A VOICE USER INTERFACE DEMONSTRATION SYSTEM FOR MEXICAN SPANISH

Carmen Garcia-Mateo, Qiru Zhou**, Chin-Hui Lee**, Andrew Pargellis***

* E.T.S.I. de Telecomunicación, University of Vigo, SPAIN (carmen@tsc.uvigo.es)
** Dialogue Systems Research Department, Bell Laboratories, Lucent Tech. USA
({qzhou,chl,anp}@research.bell-labs.com)

ABSTRACT

We present a Mexican Spanish voice user interface demonstration system. It was built on a speech research platform developed at Bell Labs, which provides major speech technology and interface components, including automatic speech recognition, text-to-speech synthesis, audio input/output functions and telephone interface. The application is written in the PERL script language with an embedded Voice Interface Language (VIL) that connects the speech and interface modules to PERL. Given the set of multilingual speech processing capabilities on the platform and the VIL, we were able to quickly develop a Mexican Spanish system using PERL with speech-enabled messaging and information access functionality similar to our English voice user interface demonstration system.

1. INTRODUCTION

Designing multilingual systems capable of handle speech processing for multiple languages and dialects is becoming a technical challenge nowadays due to the globalization of telecommunication networking and the need for expanding conventional voice and data services into multimedia applications. Therefore we are witnessing an increasing demand of universal information access that speech processing plays a critical role. One of the languages of great interest is Spanish, due to the vast speaking population with a large geographic spread and wide cultural diversity. Several dialects with many distinct feature differences also exist.

We are interested in experimenting with a research platform that meets the requirements of performing multi-lingual speech processing. We have adopted the Bell Labs Speech Technology Integration Platform (BL-STIP) [1], which is a modular client/server architecture with a well-defined system interface. It makes the set of advanced speech technology components easily accessible and enables quick integration and prototyping of speech applications.

BL-STIP is also a language independent architecture and all the language dependent components of the system are introduced through tables and files. Therefore it is only required to modify a few language dependent files in order to implement the Mexican Spanish version of the voice user interface (VUI) system that we originally developed for American English [2]. In the following, we focus our discussion on the effort in developing these language dependent components, specifically the automatic speech recognition (ASR) and the utterance verification (UV) modules for Mexican Spanish. We will also briefly touch upon the text-to-speech (TTS)[3-4] synthesis module which is not only needed to produce Mexican Spanish speech but also used to generate the lexicon for all the vocabulary words needed in ASR and UV and for training the acoustic models of Spanish phones. It is important to realize that porting the English VUI system to Mexican Spanish is rather straightforward. More attention should be paid to the flexible methodology we adopted in our design.

2. SPEECH RECOGNITION

In order to put together the demonstration systems it was necessary to design the recognition system in Mexican Spanish. In this section, we describe the procedure for the design of a subword based speech recognizer in Mexican Spanish. The tasks to accomplish are:

- Training database selection
- Context-Independent Subword Unit Set design.
- Context-Dependent Subword Unit Set design
- Training of the acoustic HMMs models
- Initial assessment of the recognizer

2.1. Some Language Considerations

Several accents and dialects can be found for the Spanish language and presently there are no relevant results stating at what extent the dialect plays a role in the recognizer performance. The lack of speech databases for Spanish could be a key factor for that. So far most of the work in speech recognition for Spanish has been done for what is

known as “Continental Spanish” (Spanish spoken in the central area of Spain). The SpeechDat database for Spanish spoken in Spain is one clear example of it [5].

In contrast, spoken Spanish has been less studied in other areas, such as in certain regions of the USA, Mexico and South America. Nevertheless, there is enormous interest in deploying speech technology products in the American countries where Spanish is spoken. The recently launched SALA (SpeechDat across Latin America) project [6] is an example of such an interest among companies and academia. Its main objective is to acquire databases for the Spanish dialects in America following the stated procedure in the EU-project SpeechDat. The SpeechDat databases will be distributed by ELRA in the near future.

