Acoustic Modeling and Training of a Bilingual ASR System when a Minority Language is Involved

Laura Docío-Fernández and Carmen García-Mateo

Departamento de Teoría de la Señal y Comunicaciones
E.T.S.I. Telecomunicación
Campus Universitario de Vigo
36200 VIGO, SPAIN
ldocio@gts.tsc.uvigo.es | carmen@gts.tsc.uvigo.es

Abstract

This paper describes our work in developing a bilingual speech recognition system using two SpeechDat databases. The bilingual aspect of this work is of particular importance in the Galician region of Spain where both languages Galician and Spanish coexist and one of the languages, the Galician one, is a minority language. Based on a global Spanish-Galician phoneme set we built a bilingual speech recognition system which can handle both languages: Spanish and Galician. The recognizer makes use of context dependent acoustic models based on continuous density hidden Markov models. The system has been evaluated on a isolated-word large-vocabulary task. The tests show that Spanish system exhibits a better performance than the Galician system due to its better training. The bilingual system provides an equivalent performance to that achieved by the language specific systems.

1. Introduction

Nowadays, with the distribution of speech technology products all over the world, the demand for automatic speech recognition systems working in multiple languages is growing so much. This fact makes the development of multilingual systems of increasing importance.

To obtain an optimal performance for speech recognition, it is necessary to train the system on large speech databases. Collection of speech databases has no difficulty in getting funds for main languages; the SpeechDat project (www.speechdat.org, 2002) is an example of an effort to collect databases for widespread languages like English, French, Spanish, etc. On the contrary, when a minority or lesser used language is envisaged, no enough funding and support is found to collect and annotate a large database. Without these databases, the development of a feasible speech recognizer for such minority languages is not possible. A suitable approach to cope with this problem is to reuse or to share data with other databases if a partially-common acoustic inventory exists. Usually, multilingual systems combine the phonetic inventory of all languages to be recognized into one global acoustic model set (Waibel et al., 2000). In this way, acoustic models for similar sounds across languages are shared. Those similarities can be derived from international phonemic inventories like Sampa or IPA, which classify sounds based on phonetic knowledge, by data-driven methods, or by a combination of both.

One aim of this work is to develop a bilingual (Spanish and Galician) speech recognizer capable of decoding an utterance spoken in any of both languages. Galician is one minority language spoken mainly in the *Galicia* region of north-west of Spain. This language is spoken by just over two million people in Spain and it coexists with the Spanish. These two languages have a very similar phonological system with a big overlap; the major differences between the two languages are the following:

- There are five vowels in Spanish (/a/, /e/, /i/, /o/, /u/) and seven in Galician (/a/, /e/, /E/, /i/, /o/, /O/, /u/)[1].
- In Galician exists the voiceless palatoalveolar fricative /S/ but in Spanish does not.
- In Spanish exists the voiceless velar fricative /x/ allophone but in Galician this allophone only exists in names, surnames, city names and company names, which correspond actually to Spanish or foreign words.

It is intuitive that ASR performance improves if a great amount of training data is available. More training data implies that the recognition acoustic models can represent the speech more accurately. The work presented in this paper also addresses the problem of limited training data in a minority language (Galician) by developing approaches that reduce the amount of data needed to build an accurate speech recognition system. The extra required data is borrowed from major and resource-rich language (Spanish).

Therefore, the goal of this work is not only to build a multilingual ASR system that be able to recognize accurately several languages, but also to assess at what extent a Spanish database can be used for helping out to a fast development of a Galician ASR.

The organization of this paper is as follows. In section 2, we describe the components of the recognition systems. In section 3, experimental results are presented and discussed. Finally, we draw conclusions in Section 4.

[1] SAMPA symbols. http://www.usc.es/∼ilgas.

2. Speech Databases

To develop the bilingual speech recognition system subsets of two speech databases were used: Spanish SpeechDat (Moreno and Winsky, 1997) and Galician SpeechDat (González-Rey and García-Mateo, 2000). Next a brief description of both speech corpora is given.

2.1. Spanish database

The speech material comes from the Spanish corpus of the SpeechDat database. The utterances were recorded through the public fixed network, sampled at 8 KHz and codified by the A-law using 8 bits per sample.

As training data we have used 10,200 phonetically balanced sentences uttered by 900 speakers. This corpus comprises five hours of continuously spoken speech. This training material is task and speaker independent.

In order to evaluate the speech recognition performance we have designed two tests both based on isolated words. The first one comprises 775 speakers included in training material, and the second one comprises 100 speakers not in training data. The utterances belong to three different tasks: names of speakers, names of cities and phonetically rich words.

Table 1 summarizes the most relevant characteristics of the training and testing sets.

	Number of		
	words	utterances	speakers
training	71,360	10,200	900
test-1	1,531	1,531	775
test-2	616	616	100

Table 1: Characteristics of training and testing Spanish corpora.

