BILINGUAL SPEECH RECOGNITION IN TWO PHONETICALLY SIMILAR LANGUAGES

Vicente Alabau[*]*, Carlos D. Martínez*[*]

[*]Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera, s/n. 46071 València, Spain
{valabau,cmartine}@dsic.upv.es

ABSTRACT

As Speech Recognition Systems improve, they become suitable for facing new problems. Multilingual speech recognition is one of such problems. In the present work, the case of the *Comunitat Valenciana* multilingual environment is studied. The official languages in the *Comunitat Valenciana* (Spanish and Valencian) share most of their acoustic units, and their vocabularies and syntax are quite similar. They have influenced each other for many years. A small corpus on an Information System task was developed for experimentation purposes. This choice will make it possible to develop a working prototype in the future, and it is simple enough to build semi-automatic language models. The design of the acoustic corpus is discussed, showing that all combinations of accents have been studied (native, non-native speakers, male, female, etc.). In addition, some experiments have been conducted with this corpus that show promising results for a Spanish-Valencian multilingual speech recognizer.

1. INTRODUCTION

The quality of Automatic Speech Recognition (ASR) has improved greatly in recent years [1, 2, 3]. Some commercial products have appeared for real-world tasks, such as speech transcription systems in restricted domains and automatic call centres. However, some problems arise in these real-world tasks: recognition performance is low under adverse circumstances, and the models are very noise sensitive [4, 5].

In this paper, we design and acquire a corpus to research one of these problems: multilingual interoperability. The problem of multilingual interoperability presents several issues related to the components of a classical ASR system, like acoustic or language models.

With respect to acoustic models, ASR systems are very language-dependent, because the phone sets are different in each language. Moreover, coarticulation effects of the same phonemes may differ in each language, and even the articulation of a phoneme may have its own singularities. Some work, for example the introduction of contextual acoustic models (triphones), has already been done to find more robust acoustical units under these conditions [6].

Language models are also very language-dependent, because of their vocabulary and gramatical issues. Furthermore, vocabulary transcription is dialect-dependent as well. For example, Spanish utterances from South America and Spain differ in a noticeable way.

Language determination is also an important issue. In some tasks, the speaker's language is unknown. Thus, the system has to find the best way to determine which language it is. Moreover, when the system has to answer the speaker, the identification of the language is needed in order to be able to answer in the same language.

In multilingual environments, other difficulties are added to speech recognition, even in monolingual ASR. Languages are usually influenced by other languages that are present in the environment and by the speaker's mother tongue (e.g., the perception distortion of a non-native Dutch speaker is equivalent to a reduction of the signal-to-noise ratio of 3-4 dB for non-native Dutch speakers [7]). This interference is demonstrated by mispronunciation and the use of syntactical structures and vocabulary from the mother tongue. For anyone who has studied foreign languages, it is easy to understand that phonemes that are not present in the mother tongue are hard to pronounce. It is even possible to identify the nationality of some people by their accent. Some syntactical and vocabulary mistakes are produced by the lack of knowledge of the foreign language.

In this work, the case of the *Comunitat Valenciana* is studied. In *Comunitat Valenciana*, two official languages coexist: Spanish and Valencian. Valencian is the name for the Catalan language dialect that is spoken in the *Comunitat Valenciana*. Catalan is one of the most widely spoken minor languages in Europe. About 6.5 million people speak it actively (on a daily basis), and about 12 million people are potential speakers (they know the language but use Spanish on a daily basis). Furthermore,

Work supported by the "Agència Valenciana de Ciència i Tecnologia" under grant GRUPOS03/031, the Spanish project TIC2003-08681-C02-02 and the "Programa d'Incentiu a la Investigació 2004 UPV".

the Catalan government is making an important effort to promote the use of the Catalan language in all spheres. Therefore, there is great interest in the speech recognition technologies for Catalan.

