern Recognition and Human Language Technology) research centre from UPV (*Universitat Politècnica de València*) was involved into the speech recognition task.

Specifically, the ASR (Automatic Speech Recognition) task was developed as a distributed client-server architecture to delegate the power of computation to the server. In order to enhance the user experience, acoustic models are adapted to the speaker/device, and language model is selected by the dialogue manager from a set of language models. One of the most important capabilities of our ASR system is that it can be easily adapted to other tasks and languages, as the adaptation for the U-TAD Smart Campus. In particular, in the case of the implementation for the UJI Smart Campus application, our ASR system is simultaneously available on three languages (Spanish, Valencian, and English).

Currently, our ASR system is in operation at the UJI servers for the UJI Smart Campus application, although, it is still not available for use by the general public. The system has an acceptable user experience quality, according to the tests performed by the personnel involved in its development (PRHLT and FHC25), and the personnel involved in its integration into the UJI Smart Campus application (INIT).

The rest of this paper is structured as follows: Section 2 introduces the ASR system design and the details for the two Smart Campus implementations, Section 3 presents the ASR system architecture, Section 4 shows the models used by this ASR system, Section 5 presents the client/server communication protocols, and Section 6 contains the conclusions and future work.

## 2 Speech Recognition System Design

Our ASR system was developed on the project "*Percepción*" as an oral interface to find relevant places in a Smart City environment through a mobile application. The specific case of development, which was subsequently adapted to another case, was the UJI Smart Campus application. The functionalities of this application and the scope defined for each one of them are as following:

- Visor: permits to show elements with its available information on the map, such as university services, parking areas, and waste containers; there are about 100 elements identifiable by name.
- Finder: permits to find people, buildings, interior spaces, and routes between two points; there are 4453 spaces identifiable with a defined nomenclature, 65 different uses, and 1503 people (staff of the UJI); the start and end points are definable by coordinates on the route map.
- biciCAS: permits to calculate cycle routes in the city; the start and end points are definable by coordinates on the route map.
- Energy: shows graphs of energy and resources consumption (real time and monthly), such as electricity, gas and water.
- Space manager: permits to find spaces, to verify its availability, and to book it; it has the same scope than Finder, but only for spaces and uses.
- Service request: permits to request an intervention on a space or to report a problem or breakdown; there are limited types of intervention and problems.

The defined scope for the application is appropriate for its intended purpose and the type of interaction for which it was originally defined, where the use of speech interaction is available exclusively for the Finder functionality. However, the current state of the art of ASR techniques is far from the skills required to reach this scope through this methodology. Although there are available systems and applications that work well even for large generic vocabularies (such as those offered by Google), the Finder functionality of the Smart Campus application has features that are not solvable not even with today most advanced systems, such as the huge number of spaces and people, with their particularly nomenclature. This leads to a restriction of the system in its approach as speech recognition task.

The first restriction to the ASR system was to limit its use on the Finder to the location of relevant places (buildings, interior spaces, parking areas, waste containers, and services), leaving people search outside the scope. Other restrictions focus on the following dimensions:

– The number of accessible services for speech recognition: the variety of services with all its associated tasks require a recognition accuracy of speech unreachable with current systems. In the Smart Campus application the interaction through speech is one of the three available sources of input: speech, vision (pictures of QR codes or places), and tactile/text (gestures, digital touches, virtual keyboard text, handwriting, etc..). Therefore, the use of speech recognition for all interactions with the application is neither reasonable nor desirable. Leveraging the multimodal capabilities of the Smart Campus application, when ambiguity is reduced to 5 items, a list with all of them is displayed to the user to choose. Thus, the Smart Campus application determines in every moment what kind of interaction is the most appropriate and consequently limits speech recognition interventions to those situations where it is the most appropriate interaction modality.
– Restricting tasks on each service: vocabulary size in certain tasks (people search by name) is inaccessible with current technology. In the other tasks, the vocabulary can be divided according to the data that identify the items of the task. Consequently, we define a dialogue manager as a state machine to cover these data in a interactive way, where for each state there is a language model restricted to the data that such state seeks to fill. The dialogue manager can significantly narrow the context of the following interactions and incorporate this information to the recognition processes, reducing ambiguity and increasing the accuracy of the speech recognition.
– Restriction on speech naturalness of the user: the phenomena of spontaneous speech are reflected in an unpredictable formulation of the interactions, both phoneticaly and syntacticaly, which is still an open speech recognition problem. The ASR quality can be improved by adapting the acoustic models to the identified speaker and device.

The heart of the ASR system is the dialogue manager. The implemented dialogue manager is based on the definition of a number of repositories of data (*frames*) with a set of data associated with each repository (*slots*). The dialogue state is defined by a *frame* (or several *frames*, as there may be ambiguity) with which it is working and the value of its *slots*. In order to reach a final state with the searched item, the dialogue manager interacts with users requesting the information necessary to select a *frame* and fill its *slots*. Figures 1 and 2 show the design of the dialogue managers (UJI and U-TAD respectively) as finite state machines, where transitions depend on the values of *frame* and *slots* detected so far.

We chose to use the MLLR (Maximum Likelihood Linear Regression) technique [1] to adapt the acoustic models to the speaker/device in our ASR system. We conducted an experiment to justify this choice. We could verify that by using this technique with a single iteration and using only 5 sentences in Spanish to estimate a global diagonal transformation matrix, a significant improvement was achieved, falling from 36.8% to 19.4% of error in key words. This error rate is on the limits of what is permissible in a task such as Smart Campus, where subsequent interaction with the dialog manager may allow correct errors.

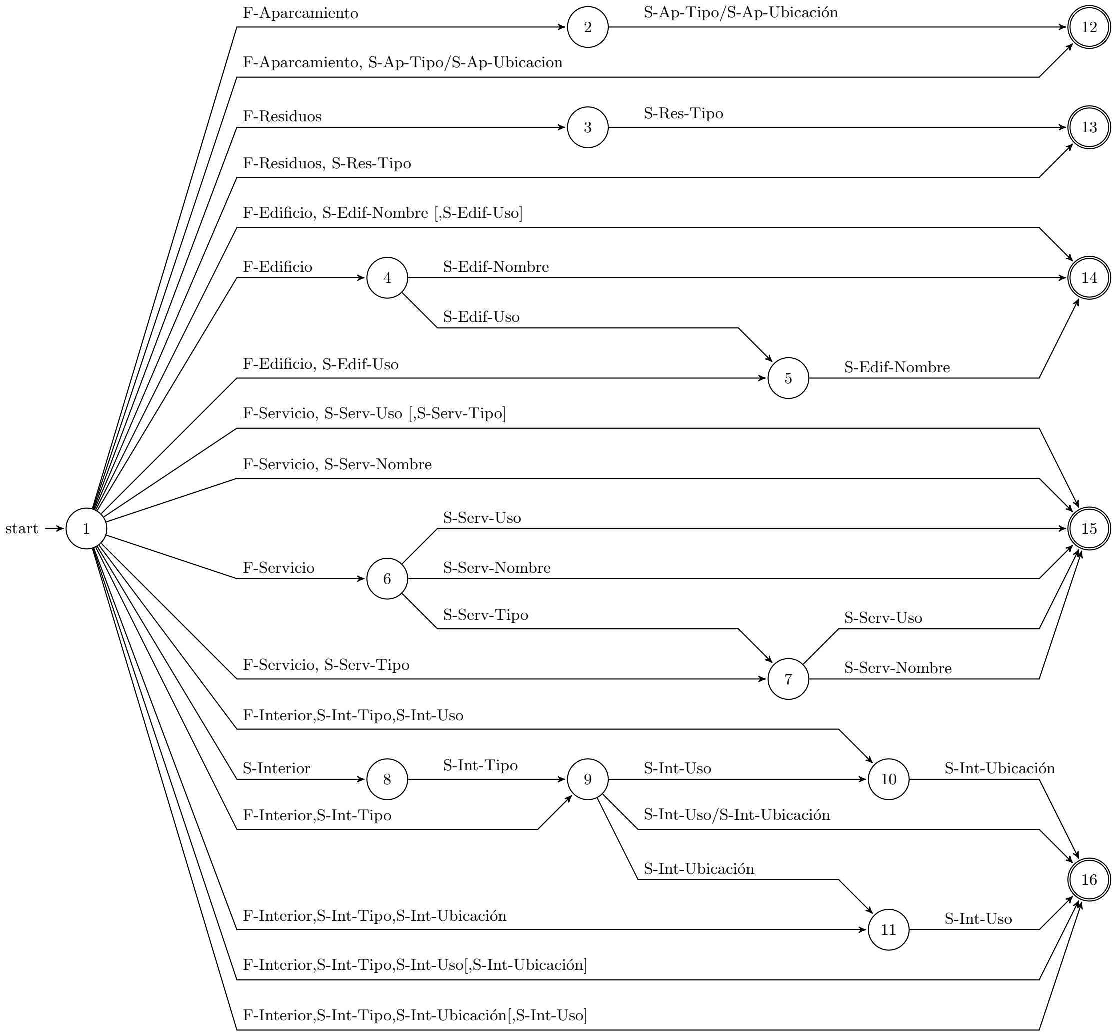


**Fig. 1.** States design for the dialogue manager (UJI Smart Campus).

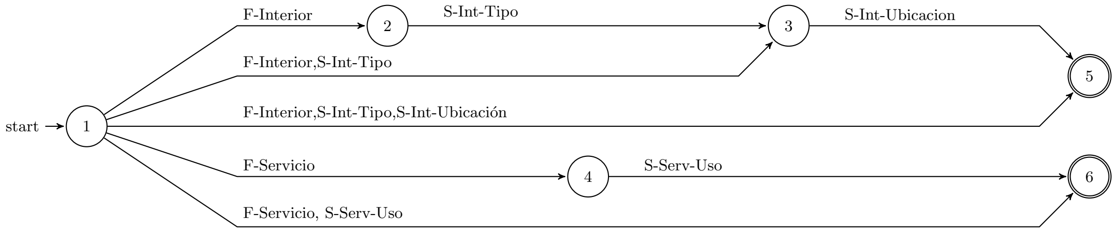


**Fig. 2.** States design for the dialogue manager (U-TAD Smart Campus).

In the following subsections, we present an overview of the two systems that were implemented with our distributed architecture.

### 2.1 Speech Recognition on the UJI Smart Campus

This task is related to the localisation of several entities in the UJI campus. There is a set of 4279 relevant locations in the database. To deal with the difficulty of differentiating each one of these items, we generated a dialogue manager (Figure 1) with five frames (*Aparcamiento*, *Residuos*, *Edificio*, *Servicio*, and *Interior*), including each one several slots. This dialogue manager has 16 states and its corresponding set of 16 language models. These language models were formed from a set of 63 categories. Lexical models include 472 words for Spanish, 443 words for Valencian, and 387 words for English. The automatically generated semantic models include only the relevant words, that are 344 words for Spanish, 347 words for Valencian, and 333 words for English.

### 2.2 Speech Recognition on the U-TAD Smart Campus

The implementation of the speech recognition module in U-TAD Smart Campus, whose goal is to support searches in the geolocation module through voice commands, includes 234 localisations distributed in a set of 5 categories. The lexical model includes 372 words for Spanish and the semantic model includes 111 relevant words. In addition, a dialogue manager has been created with six states (Figure 2) and two frames, named *Servicio* and *Interior*, including one and two slots respectively.

Additionally, in the U-TAD Smart Campus was made an improvement to facilitate the creation and maintenance of language, lexical, and semantic models, generating all of them automatically from the GIS (Geographic Information System) tool used in the project: JOSM (Java OpenStreetMap) [2]. All the relevant information is previously categorised and added as OSM (OpenStreetMap) [3] tags, representing states and key words. Therefore, all the items needed for the models are tagged within the OSM map file. Theses tags are treated and imported to a database and afterwards they are processed into the different models (language, lexical, and semantic model).

## 3 Speech Recognition System Architecture

Our ASR system has a distributed client/server architecture to delegate all the computational load to the server. The communication is done through TCP sockets and, in order to reduce the response time, the voice stream is send by streaming to the speech recogniser.

On the client side, our ASR client application is integrated into the Smart Campus application. The server side is composed by a speech recogniser, a semantic analyser, a dialogue manager, an adaptation manager, a data manager, and a database. As an example, the architecture of the implementation of our ASR system for the UJI Smart Campus application is presented in Figure 3. The main features of each component are:

- Client application: The Smart Campus client application was developed for Android; this application includes the three activities related with the ASR: `Speech` for the normal use of the ASR system, `Selection` to select the searched item when it exists a low ambiguity, and `Adaptation` to adapt acoustic models to the speaker/device (Figure 4).
- Database server: This server contains a database with all data users can inquire about.
- Speech server: The speech server contains all components of the server side of our ASR system.

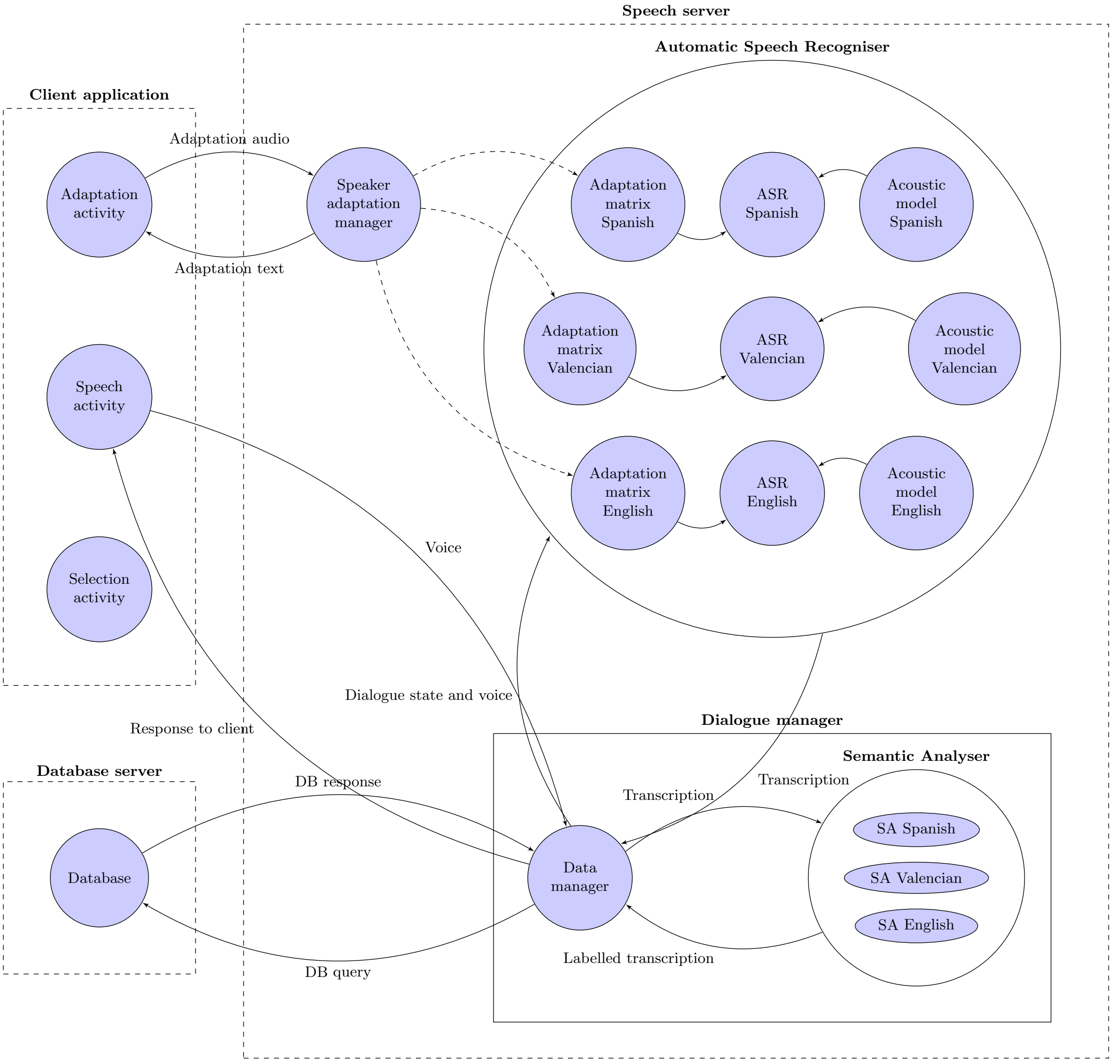
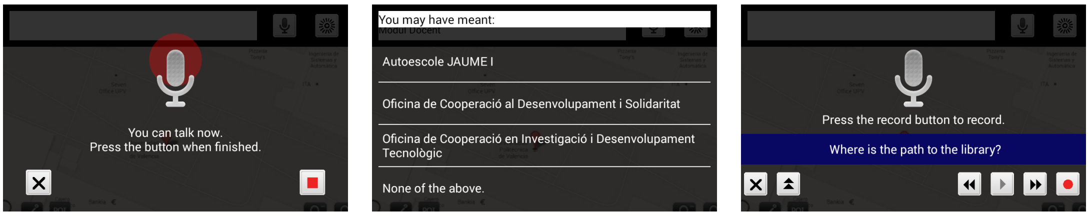


**Fig. 3.** Distributed speech recognition system on three languages with speaker adaptation (UJI Smart Campus).

- Adaptation manager: This server application sends the adaptation text to clients and receives the audio files in order to calculate the speaker adaptation matrices.
- Speech Recogniser: The used recogniser is an improved version of iAtros [4], that allows its use over TCP sockets and the adaptation of the acoustic models to speakers/devices. In the case of UJI Smart Campus application three recognisers are simultaneously working (one for each language, Spanish, Valencian, and English), each one configured with a different listening port. In contrast, on the U-TAD Smart Campus application two recognisers (for Spanish and English) are simultaneously working.
- Dialogue manager: It is responsible for determining the current dialogue state and selecting the appropriate language model for the speech recogniser. In the UJI implementation,



You can talk now.
Press the button when finished.

You may have meant:

Autoescole JAUME I

Oficina de Cooperació al Desenvolupament i Solidaritat

Oficina de Cooperació en Investigació i Desenvolupament Tecnològic

None of the above.

Press the record button to record.

Where is the path to the library?

(a) Speech activity. (b) Selection activity. (c) Adaptation activity.

**Fig. 4.** Client application screenshots.

text files are used to store the *frames* and *slots* values as well as the tracking of the user requests. In the U-TAD implementation the *frames* and *slots* values are stored using MongoDB [5] in JSON format [6], and the tracking of the user requests are stored using Redis [7]. In both systems the dialogue manager is formed by the following modules:

- Data manager: This is the core of the system and forms the dialogue manager along with the semantic analyser. It connects clients with the speech recogniser and updates the search list and semantic models from the database and the language models, respectively.
- Semantic Analyser: The semantic analyser searches the word forms in the dictionary to find out their lemmas and PoS (Part-of-Speech) tags, to allow the dialogue manager to select a *frame* and fill its *slots*. As semantic analyser we chose FreeLing [13] due to its multilingual capabilities and excellent documentation. We used the dictionary search module for PoS tagging. While in the UJI Smart Campus application it is used as an object through an API (Application Programming Interface), in the U-TAD Smart Campus application it is used through TCP sockets to improve the computational load balancing on the server.

## 4 Models

The ASR system uses four different types of models (acoustic, language, lexical, and semantic) in three languages (Spanish, Valencian, and English) on the UJI implementation and, two languages (Spanish, and English) on the U-TAD implementation. The features for the models are described in the following subsections.

### 4.1 Acoustic models

Since Castillian (Spanish) and Valencian are two different languages with phonetic similarities, we use the same acoustic models for both languages. These acoustic models were obtained by using the Albayzin corpus [8]; each acoustic model modelled a phoneme by using a left-to-right HMM (Hidden Markov Model) with 3 states and 64 gaussians per state. In the case of English, phonemes were modelled by using a left-to-right HMM with 3 states and 512 gaussians per state using the Wall Street Journal corpus [9]. These models were trained with HTK [10].

### 4.2 Language models

The dialogue manager uses a different language model for each dialogue state. The design of these language models is based on a set of finite-state models that determine a number of categories. These finite state models are in iAtros format [4].

For each dialogue state, a set of combinations of these categories was defined as the set of meaningful sentences in the corresponding state. These combinations were used for the inference of n-gram models (specifically trigrams) that employ the expansion on the defined categories. Additionally, all the vocabulary was added in order to smooth the language model. These n-grams language models were produced by the CMU-Cambridge Statistical Language Modelling Toolkit [11] in ARPA format.