Meanwhile, the VAHA (Voice Across Hispanic America) [7] database, collected by Texas Instruments for the Linguistic Data Consortium (LDC), is one of the first available public databases for Spanish collected over the phone in what is known as “Hispanic USA”. This is the database we use in this study.

From the training set of the VAHA Database, we select a subset as our training database. It consists of phonetically rich sentences and telephone numbers. Only those speakers that originated from Arizona, California, Texas, New Mexico (USA) and Mexico have been selected for this experiment. Speakers above 60 years old have not been considered. By doing so, the speaker accent may be kept at a minimum difference across the database.

We have 402 speakers with a total number of 4,457 sentences and a vocabulary of 5,774 distinct words. This data has been used to train subword and anti-subword models. The anti-subword models can be used in utterance verification [8], which we will discuss more in Section 3.

2.2. Subword Unit Set Design

The first step in the design of the subword unit set is to determine the basic number of language allophones we are going to consider. In addition to providing the synthesized speech, which is a critical component in a dialogue system, the Bell-Labs Text-to-Speech (TTS) system for Mexican Spanish, MEXTTS, is used to generate the pronunciation lexicon for the vocabulary of the VAHA task. This TTS uses 28 allophones that we have reduced further down to 25 as shown in Table 1. The reason to group some of these allophones is that their high acoustic similarity makes it unnecessary for us to use two different models for each allophone.

CI unit	TTS Symbol	Example	CI unit	TTS Symbol	Example
a	a	*c*asa	z	Z	Asno
e	e	perro	h	H	Ca*j*a
i	i	ni*ñ*o	C	C	Co*ch*e
o	o	coche	r	R	Pe*r*o
u	u	nu*n*ca	R	R	Pe*rr*o
g	g, G	Ten*g*a, a*g*ua	l	L	Co*l*a
b	b, B	Tum*b*a, a*v*ión	y	Y	mi*e*do, ca*ll*a
d	d, D	Pan*d*a, a*d*iós	w	W	C*u*erpo
p	p	*p*erro	n	N	Ca*n*a
t	t	ga*t*o	X	N	ni*ñ*o
k	k	*c*asa	N	N	nu*n*ca
f	f	ca*f*é	m	M	ca*m*a
s	s	casa			

Table 1: List of context independent (CI) units for Spanish.

This 25 phone set plus a model for the background silence constitutes our simple set of context-independent (CI) phone units. The allophone distribution over the training database is depicted in Figure 1.

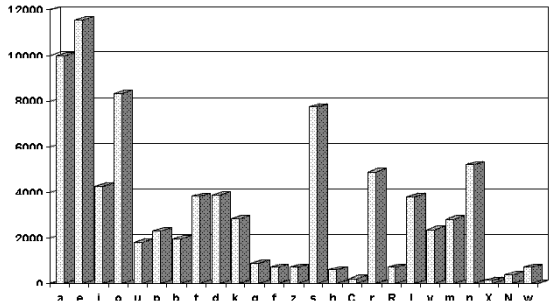

Figure 1: Allophone distribution over the training set.

In order to broaden the context coverage to deal with future unknown tasks we use right context-dependent (RCD) units [9]. Since not all RCD phone units appear in the training data set with the required minimum frequency, we use a threshold of 30 repetitions to consider a unit as trainable. This resulted in 269 RCD HMMs as opposed to the full set of 676 units. To deal with some unseen contexts, we also include in the RCD unit set the 26 CI unit set which brings our subword set to a total number of 295 units.