As official languages, every citizen has the right to know and use both Spanish and Valencian in the *Comunitat Valenciana*. However, the repression of the use and learning of Valencian in the Franco period (1939-1977) (also called Catalan Negationism) and other historical reasons, have caused that currently only 85% of the population of *Comunitat Valenciana* understand Valencian, and only 48% are active speakers[8]. This has also caused the Valencian phone set to be reduced by the extensive use of Spanish, which is true even for Valencian native speakers (a situation which has not occurred in other Catalan dialects). Thus, nowadays the Valencian phone set differs very little from the Spanish phone set.

In the following sections, we describe the design of a multilingual corpus for Spanish and Valencian, and we summarize the most common multilingual approaches presented in the literature. We also present preliminar results on this corpus that show the performance of each aproach. Conclusions and future work are presented in the last section.

2. MULTILINGUAL CORPUS DESIGN

As stated above, the Valencian dialect has special phonetic features with respect to standard Catalan. Thus, although there are a few speech recognition resources for the Catalan language there was no resource for Valencian, and a Valencian language corpus had to be acquired. For this reason, we had to acquire a specific Valencian speech corpus and a similar Spanish one. Although Spanish speech corpora are available [9], it was important to have Spanish and Valencian corpora with the same features to be able to compare them more faithfully.

Thus, we decided to acquire our own multilingual corpus specifically for experimentation purposes (i.e., not for real system development). We planned to acquire a small, simple corpus and decided to design a set of 120 medium-length sentences (60 for each language) for 20 speakers, which corresponds to approximately 1 hour of speech per language (actually, the length of the recorded signal is about 2 hours). This amount of speech signal should be enough for the experimental purposes that the corpus is going to be used for.

We chose an Information System task to design the corpus. This was done because this task is complex enough for demostration purposes, and it is simple enough to semi-automatically generate the task sentences. As there are few syntactic differences between Spanish and Valencian (especially for this task), the semi-automatic sentences could be easily translated. Dictionary translation for single words and some minor modifications were sufficient to accomplish the translation task.

The goal of this Information System was to provide information about the staff of a department by phone [10]. The possible information items the system could be asked for included timetables, office hours, phone numbers, e-mail addresses, or office locations. Some example sentences are shown in Figure 1.

Spanish

- Por favor, quiero saber el e-mail de Álvaro Rodríguez, adiós.
- Buenas noches, quería la extensión de la señorita Silvia Abrahao, muchas gracias.
- Buenos días, ¿cuál es el horario de consultas del doctor Vicente?, gracias.

Valencian

- Per favor, vull saber l'e-mail d'Álvaro Rodríguez, adeu.
- Bona nit, volia saber l'extensió de la senyoreta Silvia Abrahao, moltes gràcies.
- Bon dia, quin és l'horari de consultes del doctor Vicente?, gràcies.

English

- Please, I want to know the e-mail of Álvaro Rodríguez, goodbye.
- Good evening, I wanted to know the extension of Miss Silvia Abrahao, thank you very much.
- Good morning, what are the office hours of the Dr. Vicente?, thanks.

Figure 1. This is a selection of sentences that are representative of the corpus. The English sentences are provided for a better understanding of the examples.

This task was tested in a previous work [10] with acoustic models that were designed for other tasks. This work showed promising results in bilingual Valencian-Spanish ASR and has encouraged us to continue research in this field.

3. LANGUAGE MODELLING

Language modelling is crucial in an ASR system. Language models define which kind of sentences are allowed in the system. Therefore, any sentence said by a speaker will not be recognized correctly if it does not belong to the language model. Indeed, this sentence will be recognized as the one that is closest to one that exists in the language model.