### 4.3 Lexical models

Lexical models are formed by all the words contained in the language models with its phonetical transcription. To obtain the phonetical transcription for Spanish and Valencian we used a script based in transcriptions rules for Spanish (Valencian models were manually corrected). For English we used eSpeak [12] to get transcriptions in the International Phonetic Alphabet (IPA).

### 4.4 Semantic models

The semantic analyser FreeLing [13] needs a semantic model (a dictionary when using dictionary search) for each language in order to tag correctly the speech recogniser output. For each language, a semantic model was obtained by tagging automatically all the words contained on the corresponding set of language models. Tags link words with the *frames* and the *slots* of the dialogue manager.

## 5 Communication protocols

In our distributed speech recognition system, client applications can interact with servers in three ways: one for speech recognition (Figure 5) and two for speaker adaptation (Figures 6 and 7). First of all, clients must start connections with the correct password (in order to avoid unwanted connections) together with the ID (Identification of the Device) and the language chosen by the user. In the case of speech recognition (Figure 5) when a new connection arrives, the data manager server tests if the language, the speaker adaptation matrix, and the speech recogniser are available on the server, and informs the client about the situation. In this way, users can know if their language is not available, if they need to do the adaptation of acoustic models, or if the speech recogniser is not available at this moment.

When everything is correct, the server connects to the speech recogniser and sends to it the password, the dialogue state, the ID, and the language code. As soon as the speech recogniser accepts the connection, it adapts the acoustic models with the previously calculated user adaptation matrix, it selects the language model defined by the dialogue state, and it starts receiving the streaming of voice from the client. When the user finishes, the speech recogniser sends the recognition results to the data manager, where the dialogue manager uses the semantic analyser to “understand” the utterance and generates the answer for the user.

For adaptation we defined two different modes of connections. The first one (Figure 6) is performed when the text of the adaptation sentences is not present in the device. When a first mode connection arrives to the adaptation server, it replies with the text of the adaptation sentences. From this moment, the user can read the adaptation sentences at any time (Figure 4(c)). Once the user finishes recording the adaptation sentences, s/he can send it to the server. On the adaptation server side, when the user decides to send the audio files, a second mode connection is employed (Figure 7), which makes the adaptation server calculate the speaker adaptation matrix with all these audio files for the defined language.

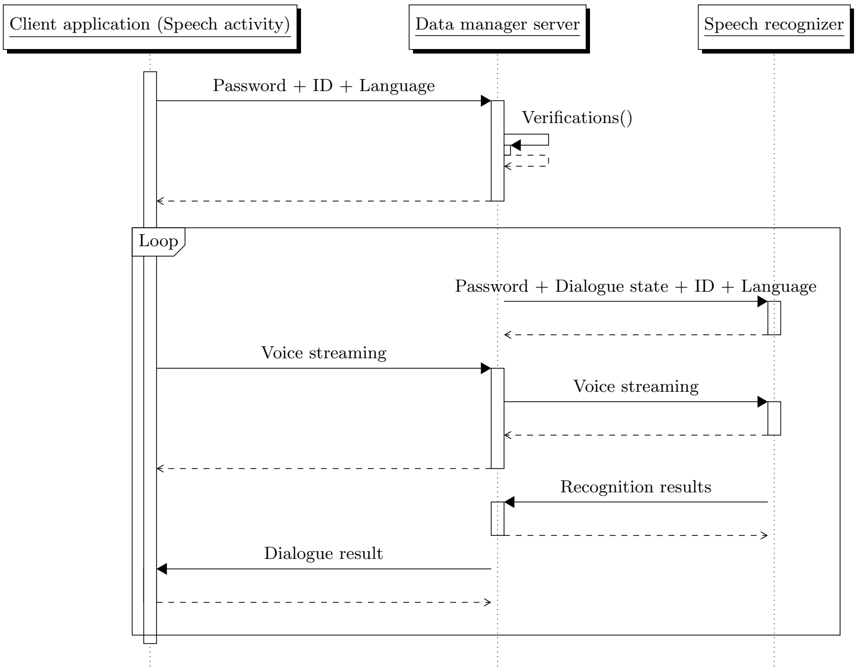


**Fig. 5.** Speech communication protocol.

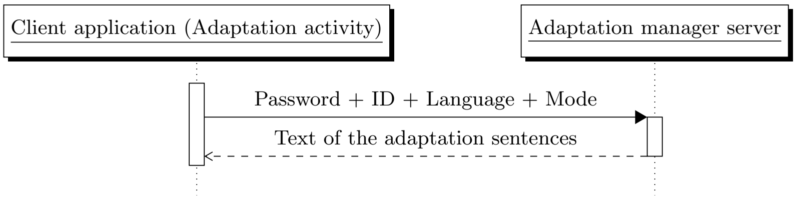

Client application (Adaptation activity)
Adaptation manager server
Password + ID + Language + Mode
Text of the adaptation sentences

**Fig. 6.** Communication protocol for the first mode connection for adaptation.

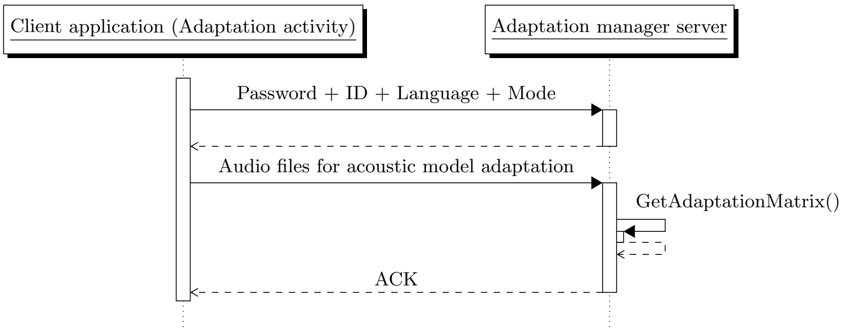

Client application (Adaptation activity)
Adaptation manager server
Password + ID + Language + Mode
Audio files for acoustic model adaptation
GetAdaptationMatrix()
ACK

**Fig. 7.** Communication protocol for the second mode connection for adaptation.

## 6 Conclusions and future work

This paper presented the distributed speech recognition system implemented for the *Percepción* project. The system includes a speaker adaptation module and a dialogue manager to enhance the quality of the recognition.

In the future, the recognition accuracy could be improved considerably by using the geographical and personal context on the dialogue manager, and by adapting the acoustic models to the different locations (noisy area, not noisy, interior, ...). Some other points that can be improved in the future are: the use of more robust acoustic models, as tandem or hybrid with neural networks, and the extension to other languages.

## References

1. C.J. Leggetter, and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer Speech & Language, vol. 9(2), pp. 171-185, 1995.
2. JOSM - Extensible editor for OpenStreetMap written in Java 7, `https://josm.openstreetmap.de/`, 2014.
3. M. Haklay, and P. Weber, “OpenStreetMap: User-Generated Street Maps”. IEEE Pervasive Computing, vol. 7(4), pp. 12-18, 2008.
4. M. Luján-Mares, V. Tamarit, V. Alabau, C.D. Martínez-Hinarejos, M. Gadea, A. Sanchis, and A.H. Toselli, “iATROS: A speech and handwritting recognition system”, V Jornadas en Tecnologías del Habla (VJTH2008), pp. 75–78, Bilbao, Spain, November 12-14, 2008.
5. E. Plugge, P. Membrey, and T. Hawkins, “The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing”, Apress, 2010.
6. T. Bray, “The JSON Data Interchange Format”, Standard ECMA-404, 2013.
7. S. Sanfilippo and P. Noordhuis. Redis. `http://redis.io/`, 2014.
8. A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Marino, and C. Nadeu, “Albayzin speech database: Design of the phonetic corpus”, Proceedings of the Third European Conference on Speech Communication and Technology (EuroSpeech’93), pp. 175-178, Berlin, Germany, September 22-25, 1993.
9. D.B. Paul, and J.M. Baker, “The design for the Wall Street Journal-based CSR corpus”, In Proceedings of DARPA Speech and Natural Language workshop, pp. 357-362, Harriman, New York, February, 1992.
10. 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 (for HTK version 3.4)”. Cambridge university engineering department, 2009.
11. P. Clarkson, and R. Rosenfeld, “Statistical language modeling using the CMU-cambridge toolkit. In Fifth European Conference on Speech Communication and Technology, Eurospeech, vol. 97, pp. 2707-2710, Rhodes, Greece, September 22-25, 1997.
12. J. Duddington, “eSpeak Text to Speech”, `http://espeak.sourceforge.net/`, 2012.
13. Ll. Padró, and E. Stanilovsky, “FreeLing 3.0: Towards Wider Multilinguality”, In Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May 21-27, 2012.

# Introducing the SpeechTech4All Web-based Platform for the Subtitling and Dubbing of Audiovisual Contents

Marta Martínez[1], Rocío Varela[1], Carmen Garcia-Mateo[1], Antonio Bonafonte[2], and
Inma Hernáez[3]

[1] Multimedia Technology Group(GTM), AtlantTIC Research Center, Universidade de Vigo, Spain
{mmartinez,rvarela}@gts.uvigo.es,carmen.garcia@uvigo.es
[2] Universitat Politecnica de Catalunya, Barcelona, Spain
antonio.bonafonte@upc.edu
[3] Aholab, University of the Basque Country, Bilbao, Spain
inma@aholab.ehu.es
http://speechtech4all.uvigo.es

**Abstract.** This paper describes the design guidelines of one of the two demonstrator platforms that are being developed in the SpeechTech4All research project. This demonstrator consists of a platform to automate or facilitate the production of subtitling and dubbing of audiovisual documents. Castilian, Catalan, Basque and Galician are fully integrated, also considering English as source language. The platform includes a structured database, a graphical interface and a number of processing tools. The use of a database enables to perform search in a simple and fast way based in a number of different criteria. The web-based user interface facilitates users the access to the different content. A number of tools for extracting metadata information regarding the speaker and the environment are under development.

**Keywords:** labelling tool, audiovisual content, multilingual processing

## 1 Introduction

SpeechTech4All is a three year project focused on advanced research in all core speech technologies (speech recognition, automatic translation and text to speech conversion) for the official languages of Spain. It is also aimed at the recognition of the speaker emotional state and the building of multimodal (speech and face) and multilingual (Spanish, Galician, Catalan and Basque) experimental work-frames. It is being developed by three research groups of the Universidade de Vigo, Universitat Politécnica de Cataluña and Universidad del País Vasco respectively.

As a result of the project significant research advances will be obtained in each of the above mentioned technologies. Some of these advances would be:

universal access to the service of synthetic voices customization, development of domain adaptation techniques in automatic translation and development of emotional state detection systems based in speech and face joint processing. To show the advances achieved in all the technologies considered, as well as to manifest the social character of the project, two showcases are defined:

- The first one integrates most of the technologies considered in the project and consists on the multilingual subtitling and dubbing of audiovisual material related to education: documentaries, speeches, seminars, etc.
- The second one is focused on one of the true applications of speech technologies: providing speech to people who have a severe level of oral disability by the use of a speech synthesis system adapted to the specific characteristics of the speaker.

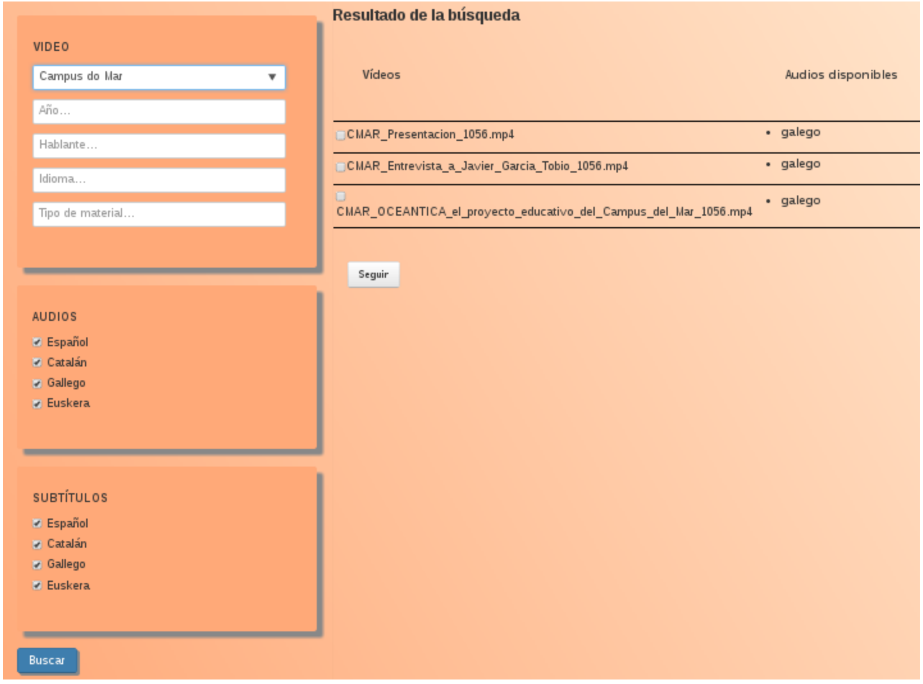
This paper describes the design guidelines of the first above mentioned showcase that is currently under construction. The goal of this demonstratior is to show the achievements of the project in the above mentioned technologies and their successful joint integration. The platform integrates several speech and language engines in different languages. A multimedia digital library is being created created from sources in all the official languages in Spain and English. The documents include News and debates TV/radio programs, newspapers and magazines. Next we will describe the different components that are being included in the tool and the various technologies used in its development.

## 2 Overview of the demonstration tool

This software repository includes a structured database that allow to organize the information, a graphical interface which enables the user to interact with the application and a number of processing tools. Next each component is described.

### 2.1 Database and software repository

SpeechTech4All training material and outputs of each language/video processing tool are hosted in a repository with three kinds of files interconnected each other: video, audio and subtitles. Video files are the core of each ensemble, so that several audio and subtitle files in different languages can be attached to one single video. The web-based interface lets the user to perform searches across the information stored in the database and return useful information about group of files to display (Fig. 1). Besides searching section, it's also possible to upload or remove files into the server (Fig. 2). To accomplish this operation, users have to log in with privileges. After that, they can see in the interface an area to drag and drop files that will be uploaded into the repository. In the other hand, files that are currently stored in the server can be displayed through three buttons that provide information about names and sizes. There is also a delete option for each file to remove it from the server.



VIDEO: Campus do Mar

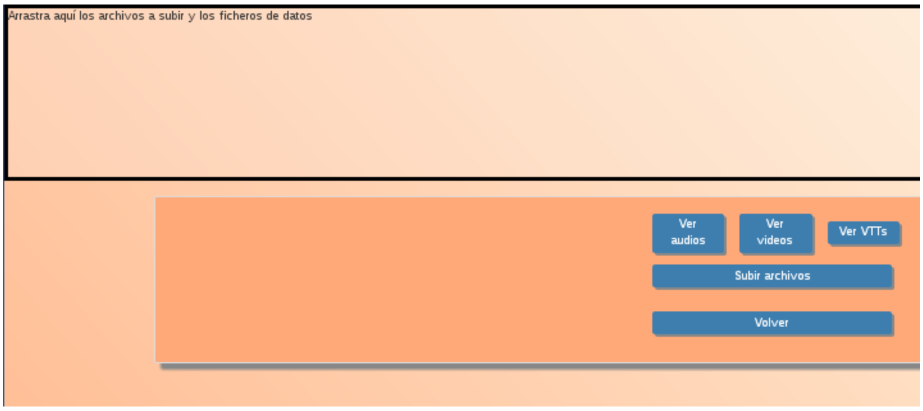
AUDIOS: ☑ Español ☑ Catalán ☑ Gallego ☑ Euskera

SUBTÍTULOS: ☑ Español ☑ Catalán ☑ Gallego ☑ Euskera

Resultado de la búsqueda

| Vídeos | Audios disponibles |
|---|---|
| □ CMAR_Presentacion_1056.mp4 | • galego |
| □ CMAR_Entrevista_a_Javier_Garcia_Tobio_1056.mp4 | • galego |
| □ CMAR_OCEANTICA_el_proyecto_educativo_del_Campus_del_Mar_1056.mp4 | • galego |

**Fig. 1.** Search section

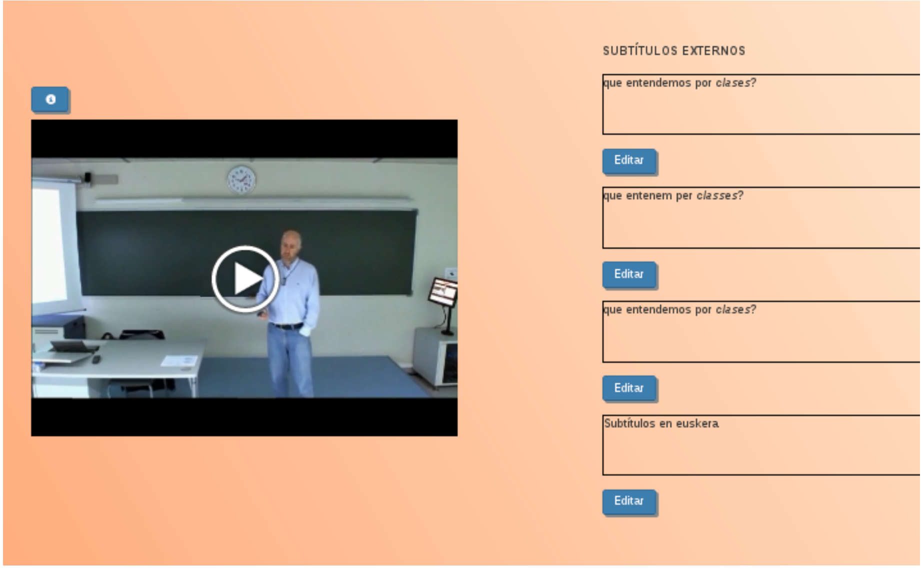


Arrastra aquí los archivos a subir y los ficheros de datos

**Fig. 2.** Administration section

### 2.2 Visualization section

This section is where results from different research areas merge. Visualization page provides to the users an interface to display video with external or integrated subtitle files (Fig. 3). It is also possible to choose full screen mode for video player. If audio is translated to more than one language, so the users are able to commute between different sources.



que entendemos por *clases*?

que entenem per *classes*?

que entendemos por *clases*?

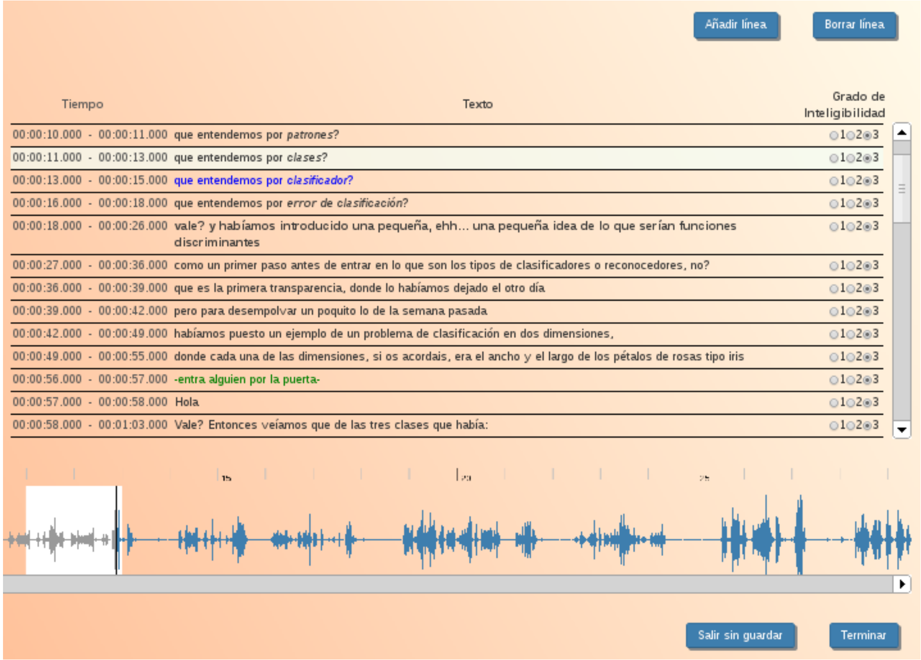
Subtítulos en euskera

**Fig. 3.** Visualization section

### 2.3 Edition and labelling section

One of the most important component of this platform is the subtitling and labelling tool. In this section, users can see a subtitle table, video player progress and a waveform advance (Fig. 4). The editable table is located next to the video and contains its transcription text in the selected language. All these elements are synchronized each other, thereby helping to change subtitle contents or time marks. In this way, when user clicks on a subtitle line, content modification is enabled. Also the user can listen to the audio loop area to help himself to transcribe.