2.2. Galician database

The speech material comes from the Galician corpus of the SpeechDat database. As in the Spanish database the utterances were recorded through the public fixed network, sampled at 8 KHz and codified by the A-law using 8 bits per sample.

As training data we have used 3,474 phonetically balanced sentences uttered by 456 speakers. This corpus comprises two hours of continuously spoken speech. This training material is task and speaker independent.

In order to evaluate the speech recognition performance over this language we have also designed two tests both based on isolated words. The first one comprises 300 speakers included in training material, and the second one comprises 100 speakers not in training data. In this case the utterances belong also to three different tasks: names of speakers, names of cities and phonetically rich words.

Table 2 summarizes the most relevant characteristics of the training and testing sets.

As we can see from tables 1 and 2 there is twice as much Spanish than Galician data. We will examine the effect of large or reduced amounts of speech data when the acoustic models of the recognition system are trained.

	Number of		
	words	utterances	speakers
training	6,950	3,474	456
test-1	1,650	1,650	300
test-2	366	366	100

Table 2: Characteristics of training and testing Galician corpora.

3. Acoustic modeling strategies

3.1. Sound inventory

In our work we have defined a global phoneme set based on the SAMPA scheme. Such phoneme set consists of 25 different phonemes plus silence and four noise models for background effects. This phoneme set includes the fricative /S/ which does not exist in Spanish words, and the fricative /x/ which does not exist in Galician words but that has been artificially added to the Galician training database through names and surnames.

Furthermore, in this phoneme set some allophonic variations are merged in an unique phone. Thus, for both languages:

- /D/ and /d/ are merged in /d/
- /B/ and /b/ are merged in /b/
- /G/ and /g/ are merged in /g/

In addition for the Galician language the following merging strategies are also made:

- /E/ and /e/ are merged in /e/
- /O/ and /o/ are merged in /o/

The main reason behind this merging strategy is the difficulty to guarantee a correct phonetic transcription from the written sentences when any of these allophones is in the utterance. In many cases, the phonetic transcription is speaker or dialect dependent, so to make a hard a-priori decision about the phonetic transcription is not recommended.

3.2. Recognition system

For experimental work the HTK v2.2 recognition system (Young et al., 1999) was used. The recognizer makes use of continuous density hidden Markov models (CDHMM) with Gaussian mixture for acoustic modeling. The acoustic units are demiphones (Mariño et al., 2000), i.e., context dependent units.

Each demiphone consists of a fully continuous density 2-state HMM. Each HMM-state is modelled by a mixture of 4 Gaussian distributions with a 39 dimensional feature space: 12 mel-frequency cepstrums (MFCC), log-energy, and their first and second time derivatives.

For the Spanish tests a lexicon of 15,819 words was considered, and for the Galician tests the vocabulary size was 9,183. For some words we have included pronunciation variations. The multiple pronunciations of a word are

obtained through rules. As the envisaged task is isolated-word based, no language model is needed and a simple grammar which consists of all the words in parallel is used.

Table 3 shows the number of *significant* demiphones that appear in the training corpus of each language. Significant demiphones were defined as those demiphones that appear at least 50 times in the training data.

	Spanish	Galician
training	562	571

Table 3: Total number of different demiphones for each training corpus.

3.3. Training of the acoustic models

Acoustic models are built in a series of steps. A first set of seed models is used to segment and label the training data using Viterbi alignment of the text transcription and a lexicon containing one or more pronunciations per word. The chosen phone sequence and segmentation are then used to construct a set of context dependent demiphones.

First, for a baseline recognition system we have developed two monolingual systems: one for Spanish an another for Galician (*Galician-1*), based on their specific training data.

As we could see above the Galician language has a limited amount of training data. Therefore, we have also examined what performance can be achieved when the Galician monolingual system is built from the Spanish system using MLLR adaptation (Legetter and Woodland, 1995) (*Galician-2*). We have applied supervised MLLR adaptation with a regression class tree with 20 terminal or leaf nodes. The regression class tree was constructed so as to cluster together components that are close in the acoustic space.

For multilingual speech recognition we wish to combine acoustic models of similar sounds across languages into one multilingual phoneme set. Based on the 25 phonemes in the global set we train two different multilingual systems that differ basically in the HMMs topology. In the first one (*Multilingual-1*) each acoustic-phonetic unit is first trained separately for each language. Then, the phonetic units that are present in both languages are modelled in the multilingual system with a topology as the shown in Figure 1. That is, the acoustic unit is modeled by a HMM with four states, two states belong to the Spanish acoustic unit (e.g. the top branch) and the other two belong to the Galician acoustic unit (e.g. the bottom branch). Acoustic units that are modelled by only one of the languages preserve the original demiphone-HMM topology.

In the second multilingual system (*Multilingual-2*) acoustic models are trained from a corpus composed gathering the available corpus of both languages, i.e., no language information is preserved in the system training.