The language model of this corpus was designed to suit our experimentation needs. That is, it should be

block	greeting
Spanish	por favor, buenas noches, buenos días
Valencian	per favor, bona nit, bon dia
English	please, good evening, good morning
block	question
Spanish	quiero saber, quería saber, cuál es
Valencian	vull saber, volia saber, quin és
English	i want to know, i wanted to know, what is
block	information
Spanish	el e-mail, la extensión, el horario de consultas
Valencian	l'e-mail, l'extensió, l'horari de consultes
English	the e-mail, the extension, the office hours
block	title
Spanish	señorita, doctor
Valencian	señoreta, doctor
English	Miss, Dr.
block	person
All languages	Álvaro Rodríguez, Àlvar Rodríguez, Álvaro, Àlvar, Rodríguez, Silvia Abrahao, Silvia, Abrahao,Vicente, Vicent
block	farewell
Spanish	gracias, muchas gracias, adiós
Valencian	gràcies, moltes gràcies, adeu
English	thanks, thank you very much, goodbye

Table 1. This table shows examples of phrases belonging to the blocks for Valencian and Spanish. English phrases are provided for a better understanding of the examples.

able to model Valencian and Spanish separatedly, but it should also be able to model a mixture of both languages. The latter is due to the fact that non-native speakers may use words of their native language when the correct word is unknown. This fact is known as barbarism. The modelling of barbarism is not only relevant to multilanguage environments but also to communities with a large number of immigrants.

In order to provide barbasism tolerance to some extent, the sentences were divided into six blocks, each of which represents a concept in the sentence. As we will see below, this allowed us to construct an automaton that could switch between languages (block-combined language models). The blocks were: greeting, question, information, title, person, and farewell. A set of frequently used phrases was used to build an acceptor automaton for each block. An acceptor automaton accepts only a set of given sentences, in this case, the phrases of the block. Samples of these phrases for the sentences in Figure 1 are shown in Table 1.

Finally, these block-oriented automata were used to build the final automata. Two methods were applied in this task:

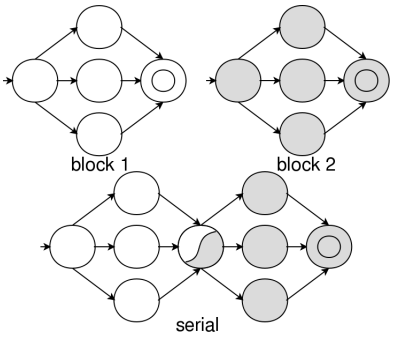

Figure 2. Illustration of the serialization process.

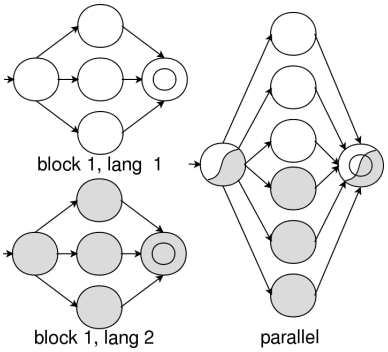

Figure 3. Illustration of the parallelization process.

- Separate language models: an automaton was build for each language. It was made by joining the block-oriented automata in a series. For every two consecutive automata, the final states of the first automaton were merged with the initial states of the second one. Figure 2 shows an example of the serialization process.
- Block-combined language models: a single automaton was built by joining two automata. The automata were joined in parallel on a block-basis manner. Thus, the initial states (and the final states) of each language were merged for each block. Figure 3 shows an example of the parallelization process. Afterwards, the joint blocks were also joined in series. Figure 4 shows an example of the parallelization process for the joint automata.

The automaton corresponding to the block 'person' was, in both cases, the list of all the people in the two languages. This reflects the natural tendency of speakers to call people the way they are used to doing so. Moreover, the names and surnames were allowed separately as well.