Subtitle files are generated in Webvtt format, similar to srt format, when user press the button *Terminar*. The main advantage of this is the ability to add styles to subtitles that will be displayed in the video. Thanks to that, subtitles can be labelled with different tags to emphasize them in visualization or even to modify their position. Word spotting in subtitle lines is implemented such that



| Tiempo | Texto | Grado de Inteligibilidad |
|---|---|---|
| 00:00:10.000 - 00:00:11.000 | que entendemos por *patrones?* | ○1○2●3 |
| 00:00:11.000 - 00:00:13.000 | que entendemos por *clases?* | ○1○2●3 |
| 00:00:13.000 - 00:00:15.000 | que entendemos por *clasificador?* | ○1○2●3 |
| 00:00:16.000 - 00:00:18.000 | que entendemos por *error de clasificación?* | ○1○2●3 |
| 00:00:18.000 - 00:00:26.000 | vale? y habíamos introducido una pequeña, ehh... una pequeña idea de lo que serían funciones discriminantes | ○1○2●3 |
| 00:00:27.000 - 00:00:36.000 | como un primer paso antes de entrar en lo que son los tipos de clasificadores o reconocedores, no? | ○1○2●3 |
| 00:00:36.000 - 00:00:39.000 | que es la primera transparencia, donde lo habíamos dejado el otro día | ○1○2●3 |
| 00:00:39.000 - 00:00:42.000 | pero para desempolvar un poquito lo de la semana pasada | ○1○2●3 |
| 00:00:42.000 - 00:00:49.000 | habíamos puesto un ejemplo de un problema de clasificación en dos dimensiones, | ○1○2●3 |
| 00:00:49.000 - 00:00:55.000 | donde cada una de las dimensiones, si os acordais, era el ancho y el largo de los pétalos de rosas tipo iris | ○1○2●3 |
| 00:00:56.000 - 00:00:57.000 | -entra alguien por la puerta- | ○1○2●3 |
| 00:00:57.000 - 00:00:58.000 | Hola | ○1○2●3 |
| 00:00:58.000 - 00:01:03.000 | Vale? Entonces veíamos que de las tres clases que había: | ○1○2●3 |

**Fig. 4.** Edition section

it's possible to select words and store them into a key word file into the server for later use.

## 3 Technology description

Since users have to access from different locations, this SpeechTech4All demonstration tool was developed as a web-based software, so that it will run inside a web browser. Despite the fact that other browsers can be used, the tool is optimized for Google Chrome browser.

### 3.1 Client side

The programming language in the client part is HTML5[1]. The decision of using it was made on account of its good features to handle mutimedia information. One of the main advantages is that video and audio sources can be added easily with native code. In addition, it's also possible include subtitle tracks in video labels to display them over the sequence. There are two different kind of formats that can be added to the track label: traditional srt files or Webvtt[?] files. In this project we decide to use Webvtt file format because it provides support for labelling multimedia material with different CSS styles and move the subtitles along the display area of the video. These styles can be indicated in a CSS stylesheet or directly in the end of the time line of the Webvtt file.

Bound by the use of HTML5, we also use Javascript[7] programming language with JQuery library to handle DOM elements easily. In addition, some JavaScript APIs created to interact with HTML5 were very useful for these project: JavaScript Audio API[3] and Text Track API[4], to handle audio and subtitle files from JavaScript.

In order to improve the behaviour of the HTML5 video labelling, we include Mediaelement[5] plugin, coded in JavaScript. It improves video player controls with different options such as chapter division or more than one subtitle track selection.

Likewise, another JavaScript plugin called Wavesurfer[6] is used. It provides the interface support to draw the waveform and synchronize it with the rest of multimedia sources.

The interface also provides a notification area that keeps the user informed of what is happening on the server.

Finally, to dress the user interface, CSS stylesheets had been added to the project. Many of them are provided by Twitter Bootstrap framework and the others are customized styles that written in own stylesheet.

### 3.2 Description of the repository

The repository core was handle with MySQL database[8], so it's managed performing SQL language queries. This relational database has four tables:

- Video: contains video description and information (path in the server, video language, recording year, speaker names, video source, kind of record). This is the main table and their identifiers are used as foreign keys in other tables. It allows cascade delete avoiding the inconsistency of data.
- Audio: includes audio information (associated video identification, language and path in the server).
- Vtt: stores subtitle files information (associated video identification, language and path in the server).

There is also one additional table to handle big file uploading synchronization (such as videos or audios), due to the fact that they are sliced at client and recomposed on server.

Instead of storing files inside the database, the repository is hosted in different paths on the server. Thus, when we make a selection in the database, we receive a path where real file is stored and we use that file from its original folder.

### 3.3 Server side

Server side is written in PHP[9] programming language. It basically provides access to database information and manages client queries.

Apache web server was used to hold the application server part. It is installed in a Linux Debian operation system, but it could be exported to any other platform.

Communication between server and client was implemented using AJAX. This kind of communication is based in the use of asynchronous request between client and server. In this way, client can do different tasks while is waiting from server response, providing a good exploitation of application time while files are being uploaded or users are waiting for processed information.

## 4 Summary and further work

As a result, an interface to share a repository and display stored information had been developed. Furthermore, the demonstration platform integrates a subtitling and labelling tool for multimedia files, including software to perform its segmentation and classification. The collection of information to create databases required to train models is a challenge task that was assisted thanks to this software. In this way, information is stored in an ordered format and database contents are highly controlled by users. Moreover, it simplifies the human intervention to improve transcriptions. Furthermore, we are trying to improve labelling techniques to make them easier and automatized for the users and import new technologies that can be related with these project.

## 5 Acknowledgements

This work has been supported by the European Regional Development Fund and the Spanish Government ( SpeechTech4All Project TEC2012-38939-C03-01,TEC2012-38939-C03-02,TEC2012-38939-C03-03).

## References

1. Lubbers, P., Albers, B., Salim, F. Pye: Pro HTML5 Programming. (eds.) Apress 2011 http://www.w3.org/html
2. WebVTT: The Web Video Text Tracks Format, http://dev.w3.org/html5/webvtt/
3. Web Audio API, http://www.w3.org/TR/2011/WD-webaudio-20111215/
4. Media Multiple Text Tracks API, http://www.w3.org/WAI/PF/HTML/wiki/Media_Multiple_Text_Tracks_API
5. Mediaelement Plugin, http://mediaelementjs.com/
6. Wavesurfer Plugin, http://www.wavesurfer.fm/
7. Resig, John: Secrets of the JavaScript Ninja. (eds.) Manning Publications 2009 http://jquery.com
8. Ullman, Jeffrey D., Garcia Molina, H. and Windom, J.: Database Systems: The Complete Book. (eds.) Prentice-Hall 2002 http://www.mysql.com
9. Ullman, Larry: PHP and MySQL for Dynamic Web Sites. (eds.) Peachpit Press 2011 http://www.php.net

# The IRIS Project: A Liaison between Industry and Academia towards Natural Multimodal Communication

João Freitas[1,2], Sara Candeias[1], Miguel Sales Dias[1,3], Eduardo Lleida[4], Alfonso Ortega[4], António Teixeira[2], Samuel Silva[2], Cengiz Acarturk[5], Verónica Orvalho[6]

[1] Microsoft Language Development Center, Lisboa, Portugal
[2] Dep. Electronics Telecommunications & Informatics/IEETA, University of Aveiro, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal
[4] Universidad de Zaragoza, Spain
[5] Middle East Technical University, Turkey
[6] Faculdade de Ciências da Universidade do Porto, Portugal
[7] Instituto de Telecomunicações, Portugal

{t-joaof, t-sacand, midias}@microsoft.com,
{lleida,ortega}@unizar.es, {ajst,sss}@ua.pt,
acarturk@metu.edu.tr, veronica.orvalho@dcc.fc.up.pt

**Abstract.** This paper describes a project with the overall goal of providing a natural interaction communication platform accessible and adapted for all users, especially for people with speech impairments and elderly, by sharing knowledge between Industry and Academia. The platform will adopt the principles of natural user interfaces such as speech, silent speech, gestures, pictograms, among others, and will provide a set of services that allow easy access to social networks, friends and remote family members, thus contributing to overcome social-exclusion of people with special needs or impairments. Application of these features will be performed in the context of serious games, virtual reality environments and assisted living scenarios. The project will be executed in the scope of the Marie Curie Action Industry-Academia Partnerships and Pathways and will bring together the knowledge of five partners, from three different countries, Portugal, Spain and Turkey. This synergy will be based on transfer of knowledge mechanisms such as regular assignments and seminars, and is expected to have repercussions in scientific, technological, social and economic domains.

**Keywords:** Communication, Multimodal, Natural Interaction, Speech impairments, Knowledge Transfer.

## 1 Introduction

Communication is a process that facilitates integration into society, allowing humans to bond and to express their needs or desires. Thus, failing to do so may lead to social exclusion and, consequently, to psychological issues.

Speech is the main way for humans to communicate and in the last decades we have seen the rise of speech technologies, mainly in desktop and mobile devices, providing a more natural interaction with this type of machines. However, this kind of technologies is not yet adapted for users bearing speech impairments or users with more dissimilar speech patterns such as children or elderly.

In this context, focusing on the human communication process, we have created a project called IRIS, inspired in the ancient Greek goddess, who was a messenger of the gods. The overall goal of IRIS is to provide a natural interaction communication platform accessible and adapted for all users, especially for people with speech impairments and elderly in indoor scenarios, by sharing knowledge between Industry and Academia. Human-Computer interaction with this platform will adopt the principles of universal design and natural user interfaces such as speech, silent speech, gestures, gaze, tactile devices, pictograms, animated characters and personalized synthetic voices. The platform will provide a set of services that allow easy access to social networks, friends and remote family members, fighting social-exclusion of people with special needs or impairments. Application of these features will be performed in the context of serious games, virtual reality environments and assisted living scenarios. We will also explore the use of personalized avatars (that resemble the user) in asynchronous human-human and human-machine communications, in situations where the user is deprived of his/her voice and in scenarios where it is not possible to have a video signal transmission, due to low bandwidth or privacy reasons. Biometrics will complement the platform, in the sense that authentication and authorization are fundamental aspects for assuring secure access to personal information in a natural way. Figure 1 summarizes the research disciplines and potential outcomes from IRIS.

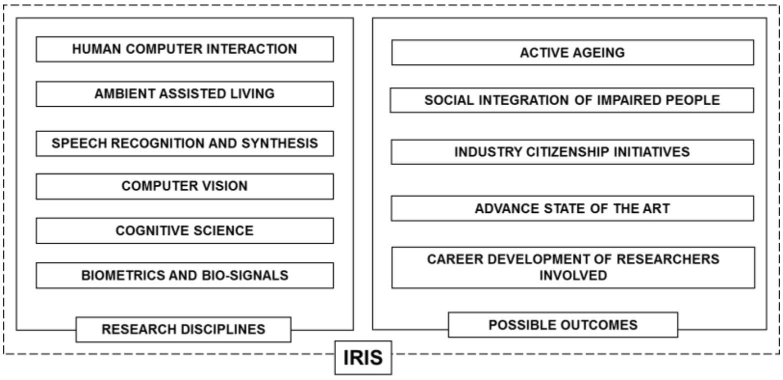

HUMAN COMPUTER INTERACTION
AMBIENT ASSISTED LIVING
SPEECH RECOGNITION AND SYNTHESIS
COMPUTER VISION
COGNITIVE SCIENCE
BIOMETRICS AND BIO-SIGNALS
RESEARCH DISCIPLINES
ACTIVE AGEING
SOCIAL INTEGRATION OF IMPAIRED PEOPLE
INDUSTRY CITIZENSHIP INITIATIVES
ADVANCE STATE OF THE ART
CAREER DEVELOPMENT OF RESEARCHERS INVOLVED
POSSIBLE OUTCOMES
IRIS

**Fig. 1.** Research disciplines and potential outcomes for IRIS.

This project will be carried out in the scope of the Marie Curie Actions framework under the Industry-Academia Partnerships and Pathways (IAPP) category. This research action aims at improving industry-academia collaboration in terms of research training, career development and knowledge sharing [1]. Thus, the proposed goals will need to rely on a successful transfer of knowledge between 5 partners, 2 from Industry and 3 from Academia. More details about the program can be found in [2].

The remains of this paper is structured as follows: Section 2 describes some of the target research topics and how they relate to IRIS; Section 3 presents the methodology adopted for this project, including how to leverage the mechanisms inherent to this type of projects (i.e. IAPP) to reach the designated goals; and finally, Section 4 describes the potential project impact from several perspectives.

## 2 Research Areas

IRIS is a multidisciplinary project that will leverage and advance the knowledge from multiple research fields. Overall, the main innovative aspects of this project are the following: a) provide an easy and natural way of accessing daily communication tasks through a multimodal interface that works for a diverse and distinct range of pathologies with particular emphasis on speech impairments and elderly people; b) analyze the impact of end-to-end communication scenarios with these interfaces, in virtual and real world scenarios, with users that otherwise were not able to conduct such tasks due to disabilities; c) create a solution and a set of resources that span and apply to multiple countries.

We aim at going beyond the state-of-the-art in a wide range of topics.These include elderly speech, silent speech interfaces, pictograms, voice based and multimodal biometrics, speech pathologies, and biological and psychophysical measures in the context of human-computer interaction (HCI) and cognitive science studies. These topics will be addressed on an individual basis and as a whole, with the ultimate goal of synchronizing all the advances into a single solution. In the sections below, we describe some of the related work in these areas and how they relate with IRIS.

### 2.1 Elderly-Computer Interaction

Elderly population individuals have developed resistance to conventional forms of HCI [3], like the keyboard and mouse, therefore making it necessary to test new natural forms of interaction such as speech, silent speech, touch and gestures. In addition, elder people often have difficulties with motor skills due to health problems such as arthritis. Therefore, proposing solutions that avoid small equipment, difficult to handle, may be an advantage over current solutions. It is also known that due to ageing, senses like vision become less accurate, hence difficulties in the perception of details or important information in conventional graphical interfaces may arise since current interfaces, most notably in the mobility area, are often not designed with these difficulties in mind.

There is also evidence that the European Union (EU) population is ageing rapidly [4]. This means that it is necessary to create solutions that allow overcoming the difficulties age brings to people who want to use new technologies in order to remain socially active. Elderly people who are connected to the world through the internet are less likely to become depressed and have greater probability of becoming socially integrated [5].

### 2.2 Speech Pathologies

Recent advances in machine learning and signal processing allow the detection of laryngeal pathologies through an automatic voice analysis. This is a promising field for speech therapists, mainly due to its non-invasive nature and objectivity for making decisions. Even the performance of automatic systems is still not perfect, thus they can be used as an additional source of "IRIS" information for other laryngoscopial exams [6]. Researching activity in this area is focused on: finding new features [7, 8], assessing quality of classification results [9], and researching newer classifier approaches [10].

In IRIS, a new vector space for voice pathology detection will be proposed, known as iVector space [11] that has become the state of the art in the fields of speaker and language recognition. This space is created as a factorization of the acoustic space in two terms. The first term accounts for the common information of the whole training database. In the second term, all sources of variability are taken into account and it is the one referred as iVector space or total variability space. Models for the normal and pathological classes will be trained by grouping iVectors from each of them separately and assuming they follow a Gaussian distribution. Speech pathology detection and classification will also be used on the IRIS project to gather speech information to build the user profile.

### 2.3 Silent Speech Interfaces

A Silent Speech Interface (SSI) performs Automatic Speech Recognition (ASR) in the absence of an intelligible acoustic signal and can be used as an HCI modality in high-background-noise environments such as living rooms, or in aiding speech-impaired individuals such as elderly persons [12]. By acquiring data from elements of the human speech production process – from glottal and articulators activity, their neural pathways or the central nervous system – an SSI produces an alternative digital representation of speech, which can be recognized and interpreted as data, synthesized directly or routed into a communications network . Recently, conventional ASR systems rely only on acoustic information, making them susceptible to problems like environmental noise, privacy, information disclosure and also excluding users with speech impairments (e.g. persons who have undergone a laryngectomy) or elderly citizens for whom speaking requires a substantial effort. To tackle this problem in the context of ASR for Human-Computer Interaction, we envisage for IRIS a novel multimodal SSI.

In view of the objectives established for IRIS, we have chosen a set of modalities that combine less invasive with the some of the most promising approaches found in the state-of-the-art. As such, it is our aim to build a multimodal SSI that uses Video and Depth information (i.e. RGB-D), Ultrasonic Doppler sensing, and Surface Electromyography. By combining multiple modalities we expect to address some of the challenges listed by Denby et al. [1], such as robustness, sensor positioning and nasality.

### 2.4 Pictograms

The IRIS platform is being designed by the principle of "design-for-all" which includes the population with spoken language difficulties. Communication and language are essential to every human being, to relate to others, to learn, to enjoy and to participate in society. For this reason, all people, whether children, youth, adults and the elderly, who for whatever reason have not acquired or have lost a level sufficient to communicate satisfactorily speech need to use an augmentative and alternative communication system.

The Augmentative and Alternative Communication Systems (AAC) are different forms of expression to the spoken language, which aim to increase (augmentative) and/or compensate (alternative) communication difficulties of many people with disabilities. AAC includes various symbol systems, both graphics (photographs, drawings, pictograms, words or letters) and gestural (mime, gesture or sign language).

IRIS makes use of pictograms for non-verbal communication. Pictograms are used as both, input and output modalities, in the way that a user can write sentences using pictograms and the system can communicate with users by means of pictograms. Pictograms are one of the preferred means to obtain information of the environment for those people that take special benefits from the visual channel, such as people with Autism Spectrum Disorders (ASD). IRIS will make use of the pictogram set provided by the Aragonese Portal of Augmentative and Alternative Communication (http://www.arassac.es).

### 2.5 Voice based and multimodal biometrics

Voice based biometrics or speaker identification begun in the 60s, with the study of several pattern recognition and feature selection techniques and has been receiving an increasing amount of attention from the research community [11].

The most widely used speech features are inspired in the human speech production system, which can be decomposed in the vocal tract and the excitation source (glottal pulse or noise generator). Recently, non-segmental acoustic features have been combined with spectral ones such as Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Coding (LPC), providing improved performance. Regarding the modelling side, the GMM-UBM (Gaussian Mixture Model-Universal Background Model) approach [13], has been recently substituted by the new techniques based on Joint Factor Analysis (JFA) [14] [15] solutions due to the limitations of the former for modelling channel variability in an appropriate way. Simpler and robust systems can be obtained with improved performance with respect to the JFA approach. Nevertheless, the accuracy of these systems can be considerably degraded if the quality of the speech signal is below the expected level or if the speech signal under analysis differs substantially from the speech samples used in the development process. Speech signal quality can be degraded due to several factors: low signal to noise ratio, saturation, acoustic artefacts, or presence of pathological speech. Therefore, the use of unimodal biometric systems presents limitations if the incoming data is noisy or if the variability is very high. Some other reasons to move to multimodal approaches are the potential non-universality of the unimodal variety (i.e. problems of speech production for some individuals, mobility impairments, etc.) or the robustness against spoof attacks.

In a multimodal biometric system, such as the one targeted in IRIS, the combination of the information coming from different sources of knowledge can be performed at different levels: Fusion at data or feature level, fusion at the match score level or fusion at the decision level. The multimodal approaches that integrate information at an early stage are usually preferred since they can offer improved performance and effectiveness [16]. Among the different modalities that can be used to build a multimodal biometric system, IRIS project will focus on the fusion of gesture biometrics with voice biometrics.

### 2.6 Multimodal Interfaces

Multimodal interfaces try to combine different modes of communicating, such as speech and gesture, in order to attain a more natural interaction and provide applications that are more engaging, exhibit more flexibility and contribute to improved accessibility, by different user groups in different scenarios, to tackle, for example, age-related impairments (e.g., loss of mobility, vision and hearing). Important aspects of multimodal interface development include: a) Modality Selection, guided by a full specification of requirements including a user profile, as complete as possible (addressing cognitive abilities, experience, physical characteristics, etc.), and characterizing the application domain, tasks and potential contexts of use [17]; b) Modality Mapping to different tasks and data types should be performed considering the characteristics of the human perceptual system and how each sense can respond to different task demands; c) Modality Combination is an important issue since assigning modalities to different tasks and data should not forget how modalities might interact, influencing how the user perceives information (e.g. to avoid sensory overload [18]).