2.3 Recognizer Description

The speech input is sampled at 8 kHz and pre-emphasized using a first-order filter with a coefficient of 0.97. Frames are 30 ms long with a frame shift of 10 ms. LPC analysis of order 10 is conducted every frame. The recognizer feature set consists of 38 parameters that include 12 cepstral coefficients, 12 delta cepstral coefficients, 12 delta-delta cepstral coefficients, delta log energy and delta-delta log energy [9]. Except for the background silence unit, each subword unit is modeled by a 3-state left-to-right HMM with no state skip. A mixture Gaussian state observation density characterizes each state. Training is performed with an iterative segmental ML algorithm [9] in which all utterances are first segmented into subword units. The Baum-Welch algorithm is then used to estimate the parameters of the mixture Gaussian densities for all states of subword HMMs. Recognition is accomplished by a frame synchronous beam search algorithm to determine the sequence of words (phones) that maximizes the likelihood of the given utterance.

In Table 2, we show the phone recognition rates over the training set for the CI and RCD models. As expected, the RCD units outperformed the CI unit set. It is noted that there was no attempt at balancing the phone insertion and deletion errors.

# units	#mix.	% Corr.	% Subs.	% Del.	% Ins.	% Error	%PA
26	32	63.4	23.8	12.8	12.3	48.8	**51.2**
295	16	79.0	15.3	5.8	17.6	38.6	**61.4**

Table 2: Phone accuracy over the training set

We also conducted a continuous speech recognition experiment, which had a 34-word vocabulary. The syntax is represented by a deterministic finite-state grammar, which was defined from a subset of 31 short sentences, each with at most 5 words. 153 utterances from a number of female speakers were used for testing on the set of 295 RCD unit models. A 93.5% sentence accuracy was achieved on this simple task. This high ASR performance gives us confidence that a good dialogue system can be designed.

3. UTTERANCE VERIFICATION

Utterance verification [8] is used to associate confidence measures to recognized words and phrases. These measures enable us to mimic an intelligent human-machine user interface. What's needed is a *confidence measure* (CM) on any recognized word or phrase that gives an indication how well it is recognized. Based on the defined CM, the dialogue interface decides how much to confirm and what to re-prompt. It also helps with designing partial understanding strategies that are critical to handle ill-formed utterances. We have experimented with a number of CMs which all have the following form [10],

$$CM_W(O) = f(\{LLR_i(O_i)\}), \qquad (1)$$

where f is a function, O is the speech utterance associated with the recognized event W, and O_i is the speech segment that corresponds to the i^{th} subword in W. LLR_i is the log likelihood ratio score of the i^{th} subword, evaluated as

$$LLR_i(O_i) = \log P(O_i \mid \lambda_i) - \log P(O_i \mid \eta_i), \qquad (2)$$

with λ_i being the HMM, and η_i being the anti-HMM, for the i^{th} subword unit respectively, i. e. η_i is represented by an HMM trained with data from the "most competitive" units to unit i. This set of competitors is called a phone *cohort* set. One cohort-set model is trained for each subword i to characterize η_i.

In our implementation, for each subword i from the CI unit set, except for the background silence unit, the cohort set of size 5 is determined. The PDF of the *LLR* measure is computed for both the true segments and the other competing segments. See Figure 2 for an example of the PDFs of the *LLR* score for the vowel unit "e" in Table 1 and its corresponding cohort set.

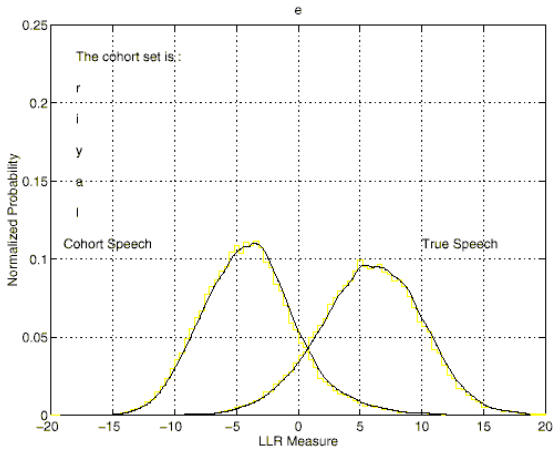

Figure 2: PDFs of the *LLR* score for the subword "e".