4. CORPUS ACQUISITION

The corpus should have about 1 hour per language in order to make a quick acquisition and to be long enough

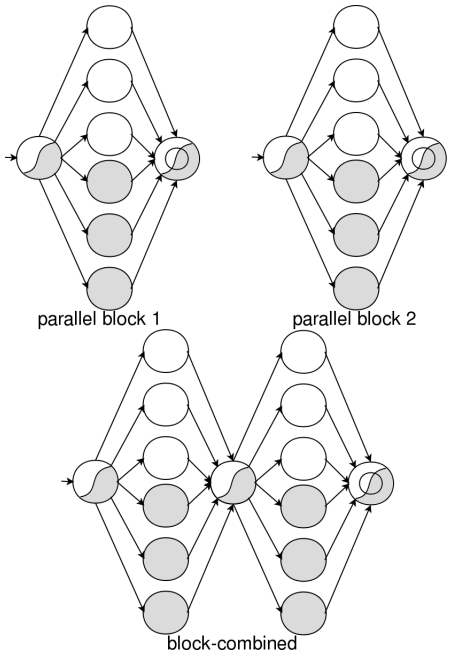

Figure 4. Illustration of the combined parallelization and serialization process.

to train reliable acoustic models in future experiments. Assuming that the average length of an utterance is 3 seconds, we decided to design a set of 120 sentences (60 for each language) for 20 speakers. This provides approximately 1 hour of speech signal per language.

The separate language models were used to generate the corpus. However, a human reviewer was needed to correct the syntactic inconsistencies introduced by the block-oriented automata development, such as gender and number agreement.

The corpus acquisition was developed on the telephone line. Half of the volunteer speakers were native Spanish speakers and the other half were native Valencian speakers. Both languages were acquired from all the participants; thus, non-native speech was recorded for both languages. Male and female speakers were equally distributed in these groups.

In the final design of the corpus, there were five groups of people with four people per group. Each group contained people of all types (men/women, Spanish/Valencian). With this distribution, we ensured a balanced distribution of native and non-native utterances, along with male and female utterances, for both languages.

The Spanish phone set was formed by 26 phonemes in the phonetical scheme that we used. Transcriptions were automatically performed following the rules described in [11] for the SAMPA phonetic alphabet [12]. However, Valencian pronounciation does not follow clear, simple rules as Spanish does. No studies have been done to help us transcribe the sentences automatically. Therefore, the Valencian transcriptions were performed manually for each word of the vocabulary, including all the known phonetic variations. The Valencian phone set we used differs by only one phoneme from the Spanish set. The Spanish phoneme $/c/$ (as in *zapato* $/capato/$) is not present in Valencian, but $/\int/$ (as in *roig* $/ro\int/$) is. The remaining phonemes are shared between the two languages.

Each acquisition session lasted an average of 50 minutes. Although literal reading was compulsory, the speakers were allowed to pronounce Valencian and Spanish as they normally do.

Nearly 2 hours of speech signal were actually acquired (including the silences) for each language. The signal was recorded with a GSM encoding at 8000 Hertz using a 3COM U.S. Robotics modem [13]. Although the GSM encoding signal provides worse quality than a-law or mu-law encoding, the fact is that the GSM encoding is currently being widely used in mobile telephony. As the mobile phone market is rising sharply, most of the potential users of these systems will use mobile phones and, other encoding schemes will not improve the signal quality.

The ambient noise was the typical noise found in a computer laboratory with the occasional mobile phone intefering with the phone line. Silences at the beginning and end of the speech signal were not removed.

5. EXPERIMENTS

In this section, preliminar experiments on the adquired corpus are presented. The aim of these experiments was not to measure the recognition accuracy of the models but to observe the impact of multilingual modelling versus monolingual modelling. Not surprisingly, the results presented in this work will not be very accurate if it is taken into account the small corpus provided. Nevertheless, the expriments will show the trend of the applied techniques.

Three evaluation measures have been used in the experiments. These measures have been selected in order to show the most interesting points of this work:

- *Word Error Rate (WER).* The WER is a measure of the ASR quality given a reference sentence. It computes the edition distance between the recognized sentence and the reference sentence.
- *Semantic Word Error Rate (SemWER).* The SemWER is a measure of the recognition quality for the semantic fields. In this work, the semantic fields are the required information (timetables, office hours, phone numbers, e-mail addresses, and office locations) and names. SemWER is computed as the WER for these fields.
- *Language Identification Rate (LIR).* This rate measures the language identification performance of

		Spanish	Valencian
Training	Sentences	240	240
	Running words	2887	2692
	Length	1h 33m	1h 29m
	Vocabulary	131	131
	Perplexity	3.32	3.70
Test	Sentences	60	60
	Running words	705	681
	Length	23m	21m
	OOVs	0	2
	Perplexity	5.80	6.14

Table 2. Corpus statistics. OOV (Out Of Vocabulary) words are that ones which have been observed in the test corpus but not in the training one.