The task of combining input modalities, to attain a single action is supported by d) Modality fusion engines [19]; e) System Adaptability also plays an important role for two main reasons: adapting to different user profiles (e.g., age, preferences and physical skills) in order to improve performance and satisfaction and adapting to changes in task and context [20]; f) Usability Evaluation of multimodal interfaces [21] should not forget some specific issues such as the importance of testing using context changes [22] and cognitive workload [23] conditions. The social impact [24] of the proposed systems must also be carefully assessed: the user might feel ashamed to use it in public if it somehow makes their difficulties or impairments more noticeable.

The main contributions of IRIS to the field of multimodal interfaces research can be described at three different levels: 1) Multimodal System, providing adequate/meaningful combinations of input and output modalities to support users in different communication tasks, adaptable to different usage contexts and user profiles, namely different pathologies and usage patterns; 2) Multidisciplinary approach provided by the different partners and their different areas of expertise allows approaching multimodal interfaces at different levels: software engineers, human computer interaction and usability experts, both the industry and academia points-of-view, and a set of infrastructures, namely a CAVE and an usability lab, allowing a strong emphasis on technically innovative approaches, usable/useful systems and market oriented solutions; 3) Design Options Validation: even though some empirical studies have been performed, this field clearly needs further contributions to validate design options on different application scenarios and, by doing that, contribute to increased knowledge to support further research and feed the industry with clear/proven guidelines to apply to new products. The conducted usability evaluations will provide valuable data to support or discard different design options leading to new or more specific guidelines for multimodal interfaces.

### 2.7 Biological and Psychophysical Measures

The broad range of recent research methods in HCI and cognitive science cuts across different research domains, such as cognitive psychology, educational psychology, psychophysics, and linguistics. These methods involve controlled experiments, questionnaires, in-depth interviews, focus groups, and usability inspection methods and cognitive modeling [25, 26]. Controlled experiments, being used widely to study human cognition and to evaluate human-computer interfaces for usability, measure different aspects of perceptual and cognitive processes by means of biological and psychophysical measures, such as response time, Electromyography (EMG: measures electrical activity produced by skeletal muscles) and Electro-dermal Activity (EDA: measures changes in the skin's ability to conduct electricity), eye/gaze tracking, and optical brain imaging (fNIR).

The basic motivation for studying eye movements is the relationship between eye movements and cognitive processes; in particular, the relationship between eye movements and visual attention. Eye tracking provides researchers with robust experimental data about online comprehension processes [27, 28]. Therefore, it has been widely used in HCI research since the past two decades. Eye movements, being indicators of cognitive processes, are used for the analysis of processing difficulties in visual displays, which in turn lead to guidelines and principles for design.

IRIS has the potential to contribute to the findings on the interdependency between the recent state-of-the-art biological and psychophysical methods, as well as on the investigation of the contribution of the complementary methods, i.e. interaction analysis and Quality of Experience (QoE) modelling, in evaluating multimodal interfaces.

## 3 Methodology

To support the framework behind this project it is necessary to define clear processes for knowledge transfer. Industry and Academia have different methods and different goals, thus a common ground between institutions needs to be found. Hence, more than defining which tasks lead to the stated objectives, we need to know how to execute these tasks using the resources available at IAPP projects and encompassing any restriction inherent to the institutions.

At an initial stage, it is important to have a clear idea of the knowledge held by each partner and how that knowledge is going to be integrated in the project. Thus, the first secondments are organized in way that there is an intersection between secondments from different institutions at the same destination. This method allows for researchers to share their experience and adequately plan future work. When the secondment ends, these researchers can take the acquired knowledge back to their institution and have a detailed overview of the project. For some areas, there is also the need to organize more specific courses given by experienced researchers, also open to the research community.

In terms of research plan, the work will be split into individual experiments/tasks that evolve towards the final objectives. The experiments will have multiple researchers assigned to them and should result in collaborative publications in an international conference, journal or book from the respective area.

Due to the multiple research disciplines encompassed by the project, a single research method (e.g. quantitative, qualitative) cannot be established. Therefore, these methods will be specified in a collaborative manner, by the researchers, before work on a specific task begins.

On a more generic level, the experiments will follow an iterative approach composed by the following stages: 1) problem identification; 2) problem assessment and state-of-the-art analysis; 3) development of a hypothesis/prototype; 4) application of test suite; 5) analysis and results comparison; 6) validation and conclusion formulation. To facilitate technological integration of all contributions, careful design and interoperability considerations are of paramount importance from the start of the project.

## 4 Impact

IRIS has potential repercussions in scientific, technological, social and economic domains, as depicted in Figure 2. From a scientific perspective IRIS aims to build the grounds for new ways of human computer-interaction and mediated human-human interaction. By combining multiple research disciplines and considering the defined objectives we expect not only to present novel ideas and solutions but also to make these applicable in the real world. As such, additional effort must be made towards realist but also innovative solutions. We also expect to create a strong impact in society, especially among those whose imparities somehow cause info-exclusion or avoid them from being more active in nowadays society. This aspect will be explored through existent partnerships with institutions and organizations that focus on social integration of people with impairments, senior universities that have daily contact with elderly citizens and special education schools that focus on the education of young students with special needs.

From an economical perspective IRIS has the potential to indirectly make the difference. By creating the mechanisms for a more informed and linked society, we expect to enable access not only to other people but also to products and goods available online. Likewise, we are opening the door for people that so far were not able to expose their ideas and skills to the community by lack of communication means. The involvement of both academia and industry in this project also allows establishing a bridge between research and concrete solutions that may strive into market.

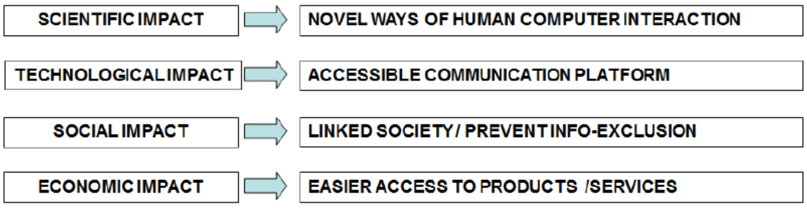

SCIENTIFIC IMPACT
NOVEL WAYS OF HUMAN COMPUTER INTERACTION
TECHNOLOGICAL IMPACT
ACCESSIBLE COMMUNICATION PLATFORM
SOCIAL IMPACT
LINKED SOCIETY / PREVENT INFO-EXCLUSION
ECONOMIC IMPACT
EASIER ACCESS TO PRODUCTS /SERVICES

**Fig. 2.** Impact areas for IRIS

**Acknowledgements**. This work was funded by Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP).

## References

1. Commission, E.: Work Programme: Marie Curie Actions. (2013).
2. Commission, E.: Guide For Applicants: IAPP Call Specific. (2013).
3. Phang, C.W., Sutanto, J., Kankanhalli, A., Li, Y., Tan, B.C.Y., Teo, H.-H.: Senior citizens' acceptance of information systems: A study in the context of e-government services. IEEE Trans. Eng. Manag. 53, 555–569 (2006).
4. Commission, E.: Europe in figures – Eurostat yearbook 2010.
5. Cisek, E., Triche, K.: Depression and Social Support Among Older Adult Computer Users. 113th Annual Convention of the American Psychological Association (2005).
6. Godino-Llorente, J.I., Gomez-Vilda, P., Blanco-Velasco, M.: Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. Biomed. Eng. IEEE Trans. 53, 1943–1953 (2006).
7. Parsa, V., Jamieson, D.G.: Identification of pathological voices using glottal noise measures. J. speech, Lang. Hear. Res. 43, 469–485 (2000).
8. Gavidia-Ceballos, L., Hansen, J.H.L.: Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. Biomed. Eng. IEEE Trans. 43, 373–383 (1996).
9. Tadeusiewicz, R., Wszolek, W., Modrzejewski, M.: The evaluation of speech deformation treated for larynx cancer using neural network and pattern recognition methods. Proc. EANN (1998).
10. Markaki, M., Stylianou, Y.: Using modulation spectra for voice pathology detection and classification. Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE. pp. 2514–2517. IEEE (2009).

11. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. Audio, Speech, Lang. Process. IEEE Trans. 19, 788–798 (2011).

12. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Commun. 52, 270–287 (2010).

13. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. Speech Audio Process. IEEE Trans. 3, 72–83 (1995).

14. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of interspeaker variability in speaker verification. Audio, Speech, Lang. Process. IEEE Trans. 16, 980–988 (2008).

15. Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P.: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. Interspeech. 9, 1559–1562 (2009).

16. Ross, A., Jain, A.: Multimodal biometrics: An overview. na (2004).

17. Obrenovic, Z., Abascal, J., Starcevic, D.: Universal accessibility as a multimodal design issue. Commun. ACM. 50, 83–88 (2007).

18. Hollender, N., Hofmann, C., Deneke, M., Schmitz, B.: Integrating cognitive load theory and concepts of human–computer interaction. Comput. Human Behav. 26, 1278–1288 (2010).

19. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimed. Syst. 16, 345–379 (2010).

20. Kong, J., Zhang, W.Y., Yu, N., Xia, X.J.: Design of human-centric adaptive multimodal interfaces. Int. J. Hum. Comput. Stud. 69, 854–869 (2011).

21. Wechsung, I., Engelbrecht, K.-P., Kühnel, C., Möller, S., Weiss, B.: Measuring the Quality of Service and Quality of Experience of multimodal human–machine interaction. J. Multimodal User Interfaces. 6, 73–85 (2012).

22. Naumann, A.B., Wechsung, I., Hurtienne, J.: Multimodal interaction: A suitable strategy for including older users? Interact. Comput. 22, 465–474 (2010).

23. Leung, R., MacLean, K., Bertelsen, M.B., Saubhasik, M.: Evaluation of haptically augmented touchscreen gui elements under cognitive load. Proceedings of the 9th international conference on Multimodal interfaces. pp. 374–381. ACM (2007).

24. Rico, J., Brewster, S.: Usable gestures for mobile interfaces: evaluating social acceptability. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 887–896. ACM (2010).

25. Cairns, P., Cox, A.L.: Research methods for human-computer interaction. Cambridge University Press (2008).

26. Sears, A., Jacko, J.A.: Human-computer interaction: Designing for diverse users and domains. CRC Press (2009).

27. Liversedge, S., Gilchrist, I., Everling, S.: The Oxford handbook of eye movements. Oxford University Press (2011).

28. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., Van de Weijer, J.: Eye tracking: A comprehensive guide to methods and measures. Oxford University Press (2011).

# A Multilingual Spoken Language Understanding System

Sergio Laguna, Mayte Giménez, Marcos Calvo, Fernando García,
Encarna Segarra, Emilio Sanchis, Lluís-F. Hurtado

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
{slaguna,mgimenez,mcalvo,fgarcia,esegarra,esanchis,lhurtado}@dsic.upv.es

**Abstract.** We have developed a Multilingual Spoken Language Understanding System which is able to understand an utterance regardless of its language. In this web-based demo we have integrated different modules: language identification, automatic speech recognition, translation and speech understanding.

**Keywords:** Multilingual Spoken Language Understanding, Language Identification.

## 1 Introduction

This demo shows a Spoken Language Understanding (SLU) System which is able to semantically decode sentences from a restricted domain, which in this case is an information system about train timetables and fares, allowing us to use the DIHANA corpus [1] to train the models of the system. Furthermore, it has been endowed with mechanisms to work with languages other than the one the system was trained for. Therefore, we developed an automatic translation system to translate from new languages (French and English) to the system's language (Spanish). Also, we developed an automatic language identification module to detect which language user spoke and automatically set the translator for that language. All led to the development of a multilingual understanding prototype.

### 1.1 From a monolingual to a multilingual spoken language understanding system

We already had a language understanding system for Spanish spoken utterances [2]. But our aim is to expand it, so it would be able to understand several languages.

There are two approaches to achieve this goal:

- Train a new language understanding system for each language we want to understand. In this approach, we need to acquire and label new data and train a new system for each new language. This implies a significant effort.

- A semi-supervised approach for adapting a monolingual understanding system to sentences uttered in a new language. This, obviously simplifies porting SLU systems between languages. In a previous paper [3] we describe in deep this approach.

## 2 Description of the system

We are going to describe the Multilingual Spoken Language Understanding System we have developed.

Our process begins when the user provides an audio file to our system. This can be either and existing audio file, or can be recorded using the web browser.

Then, the system performs the following steps.

1. The system identifies the language of the audio choosing between English, French or Spanish.
2. Given a language, detected in the preceding step, the audio is recognized using a web-based recognizer.
3. If the detected language is not Spanish, the system will translate it into Spanish, as it is the language in which the models of the system are.
4. Then, a graph of words is created using the n-best hypothesis provided by the translator.
5. Finally, the system performs the semantic decoding of the graph of words, obtaining this way a set of detected concepts and a sequence of words associated to each concept.

Below, the system components are briefly described.

### 2.1 Language identification

Language identification can be stated as a classification problem.

We have developed a two phases approach to language identification:

- Acoustic-Phonetic Decoding ($APD$) of the spoken utterance using a set of Spanish phoneme models. This phase always uses the same phoneme models.
- Using the phonetic sequence from the transcription, the system assigns a language to it. It uses a language model of sequences of phonetic units learned for each language. The selection criterion is based on minimize the perplexity.

We used triphones as phonetic units, therefore we have context information. The acoustic model was learned from out-of-task corpus in Spanish. Moreover, the model of sequences of triphones used as language model was a trigram model of phonetic units.

Language models were learned from the $APD$ output of 3446 spoken sentences uttered by several native English, French, and Spanish speakers. The English and French sentences are a translation of the DIHANA [1] corpus, which domain is restricted to information of long distance trains. However, Spanish sentences belong to the ALBAYZIN corpus which is a general-domain phonetically balanced corpus.

### 2.2 Automatic speech recognizer (ASR)

The input utterance is processed by an ASR for the language identified in the previous step. We use a general purpose, free-available web ASR, which means the ASR has no specific information about the task.

### 2.3 Automatic translation

These transcriptions are translated into the target language (Spanish) using a state-of-the-art Machine Translation system: MOSES [5]. The translation models were trained without using any manually translated data. Instead, a set of free-available web translators was used to translate the training sentences of the corpus from Spanish into the different languages, thereby building a parallel training corpus. MOSES provides as output a set of translation candidates (n-best list) of the transcription supplied by the ASR.

### 2.4 Graph of words generation

Our SLU system uses the n-best sentences obtained in the previous steps to generate a graph of words [4]. Figure 1 shows how this graph is obtained. First, a Multiple Sequence Alignment step is performed, in which we use a modification of the ClustalW software [6]. Then, we build the graph using the information contained in the alignment matrix, and computing the probabilities by the Maximum Likelihood criterion.

**source utterance**

*le prix du billet de train de six heures trente-cinq*
*(the price of the train ticket for six thirty-five)*
*(el precio del billete del tren de las seis treinta y cinco)*

**multiple ASR outputs**

le prix du billet train de sezer trente-cinq
le prix du billet train de six vers trente-cinq
le prix du billet train de six onze trente-cinq
le prix du billet train des six heures trente cinq

**multiple translations**

el precio del billete de tren de sezer treinta y cinco
el precio del billete de tren alrededor de las seis treinta y cinco
el precio del billete del tren de las seis once y treinta y cinco
el precio del billete de tren de las seis treinta de las cinco de la tarde

**alignment matrix**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| el | precio | del | billete | de | tren | - | de | sezer | treinta | - | - | - | y | cinco | - | - | - |
| el | precio | del | billete | de | tren | alrededor | de | las | seis | treinta | - | - | y | cinco | - | - | - |
| el | precio | del | billete | del | tren | - | de | las | seis | once | y | treinta | y | cinco | - | - | - |
| el | precio | del | billete | de | tren | - | de | las | seis | treinta | - | de | las | cinco | de | la | tarde |

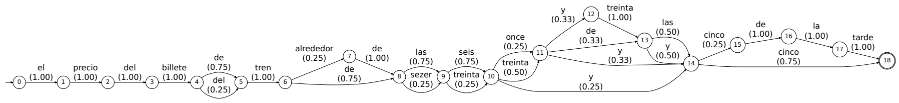


**Fig. 1.** Steps for obtaining the graph of words from the original utterance *le prix du billet de train de six heures trente-cinq*, (*the price of the train ticket for six thirty-five*).

### 2.5 Spoken Language Understanding

The goal of this module is to provide the best semantic interpretation according to the information encoded in the graph of words. This semantic interpretation is defined according to a restricted domain, which in this case is the scope of the DIHANA corpus, it is, an information system about train timetables and fares. This corpus provides a set of concepts that are relevant for the domain, as well as a set of utterances and their transcriptions, which are segmented and labeled in terms of these concepts. This information is very useful to train statistical semantic models. This way, we have trained statistical models to represent how the words are joined within a specific concept, as well as a model to represent how the concepts are chained. Table 1 shows an example of the output of the system, assuming that the input is a written sentence.

| | |
|---|---|
| **Input utterance** | *hola buenos días quería saber los horarios de trenes para ir a Madrid*<br>*(hello good morning I'd like to know the train timetables to go to Madrid)* |
| **Semantic segments** | *hola buenos días* : courtesy<br>*quería saber* : query<br>*los horarios de trenes para ir* : <time><br>*a Madrid* : destination_city |

**Table 1.** Example of the outputs of the SLU module.

The semantic decoding algorithm uses the statistical semantic models to process the graph of words. This algorithm first finds the possible attachments of sequences of words represented by a path in the graph to any of the concepts of the task. Then, it finds the best sequence of concepts using these attachments [3]. In previous works we have evaluated the behavior of this approach to multilingual SLU, achieving a 82.28% of concept accuracy when the input language is Spanish and 77.60% for French.

## 3 Experimental results

In this section, we present different experiments carried out to evaluate the performance of our system that validate the hypothesis we set.

### 3.1 Language identification minimizing the perplexity

In order to verify our approach for language identification we have conducted several experiments. We used the SRILM Toolkit [8] to estimate the phonetic language models for the classiers and the HTK Speech Recognition Toolkit [9] to perform the phonetic transcriptions.

Previously, we have published the results of this experimentation [7]. Summing-up, as we expected, lower perplexity appears when the language of the sentence and the language of the model are the same. Moreover, we evaluated the performance of the Language Identication system. The global accuracy of the system was 0.841.

Table 2 shows the perplexity of the test set for the different languages and the accuracy of the system. It shows that the use of trigrams of phonetic units learned using a corpus only in Spanish is not as critic as we a priori expected. Overall, this approach allows us to identify the language of spoken utterances with limited resources.

| | | Perplexity | | |
|---|---|---|---|---|
| | | Test language | | |
| | | French | English | Spanish |
| LM Trigrams APD | French | **8.24** | 11.62 | 12.16 |
| | English | 10.79 | **6.63** | 11.29 |
| | Spanish | 11.27 | 10.86 | **7.57** |

| | Accuracy | | |
|---|---|---|---|
| | Test language | | |
| | French | English | Spanish |
| LM | 0.793 | 0.850 | 0.960 |

**Table 2.** Perplexity of the phonetic language models and accuracy of the system *minimizing the perplexity.*

### 3.2 Changing the acoustic model

In order to performe the APD of all user utterances, we used an acoustic model trained only with Spanish audios from the TC-STAR corpus. We thought that our APD could improve with an universal acoustic model that takes into account also the acoustic variations of English and French. However, we did not have the resources to train this universal acoustic model, so we trained a new acoustic model using only audios in Spanish and English.

| | | Detected language | | |
|---|---|---|---|---|
| | | FR | EN | SP |
| TC-STAR AM | French | **67%** | 8% | 25% |
| | English | 8% | **67%** | 25% |
| | Spanish | 13% | 8% | **79%** |

| | | Detected language | | |
|---|---|---|---|---|
| | | FR | EN | SP |
| DIHANA AM | French | **82%** | 13% | 5% |
| | English | 4.1% | **92,9%** | 3% |
| | Spanish | 0,3% | 0,3% | **99,4%** |

**Table 3.** Language identification using different acoustic models

From the results shown at Table 3, a universal acoustic model improves language detection. However, when we used it with real users the results were not as good as expected.

## 4 Conclusions

In this paper, we have presented a web-based Multilingual Spoken Language Understanding system. We integrated different modules in this demo: language identification, automatic speech recognition, automatic translation and an understanding system that we adapted to extract semantic interpretations from sentences in different languages.