The overlap between the PDFs means that the *LLR* for a single subword can not be used for a hard decision rule. In order to reject or accept a word; we must compute all the *LLR*s for the subwords in the recognized string and group them to form word or phrase level CMs.

4. APPLICATION DESCRIPTION

The application is written using an embedded Voice Interface Language (VIL) that enables rapid prototyping. The VIL is implemented in the PERL script language environment. The interface functions provide an advanced speech interface control such as ASR task definition, barge-in, dynamic grammar loading, and text-to-speech synthesis. The VIL also provides functionality to write spoken dialogue applications [2].

The VUI demo system allows the user to retrieve information using the telephone. A user can gain access to the system if there is a user profile in the access control database and the user speaks the correct password. Voice control services are divided into three groups:

- Messaging services: voice, fax and e-mail messages.
- General information services: headline news, sport news, stock quotation, weather report, etc.
- Telephony: placing a call, call transfer, etc.

The user can ask for instructions at any time and the system voice prompts are designed to be clear and concise in order to avoid confusion. The dialogue is mainly menu-driven with some user initiative, since at any time the user can go across the three groups of menus and go down through another branch of the main menu. The dialogue manager is modeled by a finite state machine which remembers each dialogue turns and provide a session tracing mechanism to record the dialogue session for further studies. See reference [2] for details.

5. SUMMARY

Given the multilingual capability of BL-STIP and flexibility of VIL, we are able to quickly build a messaging and information service demo using a Mexican Spanish voice user interface. It provides a similar functionality to our English voice user interface system built on BL-STIP.

It is noted that the language and domain independent architecture of BL-STIP and the platform independent design of VIL make it possible for the quick Mexican Spanish extension.

Acknowledgments

The authors would like to thank their colleagues, Chilin Shih, and Padma Ramesh of Bell Labs, for providing the required software tools to develop this study.

Dr. García-Mateo's work has been partially supported by Spanish CICYT under the project TIC96-0964-C04-02 and by Xunta de Galicia.

6. REFERENCES

1. Q. Zhou, C.–H. Lee, W. Chou and A. Pargellis, "Speech Technology Integration and Research Platform: A System Study," *Proc. EuroSpeech-97*, pp. 621-624, Rhodes, Greece, 1997.

2. A. Pargellis, Q. Zhou, C.-H. Lee, "A Language for Creating Speech Applications," *Proc. ICSLP-'98* (presented in this conference) .

3. R. Sproat, Editor, "Multilingual Text-To-Speech Synthesis, The Bell Labs Approach," *Kluwer Academic Publishers*, 1998.

4. TTS Home Page: http://www.bell-labs.com/project/tts

5. SpeechDat Homepage: http://speechdat.phonetik.uni-muenchen.de

6. A. Moreno, H. Höge, J. Koehler, J.B. Mariño. "SpeechDat Across Latin America. Project SALA" in *Proc. of First International Conference on Language Resources and Evaluation*. pp.367-370, Granada, Spain. May 1998.

7. VAHA Home Page: http://morph.ldc.upenn.edu/ldc/catalog/html/speech.html/vaha.html

8. R. A. Sukkar and C.–H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword-Based Speech Recognition*", IEEE Trans. Speech and Audio Proc.*, Vol. 4, No. 6, pp. 420-429, 1996

9. C.–H. Lee, B.–H. Juang, W. Chou, and J. J. Molina-Perez, "A Study on Task-Independent Subword Selection and Modeling for Speech Recognition*," Proc. ICSLP-96*, Philadelphia, 1996.

10. T. Kawahara, C. – H. Lee and B.-H. Juang, "Key-Phrase Detection and Verification for Flexible Speech Understanding", *Proc. ICSLP-96*, pp. 957-960, Philadelphia, Oct. 1996.