Acoustic	Spanish	Valencian
Spanish	15.4 / 12.1	–
Valencian	–	19.5 / 13.5
Shared	14.0 / 10.5	18.9 / 14.0

Table 3. Monolingual WER / SemWER.

the models. It is computed as the percentatge of sentences identified correctly.

The Table 2 summarizes the statistics of the corpus. It should be noted that, although the perplexity of the language is very low, the size of the speech corpus is small as well. Therefore, ASR parameters obtained were not as accurate as might be desired, which causes the high error rates presented in the results.

The acoustic models were obtained using the HTK toolkit. The HMMs followed a 3-state left-to-right topology without skips. A 64 Gaussian mixture was used in each state. Furthermore, due to the small amount of data provided, only context independent phonemes were trained.

Table 3 shows monolingual ASR performance as a baseline for the experiments. It can be seen that Valencian performs worse than Spanish. It may be due to the higher variability in Valencian pronunciations, and the quality of the automatic phoneme transcriptior used in the training proccess, which was not as accurate as the Spanish one.

The WER and SemWER for various acoustic and language models is presented in the Table 4. The separate acoustic models behave especially bad for Valencian. As it has been explained before, the Valencian acoustic models have worse parameter estimation than the Spanish ones. However, the second method proposed achieves almost the same performance as the monolingual ones. As shared acoustic models have double training data, the parameters are estimated more precisely.

It might seem weird that block-combined model achieve better SemWER having such WER for Spanish. This fact can be explained by the similarity of both languages. A Spanish utterance may be easily confused with a Valencian one and viceversa by the ASR, especially in block-combined models, in which confusion may come at block levels. This would lead to WER errors as vocabularies of both languages are different. However, the semantic of the recognition would not be affected.

Models	Spanish	Valencian	Average
Sep-Sep	16.0 / 13.6	26.1 / 20.7	21.1 / 17.5
Sha-Sep	15.9 / 11.5	21.0 / 16.3	18.5 / 13.9
Sha-Blo	20.2 / 10.6	21.4 / 16.3	20.8 / 13.5

Table 4. Multilingual WER / SemWER. The column Models contains the models that were used to obtain the rates, where Sep-Sep stands for separate acoustic and language models, Sha-Sep for shared acoustic models and separate language models, and Sha-Blo for shared acoustic models and block-combined models.

Models	Spanish	Valencian	Average
Sep-Sep	99.6	85.8	92.7
Sha-Sep	97.5	99.6	98.6
Sha-Blo	91.2	96.7	94.0

Table 5. Language identification rate. The column Models contains the models that were used to obtain the rates, where Sep-Sep stands for separate acoustic and language models, Sha-Sep for shared acoustic models and separate language models, and Sha-Blo for shared acoustic models and block-combined models.

Looking carefully at the Table 5, it may be noticed that there is a high correlation between the language identification rate of each language, and its correspondent word error rate. In systems where LIR is above 99%, the ASR multilingual performance is almost equivalent to the monolingual system. It is important to observe that LIR rates very well even with these two languages so phonetically and gramatically similar.

6. CONCLUSIONS AND FUTURE WORK

Specifically, experiments were focused on two goals. The first one was to evaluate the corpus in a coupled multilingual speech recognizer which was expected to perform similarly or better than a separate recognizer. The second goal was to obtain a good ratio in speaker-language identification. Both goals were succesfully achieved, even with such a small corpus.