## References

1. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López de Letona, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proceedings of LREC 2006. pp. 1636–1639. Genoa (Italy) (May 2006)
2. Calvo, M., García, F., Hurtado, L.F., Jiménez, S., Sanchis, E.: Exploiting Multiple ASR Outputs for a Spoken Language Understanding Task. In: Speech and Computer, pp. 138–145. Springer International Publishing (2013)
3. Calvo, M., Garcıa, F., Hurtado, L.F., Jiménez, S., Sanchis, E.: Exploiting multiple hypotheses for Multilingual Spoken Language Understanding. CoNLL-2013 pp. 193–201 (2013)
4. Calvo, M., Hurtado, L.F., García, F., Sanchis, E.: A Multilingual SLU System Based on Semantic Decoding of Graphs of Words. In: Advances in Speech and Language Technologies for Iberian Languages, pp. 158–167. Springer (2012)
5. Koehn, P., et al.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proc. of ACL demonstration session. pp. 177–180 (2007)
6. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G.: ClustalW and ClustalX version 2.0. Bioinformatics 23(21), 2947–2948 (Nov 2007)
7. Sanchis, E., Giménez, M., Hurtado, L.F.: Language identification with limited resources. V Jornadas TIMM pp. 7–10 (2014)
8. Stolcke, A., et al.: SRILM-an extensible language modeling toolkit. In: INTERSPEECH (2002)
9. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al.: The HTK book, vol. 2. Entropic Cambridge Research Laboratory Cambridge (1997)

# Subtitling Tools Based on Automatic Speech Recognition

Adolfo Arguedas, José Enrique García, Eduardo Lleida, Alfonso Ortega,
Antonio Miguel

Aragon Institute for Engineering Research, University of Zaragoza, Spain
{arguedas,jegarlai,lleida,ortega,amiguel}@unizar.es

**Abstract.** Automatic Speech Recognition (ASR) has been widely used for subtitling to reduce costs derived from human intervention. When the text to be subtitled is known beforehand, the quality of the ASR-based subtitling can be further improved by means of speech and text synchronization. In this paper we present a suite of programs that using ASR can help to optimize the productivity in the subtitling process for both live broadcast news and off-line recorded programs. The suite also contains a real time monitoring tool of television channels that measures the correspondence of the broadcast subtitles with the audio, in order to detect and correct failures (i.e. subtitles emitted to the wrong channel) ensuring the quality of the subtitling task without human supervision.

**Keywords:** Automatic Speech Recognition · Text-Audio Synchronization · Subtitling System, confidence measures

## 1 Introduction

Automatic Speech Recognition (ASR) is a fundamental technology in the subtitling task, used mainly when the available human resources are limited. In live emissions whose text is not known, a large vocabulary continuous speech recognition (LVCSR) system is usually adopted [3, 4]. Some systems transcribe directly the broadcast audio, and if specialized staff is available, the transcription is done over the speech of a shadow speaker [1, 2] that repeats the broadcaster words and uses an ASR systems adapted to his voice. These LVCSR-based subtitling systems have some deficiencies, such as the long delay of the process that can reach up to 10 sec. Moreover, adverse acoustic conditions (low audio quality, high background noise or music, overlapped speech, etc.) will produce a poor subtitling quality. In addition to this, some errors can completely change the meaning of the sentence, and sometimes they can be controversial.

The quality of the subtitling can be dramatically improved if the text to be subtitled is known beforehand, as it usually occurs in news broadcast. Thus, the problem can be simplified to a more constrained one, audio and text synchronization [5]. The main advantage of the synchronization is that the text will not contain recognition errors. These systems present also negligible delay, and the quality of the generated subtitles is better since the whole text is known beforehand (even of future subtitles). The speech-text alignment problem, though simpler than LVCSR, doesn't lack challenges to investigate and solve [6]. There are two major problems to address, how to discard the audio that is unrelated to the text, and how to discard the text that is not present in the audio. An additional problem that usually face these systems is the frequent modifications of the text made by the editors, including the story currently being subtitled.

Three subtitling tools have been developed that exploiting text-audio synchronization technology, provide an efficient solution for subtitling: ZARNews for live broadcast news, ZARCaptions for previously recorded programs and ZARMonitor for monitoring broadcast subtitles in different channels of a television for quality assurance.

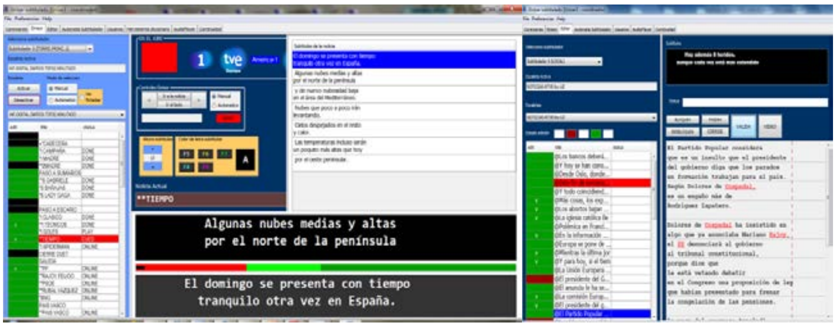
## 2 ZARNews

ZARNews is oriented to live broadcast news subtitling from texts generated by the editors of the newsroom. It is a flexible tool that allows both automatically or manually scripted subtitling, with automatic subtitling generation according to AENOR UNE 153010 quality rules [7] that include recommendations such as finishing a subtitle in punctuation marks, trying to make each subtitle as much self-contained as possible …

ZARNews (figure 1) integrates a proprietary ASR system that captures live audio and synchronizes it with the news texts coming from the newsroom system iNews. A thorough description of this system can be found in [5].

The software architecture is composed of four services:

- Communication manager. Control of the communication among services of the application and for storing and managing all system users' information.
- Subtitling service. ASR engine for audio-text synchronization and story detector [5] for re-positioning to the current story if the system gets lost.
- Subtitle generator/spelling checker service. Automatic partition of the text and assistance for checking and correcting eventual orthographic errors.
- INews text recovery client. It receives information from the newsroom system and sends it to the communication manager



**TIEMPO

Algunas nubes medias y altas
por el norte de la península

El domingo se presenta con tiempo
tranquilo otra vez en España.

**Fig. 1.** Screenshot of the scripted subtitling system (left) and editor view (right).

The automatic subtitling process starts when the continuity system indicates that a new program is ready to start. Then, all the stories and their corresponding automatically generated subtitles are sent to the Subtitling Server that will start synchronizing the news. From the communication manager the text of every story is sent to the subtitle generation and the spell checking services. In addition to this, the subtitles of all the stories of the currently emitted program are stored in the subtitling service.

Other relevant features of the system are:

- Simultaneous emission of the same subtitles to multiple TV channels.
- Reissue of broadcast subtitles for different TV channels.
- Powerful pre-broadcast edition system, newsroom system independent, with spell checker, change control, multi user editing, notification of changes in the editing system, etc.
- Multi user management, with different functionalities for different user profiles and connectivity to multiple communication managers simultaneously.
- Multiple rundowns management with automatic activation for unsupervised broadcast.

Finally, the system also includes the ability to subtitle unscripted programs by making use of a LVCSR with acoustic models adapted to the broadcaster. This system is being used for subtitling weather information in TVE.

## 3 ZARCaptions

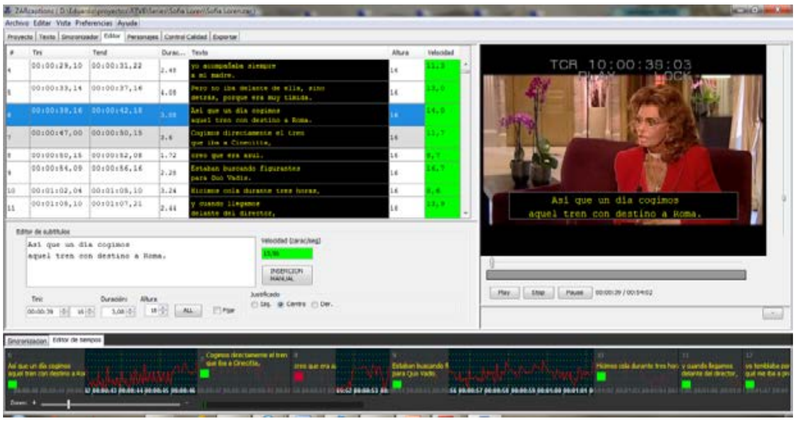
ZARCaptions is a tool that allows automatic subtitling for previously recorded programs. It includes tools to accommodate the original script to the subtitling back-end format, powerful subtitle post-editor, characters assignment capabilities and quality measures by means of confidence measures.

ZARCaptions integrates software that automatically executes a subtitling synchronization process and with real time factors that ranges from 30 to 1, depending on audio and text quality. The system guides the user over the text segments that are likely to have problems allowing manual editing and automatic subtitling refinement (figure 2).This system can be also useful to resynchronize live broadcasted subtitles (by means of re-speaking) for subsequent rebroadcast. The synchronization process can be run either locally or connected to a remote synchronization server, this way it can be easily adapted to different hardware resources.

Another interesting features of the system:

- Creation of project files for the organization of the different files that are used in the process (script, subtitles, audio, video…), auto-saving.
- Editing and text filtering tools to preprocess the script and adapt it to the subtitling process, before the synchronization step.

- Subtitles post-edition GUI, that allows a human to adjust  with more precision the subtitles time-stamps, heights, characters or colors.
- The system performs the synchronization automatically, subtitle partitioning and assignment of colors to the different characters of the program or movie, if they have been marked in the text.
- Quality control system, which can detect potential errors in the subtitle editing and temporal overlaps between subtitles.


Archivo Editar Vista Preferencias Ayuda
Proyecto Texto Sincronizador Editor Personajes Control Calidad Exportar
TCR 10:00:38:03
Play Stop Pause
Sincronizacion Editor de tiempos

| # | Tini | Tend | Durac... | Texto | Altura | Velocidad |
|---|---|---|---|---|---|---|
| 4 | 00:00:29,10 | 00:00:31,22 | 2.48 | yo acompañaba siempre<br>a mi madre. | 16 | 11,3 |
| 5 | 00:00:33,14 | 00:00:37,16 | 4.08 | Pero no iba delante de ella, sino<br>detrás, porque era muy tímida. | 16 | 13,0 |
| 6 | 00:00:38,16 | 00:00:42,18 | 3.08 | Así que un día cogimos<br>aquel tren con destino a Roma. | 16 | 14,0 |
| 7 | 00:00:47,00 | 00:00:50,15 | 3.6 | Cogimos directamente el tren<br>que iba a Cinecittà. | 16 | 11,7 |
| 8 | 00:00:50,15 | 00:00:52,08 | 1.72 | creo que era azul. | 16 | 9,7 |
| 9 | 00:00:54,09 | 00:00:56,16 | 2.28 | Estaban buscando figurantes<br>para Quo Vadis. | 16 | 16,7 |
| 10 | 00:01:02,04 | 00:01:05,10 | 3.24 | Hicimos cola durante tres horas, | 16 | 8,6 |
| 11 | 00:01:05,10 | 00:01:07,21 | 2.44 | y cuando llegamos<br>delante del director, | 16 | 13,9 |

Editor de subtítulos

Así que un día cogimos
aquel tren con destino a Roma.

Velocidad (carac/seg): 13,96

INSERCION MANUAL

Tini: 00:00:38 · Duración: 3,08 · Altura: 16 · ALL · □ Fijar

Justificado: ○ Izq. ● Centro ○ Der.

Así que un día cogimos
aquel tren con destino a Roma.

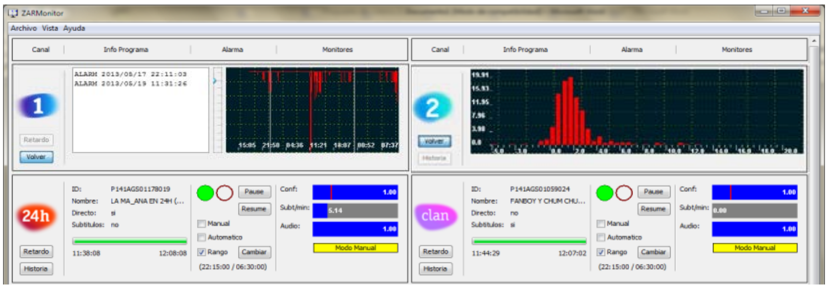
00:00:39 / 00:04:02

**Fig. 2.** ZARCaptions post-editior view.

## 4 ZARMonitor

ZARMonitor is a tool for monitoring subtitling channels of a television network. During the broadcast of a program, human errors can force the delivering of subtitles to a channel different than the one it should be. Again, text-audio synchronization can give an elegant solution for the detection of these errors. Taking audio clips and subtitles that are broadcasted in real time, a process of resynchronization is performed. The application receives real-time audio from all the channels to be monitored and broadcast subtitles and generates confidence measures on the adequacy of text to audio. If correspondence is not found for a predefined number of subtitles, the system triggers an alarm to automatically stop the subtitling, or alternativelly sends an alert to the user.

The graphic controller allows users monitoring of the status of the different broadcast channels. For each channel, we obtain information from the current program, the state of the audio and subtitles reception as well as the correspondence degree between them. We can also display a histogram with the delay of broadcast subtitles and historical information about the alarms generated in this channel (figure 3).



ALARM 2013/05/17 22:11:03
ALARM 2013/05/19 11:31:26

ID: P141AGS01178019
Nombre: LA MA_ANA EN 24H (...
Directo: si
Subtitulos: no
11:38:08 12:08:08
□ Manual
□ Automatico
☑ Rango
(22:15:00 / 06:30:00)
Conf: 1.00
Subt/min: 5.14
Audio: 1.00

ID: P141AGS01059024
Nombre: FANBOY Y CHUM CHU...
Directo: no
Subtitulos: si
11:44:29 12:07:02
□ Manual
□ Automatico
☑ Rango
(22:15:00 / 06:30:00)
Conf: 1.00
Subt/min: 0.00
Audio: 1.00

**Fig. 3.** Monitoring system graphics controller

## 5 Conclusion

In this work, we have presented a suite of programs that exploiting automatic speech recognition can definitely help to improve productivity, efficiency and quality of the subtitling task. These tools allow subtitling live broadcast news programs, previously off-line recorded programs and also monitoring broadcast subtitles quality. The presented systems are currently deployed in Aragón TV since 2008 and TVE since 2010.

## References

1. Cardinal, C. Chapdelaine, M. Comeau, P. Ouellet, F. Osterrath, “Computer-assisted closed-captioning of live TV broadcasts in French”, in Proceedings Interspeech 2006, Pittsburgh, USA, 2006.

2. J. Brousseau, J.F. Beaumont, G. Boulianne, P. Cardinal, C. Chapdelaine, M. Comeau, F. Osterrath and P. Ouellet, “Automated closed-captioning of live TV broadcast news in French”, in Proceedings of Eurospeech 2003, Geneva, Switzerland, 2003.

3. A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, “Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Program”, IEEE Transactions on Broadcasting, Vol. 46, No. 3, September 2000.

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. Ortega, A., Garcia, J.E., Miguel, A., and Lleida, E., "Real-time live broadcast news subtitling system for Spanish". In Proc. INTERSPEECH, 2009, pp.2095-2098.

6. H. Meinedo, M. Viveiros, J. Paulo and S. Neto, “Evaluation of a Live Broadcast News Subtitling System for Portuguese”, in Proc of Interspeech 2008, Brisbane, Australia, 2008.

7. AENOR (2003). Subtitling for deaf and hard-of-hearing people. Subtitling by teletext. UNE 153010:2003 v.8 (2010). Madrid: AENOR.

# The *Percepción Smart Campus* system

Carlos-D. Martínez-Hinarejos[1], Emilio Granell[1], David Rambla[2], Andrea Calia[2], Alejandro Luján[2], Guillermo Amat[3], Ángel Ramos[4], José-Miguel Benedí[1], and Alberto Sanchis[5]

[1] Pattern Recognition and Human Language Technology Research Center - Universitat Politècnica de València - Camino de Vera, s/n, 46022, Valencia, Spain
[2] Institute of New Imaging Technologies, Universitat Jaume I, Avda. Vicente Sos Baynat S/N, Castellón, Spain
[3] FHC25, Roger de Lauria 19 5-B, Valencia, Spain
[4] FHC25, Calle Rozabella, 4, Las Rozas, Madrid, Spain
[5] MLLP group, DSIC, Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain

**Abstract.** This paper presents the capabilities of the *Smart Campus* system developed during the *Percepción* project. The *Smart Campus* system is able to locate the user of the application in a limited environment, including indoor location. The system is able to show routes and data (using virtual reality) on the different elements of the environment. Speech queries could be used to locate places and get routes and information on that places.

## 1 Overview of the system

The *Smart Campus* system, developed during the *Percepción* project, was defined to take advantage of the capabilities of mobile devices. Due to the nearly-universal of mobile devices such as smartphones, tablets, an so on, they can be used as a tool for getting information in a defined environment and locate interesting points. Moreover, the use of mobile devices could improve the productivity of the society by giving the users information that allow them save time and money, e.g., the fastest route to arrive to a given place, or the availability of a given resource (a book in the library, a free parking space, etc.).

These capabilities are extensively used in the concept of *Smart City*, i.e., a populated area that provides its inhabitants and managers with the infrastructure that allows them to obtain the maximum throughput of the available resources. In the *Percepción* project, the application is reduced to the environment of a university campus. Two systems were implemented, one for a medium-size university - *Universitat Jaume I* (UJI) of *Castelló* - and other for a small-size university - *Centro Universitario de Tecnología y Artes Digitales* (U-TAD). The UJI system comprises a total of 1997 relevant locations that can be required by the user. U-TAD system has a total of 234 locations.

The system would allow the user to employ its mobile device for the following activities:

- Obtain the current location, including indoor locations

- Show, by using Augmented Reality (AR), a route to arrive to a location.
- Show, by using AR, information on people (via *QR* codes), buildings, or other entities.
- Ask for a location by using speech.

In this system description, we will focus on the *SmartUJI Project*. This project consists of a smart-maps technology which integrates all the University data into a spatial information system. The main project's objective is to improve the monitoring and management of the university resources, and provide access to all of the university's data in a unified and homogeneous way as established in the *INSPIRE Directive of the European Parliament and of the Council* (2007/2/EC) [14, 1]. The platform is accessed through a web based application[6] and a set of integrated Android applications.

The following sections describe the main features of the different modules of the system: geographical location (Section 2), augmented reality based on geolocation and markers (Sections 3 and 4), and speech interface (Section 5). Section 6 describes possible features to be included in the systems in the future.

## 2 Positioning system

The *SmartUJI* Android applications use the positioning ability of mobile devices to provide the user location. Although the inclusion of *A-GPS* in Smartphones has solved the problem of outdoor positioning, performing indoor location is still an interesting research topic since *GPS* has no coverage inside buildings. Several technologies are being used to deal with the indoor location problem: Radio-Frequency Identification (RFID) [12], Bluetooth [7], Wireless Local Area Network [9, 10], ZigBee [11], Ultrasonic [4], Variations on magnetic field [3], Dead reckoning [5], or LED lights [6], among others. Nowadays, WiFi and magnetic field are two of the most promising technologies for indoor location due to the proliferation of *Smartphones* and the embedded WiFi connectivity and magnetic sensory system. In particular, a WiFi-based fingerprint *Indoor Positioning System* (*IPS*) based on *k-Nearest Neighbor* [2] has been integrated into the *SmartUJI platform* to support the *SmartUJI* applications.

The IPS was deployed as a new service in the *SmartUJI platform*. The mobile device has only to send the current fingerprint (a list of perceived WLAN intensities) and the service returns the current indoor location. Localisation errors are raised when the location cannot be estimated. An example of full location is shown in Figure 1, where the black arrow icon shows an estimation of the user's position (longitude, altitude and floor) inside an academic laboratory.

## 3 SmartUJI AR

*SmartUJI AR* is an augmented reality application that shows geographically located content using the *MetaioSDK*[7]. It allows the user to view the gathered data in an immersive and intuitive environment. As shown in Figure 2, it uses the camera to show which facilities or points of interest (*POI*) are near the user.

[6] http://smart.uji.es

[7] http://www.metaio.com/sdk/

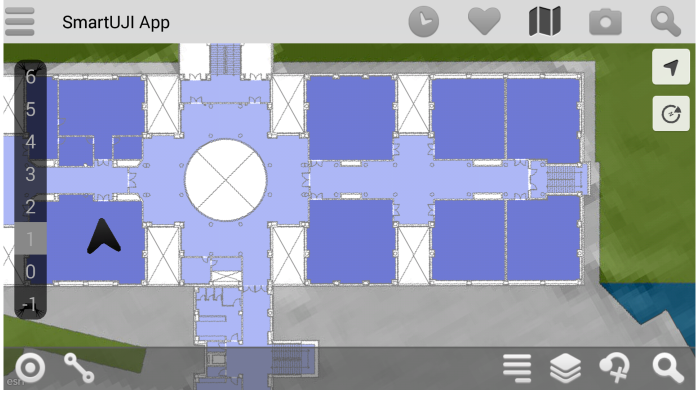


**Fig. 1.** Screenshot of the full positioning in a mobile device. The black arrow stands for the position where the user is located, this estimation is provided by the indoor positioning system.