4. Experimental Results

This section gives results for the monolingual, multilingual, and crosslingual tests based on the above systems.

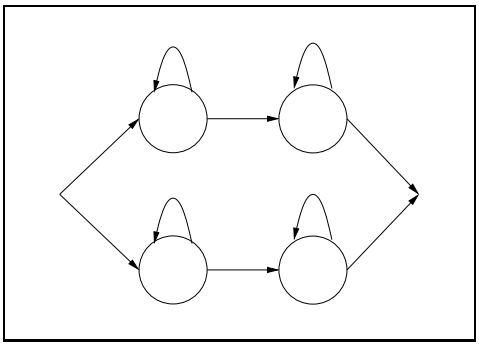

Figure 1: Hidden Markov model topology corresponding to a phonetic unit which is modelled in both languages in the Multilingual-1 system.

Despite the good coverage of contexts that demiphone provides, the problem of unseen units during the training is usually present. In this work the unseen demiphones that are contained in test vocabularies are mapped to trained demiphones through a set of rules designed from phonetic knowledge. Table 4 shows the number of unseen demiphones in each recognition test.

Recognition System	Target Language	
	Spanish	Galician
Spanish	182	175
Galician-1	220	164
Multilingual-1	158	108
Multilingual-2	146	97

Table 4: Number of unseen units in the training material.

Table 5 summarizes the performance of the analyzed recognition systems. A first glance at this table shows that the monolingual Spanish system outperforms the two monolingual Galician systems. The larger size of the training material for the Spanish language reasonably accounts for this result. Therefore we can conclude that the monolingual Galician systems are not well trained and that increasing the training data will lead to significant improvements in the recognition performance.

Table 5 also shows the crosslingual recognition tests on both languages. We observe the drastic performance decrease of the Spanish language when it is recognized with the Galician-1 acoustic models. Nevertheless, when the Spanish models are used to recognize the Galician language the recognition performance is only slightly decreased. Again, this fact shows that the monolingual Galician system has insufficient training material.

With regard to Galician-2 system we can say that MLLR adaptation of the Spanish acoustic models leads to similar recognition performance to that of the system training through bootstrapping. The advantage of MLLR adaptation is the training speed.

Finally, the recognition results obtained by the two multilingual systems show that both training methods for the Spanish language can be considered equivalent to the corresponding monolingual system. This corroborates the fact

that the Spanish language is accurately considered in all the trained systems. However, when the Galician language is concerned only a slight improvement is obtained and the scoring provided is worse than the Spanish one, this can be due to the unbalanced amounts of training material.

Recognition System	Target Language	
	Spanish	Galician
Spanish	86.10	78.05
Galician-1	78.97	79.90
Galician-2	–	78.99
Multilingual-1	85.69	80.70
Multilingual-2	85.97	80.58

Table 5: Word accuracy obtained with the different recognition systems.

5. Conclusions and further work

In this paper, monolingual and multilingual speech recognition systems are presented which can handle two languages namely Spanish and Galician. The Galician language is a minority one, then it copes with a small amount of speech and language resources. We have examined the effect of limited amount of training data.

Results show that the amount of training data is a fundamental factor to develop a accurate speech recognition system. Furthermore, it is possible to partially avoid the lack of a sufficiently large database by using a speech database available in another but phonetically similar language.

With regard to the multilingual system, results show that the performance for the Spanish is kept and for the Galician one is slightly increased.

In future experiments, we will investigate various procedures of clustering to overcome the problem of unseen phonetic units during training. Such clustering of models will also allow to share model parameters that are close in the acoustic space decreasing so the complexity of the system.

Acknowledgements

This work has been partially supported by Spanish Ministry of Science and Technology through projects TIC2000-1005-C03-02 and TIC2000-1104-C02-01.

6. References

B. González-Rey and C. García-Mateo. 2000. Diseo de una base de datos speechdat para el idioma gallego. *Procesamiento del Lenguaje Natural (SEPLN)*, (24):197–204, september.

C.J. Legetter and P.C. Woodland. 1995. Flexible speaker adaptation using maximum likelihood linear regression. In *Proceedings of the Spoken Language Systems Technology Workshop*, pages 110–115.

J.B. Mariño, A. Nogueiras, P. Pachs, and A. Bonafonte. 2000. The demiphone: an efficient contextual subword unit for continuous speech recognition. *Speech Communication*, 32(3):187–197, October.

A. Moreno and R. Winsky. 1997. Spanish fixed network speech corpus. *SpeechDat Project LRE-63314*.

A. Waibel, P. Geutner, L. Mayfield Tomokiyo, T. Schultz, and M. Woszczyna. 2000. Multilinguality in speech and spoken language. *Proceedings of the IEEE*, 88(8):1297–1313, August.

www.speechdat.org. 2002. Speechdat project.

S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valttchev, and P. Woodland. 1999. *The HTK Book (for HTK Version 2.2)*. Cambridge University Press.