In average, the best results were achieved using shared acoustic models and separate language models. It is obvious that with such similarity in phonetics, shared models behave better because more training data is being used for the same models. However, it should be noted that for languages with more different phonetics that could be not the case [14]. Following

the same reasoning, block-combined models would mix up word from different languages, because for many of them they are pronounced almost in the same way. Consequently, separate language models rate higher. Other works sustain this observation for a wider range of languages [15].

The purpose of this paper was to acquire a corpus to assess the viability of this research line. The acquisition of an acoustic corpus is a tedious task, and therefore, we decided to acquire a minimal corpus which may be a drawback for larger experiments. However, huge Spanish acoustic resources are widely available in the community. The Valencian acoustic signal of this corpus could be used for adaptation purposes, e.g., to adapt good Spanish acoustic models to the Valencian dialect by means of speaker adaptation techniques [16]. Previous work supports this approximation [17].

Finally, further experiments are planned for a larger corpus which is already being acquisited for all the Catalan dialects [18].

Acknowledgements

The authors wish to thank the volunteer speakers who participated in the multilingual corpus acquisition.

7. BIBLIOGRAPHY

[1] A. L. Gorin, G. Ricardi, y J. H. Wright, "How may i help you?," *Speech Communication*, vol. 35, pp. 113–127, 1997.

[2] J. Chu-Carroll y R. Carpenter, "Vector-based natural language call-routing.," *Computational Linguistics*, vol. 25(3), pp. 361–388, 1999.

[3] R. Billi, F. Canavesio, y C. Rullent, "Automation of telecom italia directory assistance service: Field trials results," in *Proc. IVTTA 1998*, Turin, 1998, pp. 11–16.

[4] S. Furui, "Toward robust speech recognition under adverse conditions," in *Proceedings ESCA Workshop on Speech Processing in Adverse Conditions*, Nov. 1992, pp. 31–41.

[5] H. G. Huang, "Speech recognition in adverse environments.," *Computer Speech and Language*, vol. 5, pp. 275–294, 1991.

[6] R. Eklund y A. Lindström, "Xenophones: An investigation of phone set expansion in swedish and implications for speech recognition and speech systhesis.," *Speech Communication*, vol. 35, pp. 81–102, 2001.

[7] Sander J. van Wijngaarden, "Intelligibility of native and non-native dutch speech.," *Speech Communication*, vol. 35, pp. 103–113, 2001.

[8] Institut Valencià d'Estadística, Comunitat Valenciana, *Població en vivendes familiars de 3 i més anys, segons el coneixement del valencià, l'edat i el sexe.*, 2001.

[9] J.E. Diaz-Verdejo, A.M. Peinado, A.J. Rubio, E. Segarra, N. Prieto, y F. Casacuberta, "Albayzin: a task-oriented spanish speech corpus," in *Proceedings of LREC*, 1998, vol. 1, pp. 497–501.

[10] F. Mas, J. Vicedo, y C.D. Martínez-Hinarejos, "Development of a voice-based telephone information system," in *Actas de las III Jornadas de Tecnologías del Habla*, Valencia, Spain, Nov. 2004.

[11] A. Quilis, *Tratado de fonología y fonética españolas.*, Madrid (Gredos), second edition, 1999.

[12] UCL, *SAMPA computer readable phonetic alphabet*, 1993.

[13] USRobotics, *3Com U.S. Robotics 56K Message Modem Users Guide and Reference*, 1993.

[14] J. Köhler, "Multilingual phone models for vocabulary-independent speech recognition tasks," *Speech Communication*, vol. 35, pp. 21–30, 2001.

[15] U. Uebler, "Multilingual speech recognition in seven languages," *Speech Communication*, vol. 35, pp. 53–69, 2001.

[16] C.J. Leggeter y P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models.," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[17] T. Schultz y A. Waibel, "Language-independent and language-adaptative acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[18] A. Moreno, A. Febrer, y L. Márquez, "Generation of Language Resources for the Development of Speech Technologies in Catalan," in *Proc. of LREC'06*, 2006.