*SmartUJI AR* collects all the data from the services, databases and geoprocesses provided by the *SmartUJI platform*. Then, these data are processed and properly showed to the user. The generated information is superimposed on the rear camera image in real-time. An appropriate icon is used to represent a certain facility or building in augmented reality. E.g., the *Library* (*Biblioteca*) is represented with an orange-white book icon in Figure 2. Then, the visible POI's are displayed using the user's orientation and current position via GPS or IPS (see Section 2). With this hybrid positioning system, GPS and IPS, it is possible to obtain the user's location anywhere inside the campus.

*SmartUJI AR* has a layer selector to choose what kind of POIs are going to be displayed. The main categories are: *Buildings*, *Restoration*, *Transports*, *Containers*, *Labo-*

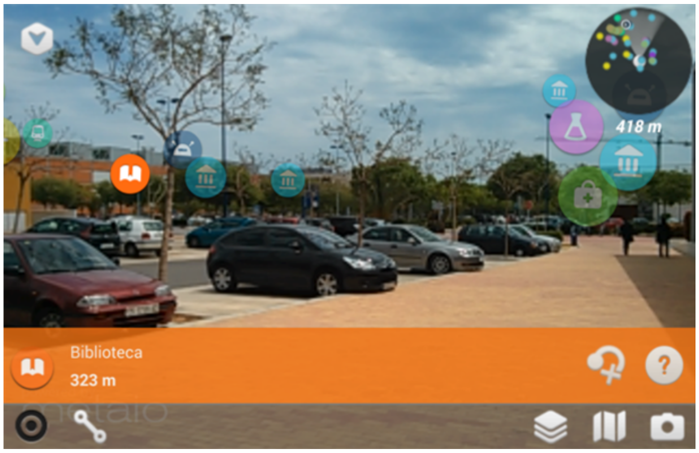


**Fig. 2.** Screenshot of the augmented reality application. It shows some facilities and the Library (*Biblioteca*) location is highlighted.

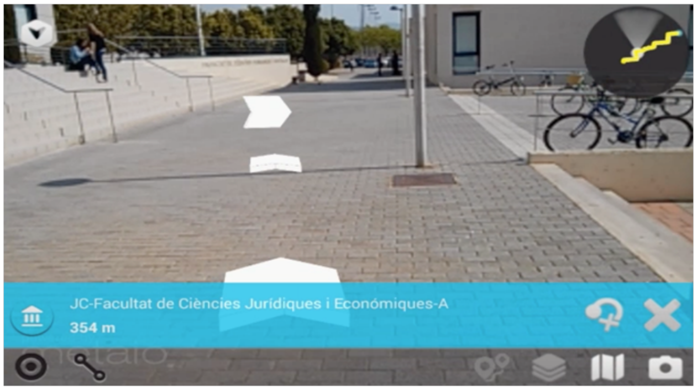


**Fig. 3.** Screenshot of the augmented reality application. It shows the path that the user should follow to reach its destination with white arrows.

*ratories*, and *Information Points*. Depending on the chosen categories, a different range of POI images will be displayed on the screen. Another functionality is the 2D radar in where all the POIs and the current vision field are represented. The user's vision fields moves at the same time as the user changes its orientation and position, so all the POIs that are inside the vision field are the same ones displayed on the screen.

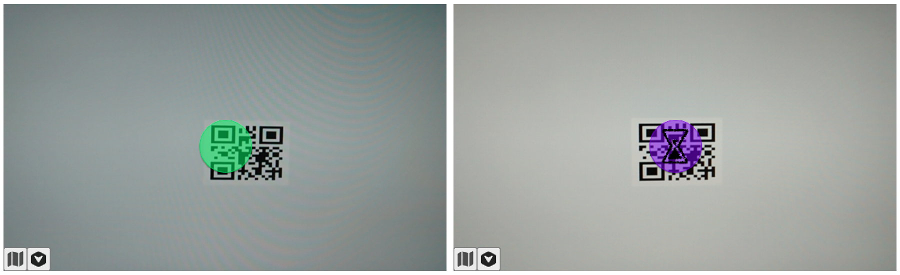
The navigation functionality uses the *SmartUJI platform* to calculates a route between the user's current location and a POI. The obtained route is placed over the camera image as an 3D-arrows route. The user just has to follow the arrows to reach its destination as shown in Figure 3.

## 4 Marker based AR

This application is the piece of *SmartUJI* that focuses on providing useful information about faculties to the user. This information is provided through a navigable augmented reality interface. To support the AR behaviour, *Quick Response* (*QR*) codes have been used to work as a beacon. Furthermore, our approach relies on *QR* codes as AR markers as they can be also used to codify and store some information. In addition, the *QR* codes may contribute to geographical location whenever that information is read.

When the application starts, the device camera is opened and a simple interface is shown to the user. As can be seen in Figure 4(a), a target point in green indicates the user where to aim the camera to help the application find the marker and read it. Once the *QR* code is located and the information that was stored in it is decoded, a request is sent to the server in order to get the GUI description along with the matching faculty data. Meanwhile, a loading icon is displayed to show that work is in progress (Figure 4(b)).

The response from the server describes the GUI that the application has to parse and render. A main interface with navigation options and representing information about the faculty is what the server sends in first place, to see the rendered scene refer to Figure 5. Through this interface, the user is able to navigate to other levels of detail.



(a) (b)

**Fig. 4.** Marker based Augmented Reality examples.(a) Target screen, showed on start-up, (b) Loading screen, waiting for data.

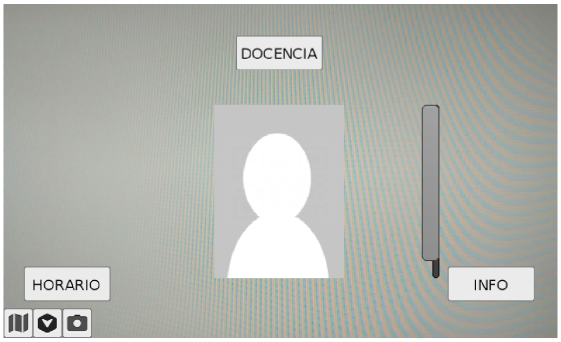
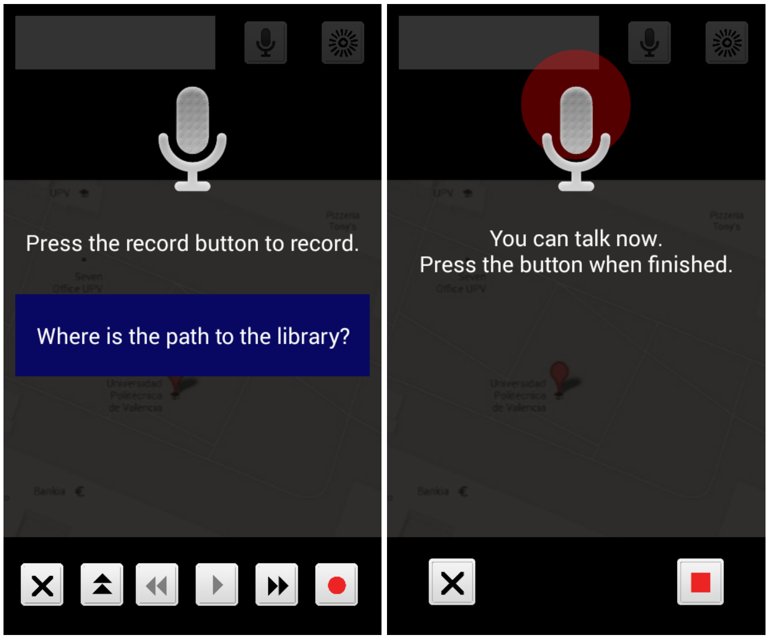


**Fig. 5.** Main screen, navigable menu, where users can get information about a university member: teaching hours and tutorials (HORARIO), courses (DOCENCIA), and more info (INFO).

## 5 Speech interface

The ASR system is used to locate points of interest in the university campus. The system is composed of a client application, a speech recogniser, a semantic analyser, a dialogue manager, an adaptation manager, a data manager and a database, with the following features:

- **Client application**: developed for Android 2.1 and higher versions; it is composed of two speech-related activities, one for speaker adaptation and the other for speech recognition.
- **Speech recogniser**: it is a version of the iATROS system [8] that supports TCP sockets audio input and speaker adaptation.
- **Semantic analyser**: FreeLing [13] was used.
- **Dialogue manager**: determines the current state of the dialogue; dialogue space is defined by a finite-state automaton; each state has associated a different language model for speech recognition.
- **Adaptation manager**: manages the speaker adaptation (generation of adaptation sentences and speaker adaptation based on the acquired audio signal).



Press the record button to record.

Where is the path to the library?

You can talk now.
Press the button when finished.

(a) Speaker adaptation activity. (b) Speech acquisition activity.

**Fig. 6.** Speech interfaces.

- **Data manager**: connects clients with speech recogniser, provides system responses, and updates semantic models according to the user queries and the database.
- **Database**: the same than for the geographical part.

The system allows the speech recognition in three languages (Spanish, Valencian, and English). The language to be decoded depends on the client device configuration. Acoustic, language, lexical, and semantic models were generated for each language.

The speech interface is launched from the main client application by touching the proper icon. When launched, detects if the device/speaker is registered for the selected language. In case it is not registered, launches the speaker adaptation activity (Figure 6(a)). This activity shows the user the text to be uttered and allows to send the acquired audio signal when all the sentences were completed.

In any case, the speech activity is launched (Figure 6(b)). This activity starts with an initial dialogue state (i.e., no data was present and all possible queries are admissible). The activity follows a push-to-talk protocol that allows the user a complete control on the speech signal that is recorded. After finishing the recording, the audio signal is sent to the server and it is recognised and processed against the database. Three different situation may appear from the user perspective:

1. An only item was retrieved from the database: the corresponding identifier is returned to the client and the speech activity sends it to the geographical manager.

2. A small number (5 or less) of items was retrieved from the database: a list with the items is generated and sent to the client, that shows it and awaits for user selection; if the "None of the above" option is selected, it returns to speech acquisition.
3. A large number (more than 5) of items was retrieved from the database: returns to speech acquisition, showing an informative text (based on the new dialogue state) and the last recognition result.

In any case, internal dialogue state is updated according to the data the user gave and the database query results.

In case any error appears (server not available, sentence not recognised, etc.), the control is returned to the main client application.

## 6 Discussion and future directions

In this article we presented the complete *Percepción Smart Campus* system for the *SmartUJI* environment. The described system complements with the other Android applications of *SmartUJI*. For example, from anywhere in the application is possible to launch the other parts. Furthermore, this application can enhance the others parts providing a precise localisation of the user. This can improve the localisation algorithm because we know the position of the user based on the known position of the *QR* code the user is scanning. Also, the application integrates with the underlying operating system in various ways.

It is worth mentioning that all the applications, functionalities and modules have been designed to interact between them, to be fully interoperable within the *SmartUJI platform*, and to enhance the user's experience. Also, the applications are integrated with the underlying operating system in various ways. E.g., *SmartUJI applications* are able to call a university's member through the operating system's dialler application.

However, the system could benefit from other data sources that are not integrated at this moment in the whole application. For example, user identification can be used to know features on the user that could enhance the interaction in both location and speech recognition: if a student is inscribed in a given center or subject, it will be more usual for s/he that her/his location is on a given part of the campus, or that the uttered speech is on locations related to the user situation (e.g., asking for a specific classroom could be centered to the classrooms of the center where the student is inscribed). There are future plans to incorporate this data on the system and improve interaction, apart from using data pertaining to systems different to the *SmartUJI* system (e.g., data on social networks that allow to locate the user or knowing her/his preferences).

## References

1. Benedito-Bordonau, M., Gargallo, D., Avariento, J., Sanchis, A., Gould, M., Huerta, J.: Uji smart campus: Un ejemplo de integración de recursos en la universitat jaume i de castelló. In: Actas de la IV Jornadas Ibéricas de Infraestructuras de Datos Espaciales (2013)
2. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13(1), 21–27 (Sep 1967), `http://dx.doi.org/10.1109/TIT.1967.1053964`

3. Guo, Y., Chen, Y., Liu, J.: Indoor location estimation based on local magnetic field via hybrid learning. In: Extreme Learning Machines 2013: Algorithms and Applications, Adaptation, Learning, and Optimization, vol. 16, pp. 189–207. Springer International Publishing (2014)
4. Ijaz, F., Yang, H.K., Ahmad, A., Lee, C.: Indoor positioning: A review of indoor ultrasonic positioning systems. In: Advanced Communication Technology (ICACT), 2013 15th International Conference on. pp. 1146–1150 (Jan 2013)
5. Jin, Y., Toh, H.S., Soh, W.S., Wong, W.C.: A robust dead-reckoning pedestrian tracking system with low cost sensors. In: Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on. pp. 222–230. IEEE (2011)
6. Kuo, Y.S., Pannuto, P., Dutta, P.: Demo — luxapose: Indoor positioning with mobile phones and visible light. In: The 20th Annual International Conference on Mobile Computing and Networking. MobiCom '14 (September 2014)
7. Li, H.: Low-cost 3d bluetooth indoor positioning with least square. Wireless Personal Communications pp. 1–14 (2014), `http://dx.doi.org/10.1007/s11277-014-1820-1`
8. Luján-Mares, M., Tamarit, V., Alabau, V., Martínez-Hinarejos, C.D., i Gadea, M.P., Sanchis, A., Toselli, A.H.: iatros: A speech and handwritting recognition system. In: VJTH'2008. pp. 75–78 (2008)
9. Machaj, J., Brida, P., Pich, R.: Rank based fingerprinting algorithm for indoor positioning. In: Proceedings of the 2nd the International Conference on Indoor Positioning and Indoor Navigation (IPIN'2011). pp. 1–6 (2011)
10. Marques, N., Meneses, F., Moreira, A.: Combining similarity functions and majority rules for multi-building, multi-floor, wifi positioning. In: Proceedings of the 3th the International Conference on Indoor Positioning and Indoor Navigation (IPIN'2012) (2012)
11. Martí, J.V., Sales, J., Marín, R., Jiménez-Ruiz, E.: Localization of mobile sensors and actuators for intervention in low-visibility conditions: The zigbee fingerprinting approach. International Journal of Distributed Sensor Networks (2012)
12. Montaser, A., Moselhi, O.: RFID indoor location identification for construction projects. Automation in Construction 39(0), 167 – 179 (2014), `http://www.sciencedirect.com/science/article/pii/S092658051300109X`
13. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Chair), N.C.C., Choukri, K., Declerck, T., Doan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)
14. Sanchis, A., Arnal, A., Molina, W., Sanchis, V., Daz, L., Huerta, J., Gould, M.: viscauji: campus inteligente como ide local. In: Actas de la III Jornadas Ibéricas de Infraestructuras de Datos Espaciales (2012)

# Web Server for Evaluation of Voice Quality: e-VOICE

Josue Cabrera[1], Jesús B. Alonso[1], José De León[2], Miguel A. Ferrer[1], Patricia Henriquez[1], Carlos M. Travieso[1]

[1]Institute for Technological Development and Innovation in Communications
University of Las Palmas de Gran Canaria
Campus Universitario de Tafira -35017,
Las Palmas de Gran Canaria, Spain

[2]Department of Otorhinolaryngology General Hospital of Gran Canaria "Dr. Negrín"
Barranco de la Ballena, 35010
Las Palmas de Gran Canaria, Spain

jalonso@dsc.ulpgc.es

**Abstract.** The need to estimate the acoustic information that comes from the patient has progressively developed methods of speech signal analysis . These methods range from the subjective estimate using certain protocols (such as GRAB, …) which allow homogenize the perceptual features of the voice, to objective estimate the speech signal based on a set of features, allowing the documentation of the speech signal in a clear and effective way. In this work, a web server (http://evoice.ulpgc.es) for the objective evaluation of voice signal using the recording of a sustained vowel is presented. This software implements a protocol based on quantifying four aspects that characterize the quality of the voice with regard to the normal pattern, obtaining a correlation between the numerical value of the measure and the physical phenomenon quantified. This website allows the remote use of the software. This implies a significant improvement in efficiency and potential for clinical application, such as e-health or preventive medicine, compared to other existing tools that develop protocols of greater difficulty in its practical application.

***Keywords***— Speech Processing, Voice Quality, Biomedical Signal Processing, e-Health, Web Server

## 1 Introduction

The voice is the main means of communication between people, it is part of our body image, and in many cases is an important tool of work 1. Between 5 and 7% of the population in industrialized countries suffers from some disorder of the vocal cords that requires the attention of the specialist, this percentage increases to 22.5 % in groups such as teachers, singers, speakers and commercial that used the voice as an instrument of work 2, 3, 4, 5.

Techniques based on methods of visual inspection of the vocal cords have classically been used for the diagnosis of voice disorders. However, these techniques have certain drawbacks, since visual exploration techniques, such as the need for an adequate training and the discomfort that this implies in many occasions for the patient. At the same time, these diagnostic methods provide considerable data on the morphological aspect of the larynx and the disease processes, but very few on the estimate of the acoustic voice, at the same time that have to be carried out exclusively by professionals in the medical field.

Currently, adequate human resources and material means are not always available to perform this estimate. It is of interest to have objective evaluation of the voice quality. So far, some methods have been developed (which are more or less standardized) based on the subjective evaluation-perceptual voice quality, such as GRABS 6, 7, RHB 8, and others 9, 10, 11.

Therefore, there is a need to objectify the voice signal based on the estimation of objective measurements to quantify numerically and automatically the voice quality from an audio recording. These techniques improve the quality of care in the medical field. To cite some examples, it is useful to doctors to perform the first screening in primary care centers, or to medical specialists in otolaryngology to document objectively the evolution of a certain treatment of a patient and for documenting the evolution of a surgical intervention.

The web server application of these techniques is useful in situations where the speaker is not present as is the case of telemedicine, rural or isolated areas, islands etc. It also opens the door to medical professionals whose professional work is oriented to the rehabilitation of patients suffering from problems in the voice (speech therapists, phonetics) to obtain tools that will facilitate their work professionally, and this is appealing to the professional for which his voice is his work tool (teachers, speakers, singers, commercial, etc.), because it enables them to have a tool to estimate of way simply if you suffer any alteration in the larynx (disease in the vocal cords and the voice production system) from a voice recording.

In this work we present a web server for the evaluation of the quality of the voice of a speaker objectively from the phonation of a voiced sound: the vowel "a".

## 2 Clinical study

With the aim of having non-invasive techniques, objective and quantifiable evaluation of the voice quality, we develop a study of evaluation system acoustics phonic apparatus for use in clinical environments coordinate by a Department of Otorhinolaryngology.

In this study of the voice quality, it indicates the need to measure four characteristics of the voice: stability in the voice, rhythm in the beating, problems on the glottal closure and irregularities in the masses 12, 13, 14. Each of the different characteristics of the voice is quantified by a single measure. Table I shows the measures used to quantify each of the four characteristics.

The four physical phenomena identified that allow you to perform an objective measurement of the quality of the voice are:

- *Stability in the voice*: This phenomenon characterizes the flow of air that is exhaled by the lungs and which gives rise to the vibratory motion of the vocal cords (movements of opening and closing).

- *Rhythm in the beating*: The harmonic structure of the voice signal is determined by the pace and force of the beating (the impact between the vocal cords). The existence of an irregular rhythm reveals the presence of certain pathologies.

- *Problems on the glottal closure*: This phenomenon characterizes the contact of both cords during phonation. In the case of absence of a proper closure of the same identifies the presence of noise in the voice signal.

- *Irregularities in the masses*: To perform a correct beating of the vocal cords it is necessary that both vocal cords carried out synchronously the different phases of the vibratory motion necessary that gives rise to the successive opening and closing of the vocal cords. This phenomenon quantifies the misalignment of the different phases of the movement between vocal cords when these are irregularities.

**Table 1.** Characteristics of the voice measures

| Characteristic | Measure |
|---|---|
| Stability in the voice | Amplitude Perturbation Quotient (*APQ*) 16, 17, 18, 19 |
| Rhythm in the beating | Average value of the first peak Cepstral of the cepstrum derived spatially 13 |
| Problems on the glottal closure | Average ratio subharmonic harmonic 15 |
| Irregularities in the masses | Average of the lower area of the integrated Bicoherence index 12 |

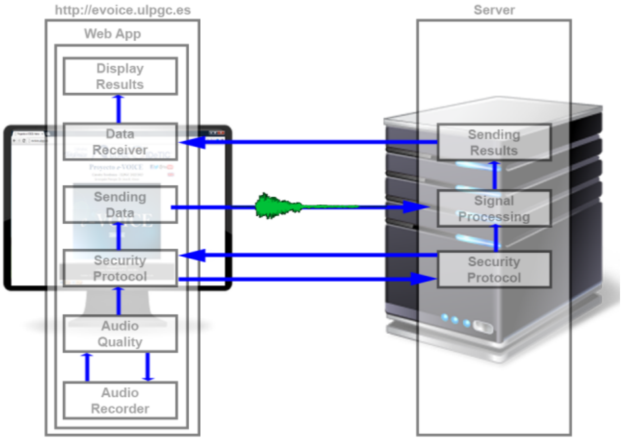


**Fig. 1.** Basic diagram of the web architecture

## 3 Web Arquitecture

We have developed the web server for the evaluation of the voice quality: e-VOICE, available at the URL http://evoice.ulpgc.es, that implements the clinical study described above, providing an objective method reliably, and easily, accessible anywhere in the world and at any time via a browser and an internet connection.

The web server uses standard technologies and programming languages, frequently used in computing and communications, such as HTTP, TCP/IP, HTML, CSS, Flash and Java.

The web architecture, shown in Fig. 1, is simple, consists of a web application built into a web site and a processing server. The web application performs the tasks of: voice recording (Fig. 2), check the voice quality, secure connection to the server, sending the voice signal, reception and representation of the results of the signal processing (Fig. 3). The server is responsible for verifying the secure connection, receive and process the voice signal, estimating the validated measures by the clinical study and sending the results to the web application.

This simple but effective architecture has the advantage of allowing multiple connections, from any location (internet), with tools of extended use (desktop browsers) using a thin client system compared to the traditional client-server system that need a heavy client application installed on your computer.

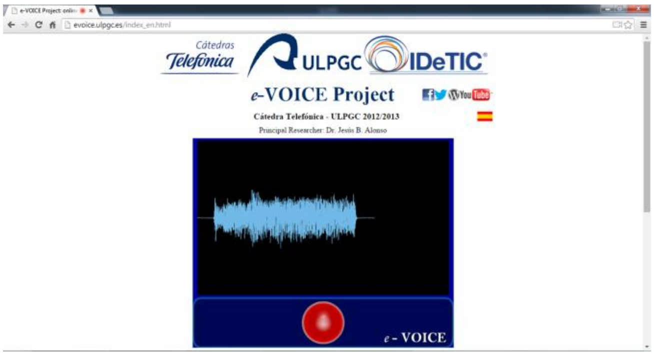

evoice.ulpgc.es/index_en.html
Cátedras Telefónica
ULPGC
IDeTIC
e-VOICE Project
Cátedra Telefónica - ULPGC 2012/2013
Principal Researcher: Dr. Jesús B. Alonso
e - VOICE

**Fig. 2.** Voice **recording**

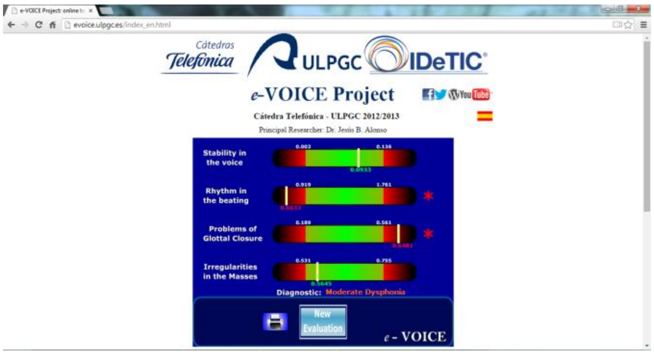

evoice.ulpgc.es/index_en.html
Cátedras Telefónica
ULPGC
IDeTIC
e-VOICE Project
Cátedra Telefónica - ULPGC 2012/2013
Principal Researcher: Dr. Jesús B. Alonso
Stability in the voice
0.002
0.136
0.0933
Rhythm in the beating
0.919
1.761
Problems of Glottal Closure
0.189
0.561
Irregularities in the Masses
0.531
0.755
0.5645
Diagnostic: Moderate Dysphonia
New Evaluation
e - VOICE

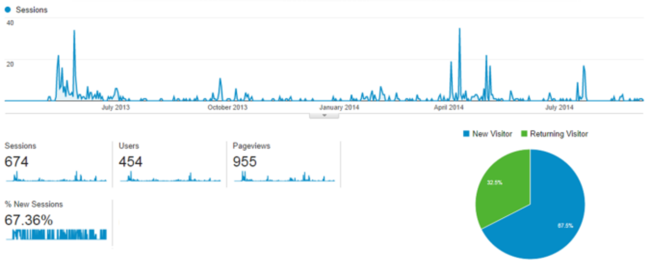
**Fig. 3.** Representation of results of an abnormal quality voice

# 4 Results

The web analytics tool integrated in the web site http://evoice.ulpgc.es enable us to obtain measures about the accessibility and possibilities for clinical application, such as e-health or preventive medicine, to have the web server for evaluation of voice quality described in this paper.

The statistical data provided by the web analytics, shown in Figure 4, reveal that in a period of approximately one year of operation of the web environment have been made to 674 sessions by 454 users with a percentage of 67.36 % of new visitors and nearly 1000 page views.

In terms of accessibility, the demographic data (Fig. 5) show connections from 85 cities in 29 countries on 5 continents, such as Spain, India, United Kingdom, Hong Kong, Australia, Cuba, Morocco,...



Sessions 674

Users 454

Pageviews 955

% New Sessions 67.36%

**Fig. 4.** Statistical data by web analytics tool

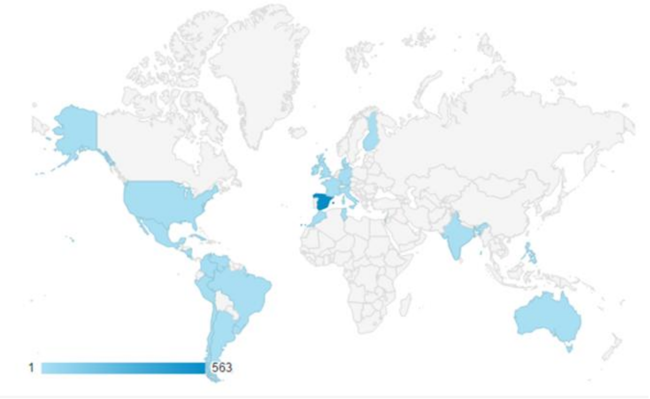


**Fig. 5.** Demographic data by web analytics tool

## 5 Discussion

It is possible to identify a particular group of computer applications that allow you to record voice signal and subsequently allow you to perform a qualitative evaluation both by means of different representations of the speech signal as quantitatively by means of a series of measures.

Some applications such as Soundscape 20, PRAAT 2122 and CSRE and have in common be designed for the analysis of sound signals, with a thorough analysis of the signal, allowing its recording, editing, and calculation of statistics of values taken from the signal analysis as: spectrograms, fundamental frequency of the signal (pitch), disturbance (jitter), the disturbance in signal amplitude (shimmer), spectral analysis of the signal by FFT and LPC, and statistics on some characteristics of the signal, such as the percentage of unvoiced and voiced sounds.

Other applications also include utilities computing oriented education and syntactic training. A good example is the system developed by Speech Technology Research (STR) Ltd., which consists of a series of tools for capturing and analysis of the speech signal. Among all of its products, it is worth highlighting the Computerized Speech Lab (CSL) 23, developed jointly with Kay Elemetrics Corporation. CSL is the reference tool in areas ranging from forensic sound until the acoustic survey of pathologies of the voice. This is due to its high performance, use of high-end hardware and to the broad base of accompanying data. The CSL provides information of multiple characteristics of the voice, by removing temporary parameters of qualities of energy and of the fundamental frequency. The evolution of these parameters can be studied in real time while recording of the voice of the speaker.

It is also possible to find computer applications aimed at patients who have disorders of hearing, speech and language with in the case of SpeechViewer 24, which is especially useful for people who want to modify their inflection, pronunciation and vocal quality. The program allows you to know and to exercise the features of the voice (tone, intensity, duration, and loudness) through a visual and auditory feedback of the verbal productions.

There are more complete computer systems that include all the aspect described-cough as those developed by the UCL division Speech Hearing and Phonetic Science 25, of the University College London, which allows a simple but comprehensive study of the voice signal. Through these tools it is possible to make the recording of voice signals, the qualitative study of the signals by means of temporary representations, spectral (spectrum and spectrogram) and cepstrals. It is possible to identify the formants and estimate the vibrations frequencies of the vocal cords.

Finally there are complex systems, such as Dr. Speech 26  developed by Tiger Electronic, that add another type of signals electroglottography and laryngoscope images. Dr Speech is oriented to the analysis and rehabilitation of pathological voices. The software has been developed primarily to be used in professional environments, particularly in the health sector. Dr. Speech gives the possibility of making an ongoing study of the evolution of the pathological voice. It contains a simple working environment so that the specialist can annotate each of its impressions of the patient, leaving its history stored in a database of patients.

All these systems have in common present a clear disadvantage of accessibility, because they need a local installation on a computer, in addition to the estimate of the quality of the voice for qualitative or quantitative techniques that do not guarantee the uniqueness of the ratings obtained from different evaluators.

This paper proposes an alternative to these systems through a web server that allows through internet to objectively evaluate the clinical voice quality of a speaker from the recording of a sustained phonation of a vowel. This system measures, objectively, four aspects of the voice quality: stability in the voice, rhythm in the beating, problems on the glottal closure and irregularities in the masses. This assigns a value to each of the different aspects that characterize the voice quality, where in addition have been identified normal ranges in a clinical study.

## 6 Conclussions

In this work we present, in the context of accessibility for the clinical evaluation of the voice quality from a recording, the different software tools that enable you to evaluate the voice quality by qualitative and quantitative method. However, these systems, whose task is not exempt from interpretation and the need for wide knowledge of the characteristics of the speech signal in the different domains of representation, are designed as desktop applications.

As an alternative, this paper presents a simple and robust web tool that lets you document the voice quality, quantifying and objectively automatically four physical phenomena that allow you to perform a measurement of the voice quality. As a result, it is possible to obtain a correlation between the numerical value of the measure and the physical phenomenon that quantifies, allowing detecting deviations from their margins of normality, where any voice of abnormal quality presents a deviation from the normal pattern of at least one of the four characteristics.

A prototype of this tool has been evaluated in a clinical study, obtaining as a result a high correlation between the presence of a laryngeal pathology and the labelling of at least one of the four quantifiable physical phenomena outside the range of normality. In the previous laboratory work 12, 13, 14 we have demonstrated that the tool is statistically significant in the discrimination between samples of healthy voices and with laryngeal pathologies, and with the clinical study we have demonstrated that the tool is clinically relevant in the evaluation and documentation of patients with laryngeal pathologies

The objective quantification of the quality of the voice track web allows you to perform functions of screening of laryngeal pathologies that can be useful in situations where the speaker is not present as is the case of telemedicine, as well as to assess treatment outcome of the same in the specialized environment.

## Acknowledgments

This work has been supported by the Telefónica Cathedra – University of Las Palmas de Gran Canaria Program 2012/2013.

## References

1. Vilkman E. Voice problems at work: a challenge for occupational safety and health arrangement. Folia Phoniat et Logop. 2000; 52, 120-125..
2. Jardim R., Barreto S. M., Assunção A. Á. Voice disorder: case definition and prevalence in teachers. Revista Brasileira de Epidemiologia. 2007; 10(4), 625-636.
3. Delcor NS, Araújo TM, Reis EJFB, Porto LA, Carvalho FM, Silva MO, Barbalho L, Andrade MA. Condições de trabalho e saúde dos professores da rede particular de ensino de Vitória da Conquista, Bahia, Brasil. Cad Saúde Pública. 2004; 20(1), 187-96.
4. Al-Saleem S., Al-Saleem M. Epidemiology of voice disorders among male school teachers in Khamis Mushayt city, Saudi Arabia. International Journal of Medical Science and Public Health. 2013; 2(2), 330-338.
5. Tobar L. N., Pérez M. C., Camacho, J. Patología benigna de laringe en docentes y cantantes. Patología. 2011; 39(1).
6. Hirano M.. Clinical Examination of Voice. New York, Springer-Verlag, 1981.
7. Dejonckere P. H.; Remacle M.; Fresnel-Elbaz E.; Woisard V.; Crevier-Buchman L.; Millet B.; Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. Revue de Laryngologie Otologie Rhinologie, 1996; 117 (2), 219-224.
8. The National Center for Voice and Speech. Available: www.ncvs.org
9. Hammarberg, B., Gauffin, J. Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects. Vocal Fold Physiology. 1995; 283-303.
10. Laver J.The Gift of Speech. Edinburgh University Press. 1991.
11. Zetterholm E. Auditory and Acoustic Analysis of Voice Quality Variations in Normal Voices. Proceedings of the XIVth International Congress of Phonetic Sciences ( ICPhS-99). 1999; 973-976.
12. Alonso J. B., Henríquez P. Speech-based Clinical Diagnostic Systems. Encyclopedia of Artificial Intelligence. IGI Global publication. 2009; 1439-1446.
13. Alonso J., De León J., Alonso I., Ferrer M.A.. Automatic Detection of Pathologies in the voice by HOS based Parameters, EURASIP Journal on Applied Signal processing. 2001, 4, 275-284.
14. Henriquez P., Alonso J.B., Ferrer M.A, Travieso C.M., Godino-Llorente J., Diaz-de-Maria F.. Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics. IEEE Transactions on Audio, Speech and Language Processing. 2009; 17 (6), 1186-1195.
15. Sun X.. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. IEEE International Conference on Acoustics, Speech, and Signal Processing. 2002; 1, 333-336.
16. Godino-Llorente J., On the selection of meaningful speech parameters used by a pathologic/non pathologic voice register classifier. Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99). 1999; 563-566.
17. Godino-Llorente J., Aguilera-Navarro S., Gomez-Vilda P.. Non supervised neural net applied to the detection of voice impairment. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00. 2000; 6, 3594-3597.
18. Godino-Llorente J.; Aguilera-Navarro S.; Gómez-Vilda P.. Automatic detection of voice impairments due to vocal misuse by means of gaussian mixture models. Procedings of the 23rd Annual EMBS International conference. 2001; 1723-1726.

19. Jo C., Kim K., Kim D., Wang S.. Screening of Pathological Voice from ARS using Neural Network. International Workshop on MAVEBA. 2001; 13-15.
20. GW Instruments, Inc., Available: http://www.gwinst.com/macsftwr/html/sos_summary.html.
21. Boersma P., Weenink D. Praat: doing phonetics by computer [Computer program]. 2013. Available: http://www.praat.org/.
22. Boersma P., Weenink D.. Praat, a system for doing phonetics by computer, version 3.4. Report, Institute of Phonetic Sciences of the University of Amsterdam.1996; 132, 1- 182.
23. KayPENTAX, Available: http://www.kayelemetrics.com
24. IBM Corporation, Available: ftp://service.boulder.ibm.com/sns/spv3/spv3supt.htm
25. UCL Speech Hearing and Phonetic Science, University College London, Available: http://www.phon.ucl.ac.uk/resource/software.php
26. Dr. Speech home page. Available: http://www.drspeech.com.

# Discriminative Methods for model optimization in Speaker Verification

Leibny Paola García-Perera[1], Juan Arturo Nolazco Flores[1]*, and Eduardo Lleida Solano[2]

[1] Computer Science Department, Tecnológico de Monterrey, Campus Monterrey
Monterrey Nuevo León, México
{paola.garcia,jnolazco}@itesm.mx
[2] Aragon Institute for Engineering Research (I3A) University of Zaragoza
Spain
lleida@unizar.es

**Abstract.** This thesis proposes a method to improve the current state-of-the-art of *Speaker Verification* (SV) systems by including a discriminative optimization. The success of traditional SV systems rely on how well the models represent the data. In this sense, the modeling is the essential part of SV. Conventional methods focus their efforts on maximizing the likelihood of those models to its corresponding data, and the error reduction comes as a consequence of the correct computation of the models. This work embraces the error as part of the optimization task. Following this idea, we explored two branches. In a first approach, we investigated the optimization of the training hyper-parameters to explicitly consider the tradeoff between false rejections and false acceptances. We believe that the enhancement of the parameters should not be limited to a single operating point, and that a more robust strategy is to optimize the parameters according to the maximization of the *area under the Receiver Operating Characteristic (ROC) curve*. In a second approach, we investigated the speech signal as a combination of attributes that contains information of the speaker, channel and noise. Conventional speaker verification systems train a single generic model for all cases. We propose a new methodology to *partition* the *data space* according to these attributes and discriminatively train separate models for each region, reducing the empirical error. Finally, we extended and merged our current methodologies to alleviate the effect of noise. The thesis can be found in: `http://www.vivolab.es/thesis.html`

**Keywords:** speaker verification, noise condition, ensemble approach, discriminative training, minimum verification error

## 1 Introduction

Over the last decade, advances in communication technology have leaded the research community efforts to focus more and more on secure and remote trans-

[*] Thanks to Cátedra de Seguridad de la Información, Tecnologico de Monterrey, Campus Monterrey.

actions over the networks [1]. User authentication has attracted major attention because of the emphasis on security issues [2]. The necessity of these systems to obtain unique information from users led to biometrics (human characteristics) [2]. From the variety of biometric signals that can be used for authentication, speech shows several advantages. For example, it is the most natural source of human communication; the acquisition of the signal can be performed using well-known non-sophisticated equipment and speech opens the possibility to perform remote transaction, among others [3].

In this research, we study SV, which main objective is to accept or reject a prospect speaker– with the lowest error rate– given a claimed identity and a speech signal. The SV schemes are based on statistical hypothesis testing theory for classification, which establishes the relationship between two hypotheses: a null hypothesis $H_s$ (accepts the speaker as legitimate) and the alternative hypothesis $H_{\bar{s}}$ (rejects him or her) [3]. Their relation is expressed as a ratio between likelihoods with respect to two models: target and imposter. The output ratio is then tagged as accepted or rejected user and later used as a score to perform a decision; the larger this score, the more "likely" to be a registered user.

Two types of errors commonly occur: *false acceptances* (FA) – incorrect decision due to accept a speaker who is not actually the target speaker, and *false rejections* (FR) – incorrect rejection of target speakers. Ideally, the probability of both types of errors must be zero; in practice, the two occur and are considered in the system design. A tradeoff between them is established so that both errors are reduced. The probability of false acceptance can be reduced at the cost of increasing false rejection. Thus, for any given system, the operating point can be manipulated to obtain a desired ratio between false acceptances and false rejections. The entire range of possible operating points is characterized mainly by the *Operating Receiver Characteristic* (ROC) curve [4] and *detection error tradeoff* (DET) curve [5].

This thesis shows how to lower these errors using a discriminative optimization approach from two different perspectives. The first one concentrates on the optimization of every operating point along the ROC curve, resulting on better performance in terms of known metrics, such as, equal error rate (EER) and minDCF (minimum Detection Cost Function). This curve provides a tool to measure system performance. It plots the relation between false positive rate vs true positive rate as shown in Figure 1. A single operating point on the ROC curve is not enough to describe the efficiency of the system. The area under the ROC curve is a more sensible measure of classifier performance. Second, we searched for more specific ways to describe the speaker space based on certain attribute. We built contiguous region models that characterize a specific attribute. The discriminative optimization enhances those models and improves the results of the traditional architectures. Both methods have their extension tackling the noisy condition case. Finally, merging both methodologies in an *integration modeling*, the system is capable to compute improved models that outperform the current state of the art.

## 2 Motivation and Scope

The success of traditional SV systems lies in computing the adequate models that can clearly classify a target speaker from an impostor. The more suited these models are to particular data in different scenarios or target speakers, the better results we can obtain [3, 6, 7]. The *maximum likelihood* (ML) approaches using generative modeling [3] and lately factor analysis [6, 7] represented a huge improvement in the systems performance. Generative modeling relies on maximizing the likelihood of a model to a current set of data. However, reduction of the *false acceptance* and *false rejection* errors is not taken as a primary objective but as a consequence of an accurate modeling of the target speakers.

Discriminative training approaches were first employed in speech recognition [8, 9], as an alternative solution to reduce the error. However, just a few studies are found in the bibliography [10–13] from the SV point of view.

One of the schemes that fits the SV theory is the minimum classification error approach. Its main goal is to minimize the empirical classification error regardless the distribution of the data [8, 14]. Consequently, we consider it to be the backbone of our research. The algorithm, by itself, focuses on the error reduction, which is the missing part in the SV framework. We broadened its capabilities and incorporated it as part of the optimization.

Finally, the challenge nowadays is to design systems robust enough to different kinds of scenarios (including environments, and noisy conditions). Hence, the motivation of this study is to deal with those scenarios and build a robust system based on discriminative modeling methods.

## 3 Thesis statement

We propose to include a discriminative model optimization as part of the current state-of-the-art systems to improve performance considering the error reduction as main objective. Two perspectives were explored separately and finally merged to obtain a more robust architecture.

In a first study, we examined a discriminative training paradigm that *explicitly* learns the model parameters to optimize the *entire* ROC curve. The solution we propose is to maximize the AUC; this naturally optimizes the performance at every operating point on the ROC curve.

In a second study, we propose an alternative approach to consider an effective decomposition of the speech signal space. Instead of addressing the problem with just a unified model (one model that can represent all possible data), a decomposition that partitions the signal space based on special attributes of the training data provides a better estimation. We propose to break down a general model into sub-region specific models. Each submodel is trained discriminatively for specificity purposes and empirical error reduction. The new optimized models are more robust and specific to the region they represent (a set of sub-models can represent, for example, channel, noise and speaker main attributes). Both techniques were examined also under noisy conditions, improving the baseline systems.

Finally, we propose a hierarchical architecture that incorporates both approaches. Data spaces (from general to specific) partition at each level and are discriminatively optimized using the AUC under the ROC curve. This approach is also naturally extended to noisy conditions scenarios.

## 4 Methodology

First, we explored the traditional modeling approaches in SV, resulting in baseline systems. Afterwards, we designed a discriminative optimization to be included as part of the current solutions: factor analysis and minimum verification error approaches. From this optimization, two branches emerged: the *AUC under the ROC curve optimization* and the *ensemble modeling*. At first, we modified the baseline methods, Joint Factor Analysis (JFA) and the Minimum Verification Error (MVE), to follow either the AUC under the ROC curve or the Ensemble approach. Then we merged both ideas to fulfill the integration goal. The algorithms at each stage work under clean and noise conditions obtaining improvements at every step.

## 5 Area under the ROC curve optimization

A way to address the problem of error reduction, is to optimize the AUC of the ROC curve. The computation of the AUC is clearly demonstrated by the Wilcoxon-Mann-Whitney (WMW) [15] statistic, which basic idea is to compute the probability of a classifier to score a random positive sample more times than random negative samples. Using the WMW statistic as a starting point, we formulate a method to reduce the AUC of the ROC curve for the SV task.

Consider a binary classifier that attempts to classify the data as belonging to a class $C$ or not. Let $\mathcal{H}$ and $\mathcal{W}$ be two sets of data belonging respectively to $C$ and $\bar{C}$ (*i.e.* not in $C$). The empirical AUC of a classifier that computes a score $\theta(\chi)$ to determine if any data instance $\chi$ belongs to $C$ is given by Equation 1 (based on the WMW theory).

$$\Upsilon(\Lambda) = 1.0 - \frac{\sum_{\chi \in \mathcal{H}} \sum_{\hat{\chi} \in \mathcal{W}} 1(\theta(\chi) > \theta(\hat{\chi}))}{|\mathcal{H}||\mathcal{W}|} \tag{1}$$

In the equation above we have written $\Upsilon(\Lambda)$ to be a function of $\Lambda = \Lambda_C \cup \Lambda_{\bar{C}}$, where $\Lambda_C$ and $\Lambda_{\bar{C}}$ are the parameters of any models associated with $C$ and $\bar{C}$.

We can optimize the AUC by employing the function given in Equation 1. Computed over the training set, Equation 1, is thus a new objective function to be minimized. To optimize it, a smooth, differentiable version of this objective function is needed. To do so, we replace the indicator functions $1(a > b)$ by a sigmoid function, $R(\theta(\chi), \theta(\hat{\chi}))$, following an approach that is commonly used in discriminative training methods, *e.g.* [14].

Thus, the modified AUC function to optimize is:

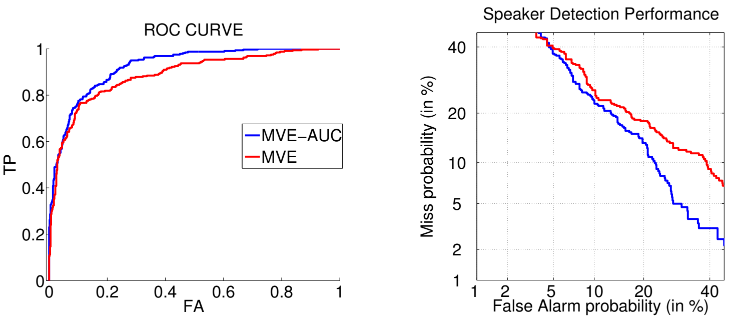
$$\Upsilon(\Lambda) = 1.0 - \frac{\sum_{\chi \in \mathcal{H}} \sum_{\hat{\chi} \in \mathcal{W}} R(\theta(\chi), \theta(\hat{\chi}))}{|\mathcal{H}||\mathcal{W}|}. \tag{2}$$

The modified AUC function of Equation 2 must be appropriately customized to the type of model being considered. It can then be optimized using the generalized probabilistic descendent (GPD) algorithm. Let $\mathbf{X}$ represent the complete set of training instances: $\mathbf{X} = \mathcal{H} \cup \mathcal{W}$. The GPD updates are performed according to the following:

$$\Lambda_{t+1} = \Lambda_t - \epsilon \nabla \Upsilon(\mathbf{X}, \Lambda), \tag{3}$$

$$\nabla \Upsilon(\mathbf{X}, \Lambda) = -\frac{1}{|\mathcal{H}||\mathcal{W}|} \sum_{\chi \in \mathcal{H}} \sum_{\hat{\chi} \in \mathcal{W}} \gamma R(1-R) \left[ \frac{\partial \theta(\chi)}{\partial \Lambda} - \frac{\partial \theta(\hat{\chi})}{\partial \Lambda} \right]. \tag{4}$$

In the above Equation $R$ is a short-hand notation for $R(\theta(\chi), \theta(\hat{\chi}))$, $\epsilon$ is a learning rate parameter. The above formalism can generally be used in all formulations of speaker verification with appropriate customization of the objective function that can be MVE or FA approaches. This method updates the model parameters iteratively for each learning step. The optimization is shown in Figure 1 (considering a MVE baseline system and a 10 iteration AUC optimization). We observe that by applying the method the AUC increases and the effect is also observed in the DET curve.



(a) ROC curve

(b) DET curve

**Fig. 1.** AUC Optimization: Score Distributions for 10dB SNR, babble Noise

## 6 Ensemble

The objective of this method is to compute specific and enhanced models that altogether represent the complete data space. The premise to follow is to separate the cluster attributes (classes) so that the error is minimized. The first challenge of this approach is to obtain suitable partitions $\Omega_1, \cdots, \Omega_P$ from the signal space (see Figure 2). Two options are possible to compute those partitions: *a)* blindly (unsupervised) or *b)* using some prior information (supervised).

### 6.1 Training the Ensemble Model

Corresponding to each of the partitions $\Omega_1, \cdots, \Omega_P$ we train a separate partition-specific background model. All background models are GMMs. We require each

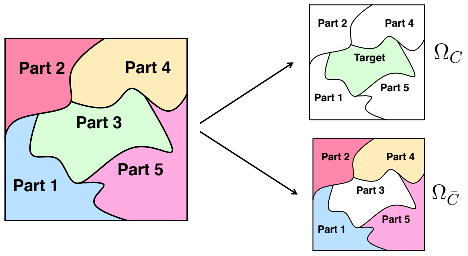


**Fig. 2.** Ensemble scheme

of the background models to be highly *specific* to the partition they represent, and not generalize to other partitions. In order to do so, we train all of them together using the following *discriminative* training procedure [16].

Let $\Lambda_C$ represent the model for a partition $\Omega_C$. Let $\chi_C$ represent all (training) recordings assigned to $\Omega_C$. For any partition $\Omega_C$, let

$$\Omega_{\bar{C}} = \bigcup_{\bar{C}' \neq C} \Omega_{\bar{C}'} \tag{5}$$

represent the *complement* of $\Omega_C$, *i.e.* the union of all partitions that are not $\Omega_C$.

Let $g(\chi; \Lambda_C) = \log P(\chi; \Lambda_C)$ represent the discriminant function (log-likelihood in this case) of any recording $\chi$ computed with the distribution for partition $\Omega_C$. We can now define $d(\chi, \Lambda_C)$, a misclassification measure for how likely it is that a data $\chi \in \chi_C$ from $\Omega_C$ will be misclassified as belonging to $\Omega_{\bar{C}}$ as

$$d(\chi, \Lambda_C) = -g(\chi; \Lambda_C) + G(\chi; \Lambda_{\bar{C}}), \tag{6}$$

$G(\chi, \Lambda_{\bar{C}})$ represents the combined score obtained from a partitions in $\Omega_{\bar{C}}$,

$$G(\chi, \Lambda_{\bar{C}}) = \log \left\{ \frac{1}{|\Omega_{\bar{C}}|} \sum_{C' : \Omega_{C'} \in \Omega_{\bar{C}}} \exp\left[\eta g(\chi, \Lambda_{C'})\right] \right\}^{\frac{1}{\eta}}, \tag{7}$$

where $|\Omega_{\bar{C}}|$ is the number of partitions included in $\Omega_{\bar{C}}$, and $\eta$ is a positive parameter. Now, we can define a new objective function for discriminative training of $\Lambda_C$. This function takes the following form,

$$\ell(\Lambda_C) = \frac{1}{|\chi_C|} \sum_{X \in \chi_C} \frac{1}{1 + \exp\left[-\gamma\left(d(\chi, \Lambda_C) + \theta\right)\right]}, \tag{8}$$

where $|\chi_C|$ represents the number of recordings in $\chi_C$, and $\gamma$ and $\theta$ are control parameters. Note that for this specific case, we consider the formulation with different competing classes (as many as the number of partitions). Hence, the optimization is performed per partition. Finally, the objective function in Equation 8 can be optimized by applying the following generalized probabilistic descent (GPD) update rule for $\Lambda_C$:

$$\Lambda_C^{t+1} = \Lambda_C^t - \epsilon \nabla \ell(\Lambda_C)|_{\Lambda_C^t}. \tag{9}$$

The procedure is included in the MVE and JFA frameworks. We obtain suitable number of target and imposter models as clusters we have. To come up with a decision, the system computes the score for a new recording, $X$, speaker $S$, and partition $C$ as,

$$\theta_C^S(X) = \log P(X|\Lambda_S^C) - \log P(X|\Lambda_C). \tag{10}$$

Next, the algorithm chooses among the following scoring options.

a) *Partition Selection (PS)*: Assign the recording to the most likely partition as, $\hat{C}(X) = \arg\max_C \log P(X|\Lambda_C)$.
b) *A priori (AP)*: If the correct partition $\Omega_C$ for $X$ is known *a priori*, then we can simply set $\theta^S(X) = \theta_C^S(X)$.
c) *Best-score (BS)*: We select the greatest score, $\theta^S(X) = \max_C \theta_C^S(X)$.
d) *Combination (CO)*: Combine the scores from the different partitions: $\theta^S(X) = \sum_C w_C^S \theta_C^S(X)$. In this case $w_C^S$ are simply the weights assigned by the SVM.
e) *Fusion (FU)*: Combine the scores from the partitions in a logistic regression fusion, $f_j = \alpha_0 + \sum_{i=1}^N \alpha_i a_{i,j}$.

## 7 Integration

We integrate the algorithms described in the two previous sections: *area under the ROC curve optimization* and the *ensemble modeling* approach. We have in mind that the final intention is to combine these techniques for both cases: clean and noisy speech.

We introduce a top down hierarchical architecture that comprises several partition levels. Each level was optimized using the previous techniques. To make it clear, in the first level the data space is partitioned in distinctive clusters for a chosen condition, not necessarily disjunct (using the *ensemble* approach). From those partitions the algorithm generates their corresponding models. Then, we maximize the distance between the competing regions using *ensemble modeling.* The clusters are next refined using the AUC optimization premises. Once we have these new mapped and refined clusters, we continue to the next partition level in the same manner. As a consequence of optimizing for every operation point along the ROC curve, the EER and the minCDF are reduced. We continue the procedure until the partitions are among speakers (final step). In the test stage we perform a hierarchical search of the most "likely" partitions from which the utterance might belong (following a top down search). The system comes up with a decision at every level using the ensemble modeling approach and the scores are computed. The fusion scoring is preferred for the final score computation. As an example, the first-level partitions may correspond to any condition (channel, noise type, SNR), and the second to the actual speakers. Note that the refinement of the models, provided by the AUC optimization, can be included at every level on the methodology to ensure that we are getting optimized models.

## 8 Contributions and Results

Our main contributions deal with the analysis of the discriminative optimization in clean and noise conditions. In this section, we show representative results for every case explored. On one hand, we can highlight the improvements considering the *AUC optimization*. For every approach considered, FA and MVE, we showed that the method is beneficial.

Figure 3 and Table 1 show the system performance applying this optimization. Moreover, the extended version, which deals with noise conditions, presents better improvements.

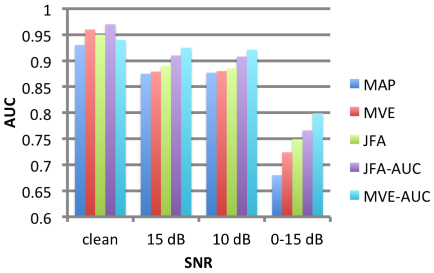


**Fig. 3.** AUC optimization: AUC results for different systems in Noise Conditions

| *System* | *clean* | *15 dB* | *10 dB* | *0-15 dB* |
|---|---|---|---|---|
| *MAP* | 15.95 | 18.01 | 17.48 | 35.7 |
| *MVE* | 13.51 | 17.67 | 17.15 | 28.1 |
| *JFA* | 12.07 | 17.23 | 16.79 | 27.3 |
| *JFA-AUC* | 11.93 | 16.51 | 16.22 | 24.0 |
| *MVE-AUC* | 13.21 | 15.93 | 15.78 | 22.8 |

**Table 1.** AUC optimization: EER of the noisy task (babble noise).

On the other hand, the *ensemble method* shows that by partitioning the space and refining those subspaces, the results improved. The new updated scores depend not only on the target and imposter modeling but on the partition models with their corresponding refinements. In here, we depicted the results for the MVE highlighting that having a prior informations is a plus. But the study opens the discussion about the benefits even when performing blind clustering.

Figure 4 and Table 2 depicts and example of this method.

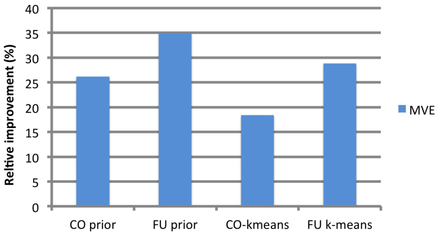


**Fig. 4.** Ensemble Modeling: results example, MVE

| | 5c-MVE | | | | |
|---|---|---|---|---|---|
| *Baseline* | 28.3 | | | | |
| *minDCF* | 15.32 | | | | |
| *Condition* | PS | AP | BS | CO | FU |
| *k-means* | 27.0 | 24.1 | 27.7 | 23.1 | 20.13 |
| *minDCF* | 14.21 | 10.48 | 14.34 | 8.52 | 7.86 |
| *prior info* | 23.7 | 22.2 | 24.4 | 20.9 | 18.4 |
| *minDCF* | 9.21 | 8.35 | 10.59 | 7.78 | 7.16 |

**Table 2.** Ensemble modeling: EER and minDCF for different clusters (noise condition), MVE.

Finally, the *integration* approach combines both methods providing an enhanced modeling. This stage is also the summary of our proposed techniques. Figure 5 and Table 3 show the improvements obtained by this method. Note that

MVE outperforms other methods like JFA and i-vector for this specific task. The improvements are more noticeable at low SNRs.

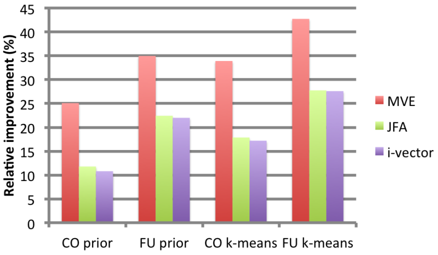


**Fig. 5.** Integration approach: result examples

| | 5c-MVE | | 5c -JFA | | 5c i-vector | |
|---|---|---|---|---|---|---|
| *Baseline* | 28.3 | | 26.3 | | 25.0 | |
| *minDCF* | 15.32 | | 14.03 | | 13.23 | |
| *Condition* | CO | FU | CO | FU | CO | FU |
| *k-means* | 21.2 | 18.4 | 23.2 | 20.4 | 22.3 | 19.5 |
| *minDCF* | 8.14 | 6.57 | 12.74 | 7.36 | 11.87 | 6.84 |
| *prior info* | 18.7 | 16.2 | 21.6 | 19.0 | 20.7 | 18.1 |
| *minDCF* | 7.23 | 6.68 | 8.11 | 7.51 | 7.31 | 6.05 |

**Table 3.** Integration approach: EER and minDCF for different clusters on noise condition, MVE, JFA, i-vectors

## 9 Conclusions and Future Work

Along this thesis we explored the discriminative model optimization from SV perspective. We explored two methods: the *AUC optimization* and the *ensemble modeling*. The former optimizes for every operating point along the ROC curve reducing the error metrics. The latter refines the models, either for any chosen attribute or for specific target speakers. The combination of them results in a hierarchical method, called *integration approach*, in which every level corresponds to a specific attribute space that is optimized by maximizing the AUC of the ROC curve. Each of our proposed algorithms can be easily included as part of the main modeling method, such as FA or MVE.

The *AUC optimization* approach presents competitive performance compared to conventional maximum likelihood or discriminative training methods. Improvements obtained are actually significantly greater on noisy speech than on clean speech. We observed that the AUC optimization naturally accounts for any shift in the data away from the intended operating point where performance is measured. For future research, other modeling techniques, e.g. i-vector and PLDA can be employed. We will also investigate other objective functions that assign weights to the ROC/DET curve so that we can control the curve at a desired operating point.

The *ensemble modeling* exhibited the best results when a priori knowledge of the factor is available. Much of this benefit may be obtained if this information is only estimated for the training data. We will investigate this in future work. Moreover, it would be useful to easily identify partitions when multiple factors must be considered concurrently and particularly when they must be estimated.

Finally, we observed that performing an *integration*, the actual baselines improve, especially under noisy conditions. In the future, this research will look for suitable techniques that can provide the appropriate hierarchical structure for each application.

## References

1. T. Weigold, T. Kramp, and M. Baentsch, "Remote Client Authentication," *IEEE Security & Privacy*, vol. 6, no. 4, pp. 36–43, 2008.
2. A. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *IEEE transactions on information forensics and security*, vol. 1, no. 2, pp. 125–143, 2006.
3. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.
4. T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
5. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech' 97*, 1997, pp. 1895–1898.
6. N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.
7. P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
8. B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *Signal Processing, IEEE Transactions on*, vol. 40, no. 12, pp. 3043–3054, 1992.
9. D. Povey, P. Woodland, and M. Gales, "Discriminative MAP for acoustic model adaptation," in *IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc. (ICASSP)*, vol. 1, 2003, pp. I–312.
10. C. Ma and E. Chang, "Comparison of discriminative training methods for speaker verification," in *IEEE Intl. Conference on Acoustics, Speech, and Sig. Proc.(ICASSP)*, vol. 1, 2003, pp. 192–195.
11. M. G. Rahim, C. H. Lee, B. H. Juang, and W. Chou, "Discriminative utterance verification using minimum string verification error (MSVE) training," *Proc. ICASSP*, vol. 6, pp. 3585–3588, 1996.
12. L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *IEEE Intl. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2011, pp. 4832–4835.
13. S. Cumani, N. Brummer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i-vector space," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011, pp. 4852–4855.
14. B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.
15. H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18:1, pp. 50–60, 1947.
16. Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE transactions on Speech*, vol. 43, pp. 781–785, August 1994.