Multidialectal Spanish Modeling for ASR

Mónica Caballero, José B. Mariño, Asunción Moreno

Department of Signal Theory and Communications
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
Universitat Politécnica de Catalunya (UPC)
Jordi Girona 1-3, Barcelona, Spain
{monica,canton,asuncion}@gps.tsc.upc.es

Abstract

This paper describes the latest advances in our ongoing work in the area of Spanish multidialectal speech recognition. This work deals with the suitability of using a single multidialectal acoustic modeling for all the Spanish variants spoken in Europe and Latin America. The objective is two fold. First, it allows to use all the available databases to jointly train and improve the same system. It also allows to use a single system for all the Spanish speakers. Our latest experiments consist of the optimization of the acoustic models applying a top-down bottom-up hybrid clustering algorithm. Overall multidialectal acoustic modeling leads to maintain the performance of the recognition system even when it's tested with an unseen dialect, that is, not seen in the training process.

1. Introduction

With more than 300.000 Million speakers worldwide, Spanish is considered as one of the global languages in the world. Spanish is also one of the most widely spoken languages; it's spoken in Europe and in America from US to the Antarctica. There are many factors that prompted the apparition of dialectal variants, the geographical dispersion, the existence of native languages in the Latin American countries where Spanish is spoken, the evolution of the language in each country along the time, etc. The dialectal variants have been described in the literature and include phonetic, lexical, semantic and cultural variations among others.

The lack of suited databases to properly train ASR in each Latin American country is being overcome. Adding to the former VAHA or CALL HOME databases available in the LDC[1], the SpeechDat Across Latin America (SALA) (Moreno, 1998) project has developed a set of telephone databases in most of the countries in Latin America allowing to train ASR systems. The SALA project was born with the objective of solving the lack of databases in Latin America using the specifications of SpeechDat (Höge, 1997) project that is considered close to a standard.

This paper deals with the suitability of using a single multidialectal acoustic modeling for all the Spanish variants spoken in Europe and Latin America. The experiments that are reported in this paper are designed to use all the available databases both to jointly train and improve a single system and to be able to use it for all the Spanish speakers.

Focusing on the importance of dialectal variants from the Automatic Speech Recognition (ASR) point of view, Moreno (1998) tries to cluster wide geographic areas or countries where phonetic similarities can be useful to have a unique ASR system, and gives the general rules to describe such phonetic differences. The dialectal variants chosen in this paper are: Spanish as spoken in Spain, Colombia, Venezuela (Caribbean) and Argentina. All the Spanish variants are considered different from the linguistic point of view and representative of a wide number of Spanish speakers.

This paper is organized as follows. Section 2 describes the databases used in the experiments, section 3 shows how is made the phonetic transcription, section 4 describes the recognition system and the clustering algorithm applied to improve the system, and sections 5 and 6 describes the experiments and results.

2. Databases

In this work, the databases used for training and testing the ASR systems belongs to the SpeechDat and SALA projects.

The database of Spanish as spoken in Spain was created in the framework of the SpeechDat project. The database consists of fixed network telephone recordings from 4000 different speakers. Signals were sampled and recorded from an ISDN line at 8KHz, 8 bits and coded with A law. Each speaker utters 45 sentences either reading or answering some questions. Speakers were selected to have a broad coverage of ages and dialectal distribution in the country. In this work, 3500 speakers were selected for training purposes and 200 speakers for testing.

The databases of Spanish as spoken in the different dialectal variants of Latin America were created in the SALA project. Each database consists of fixed network telephone recordings from 1000 different speakers. Signals from the Colombian and Argentinean databases were recorded and stored from an ISDN line at 8KHz, 8 bits and A law coded. The Venezuelan database was recorded from an analogue telephone line and stored with μ law code. The criteria for speaker selection follow the SpeechDat project. 800 speakers from each database were selected for training and 200 speakers for testing.

[1] http://www.ldc.upenn.edu

3. Transcription

The phonetic transcription used in this project is done in SAMPA symbols. For each dialectal variant considered in this project, a canonical phonetic transcription was defined and implemented. The canonical phonetic transcription of each dialectal variant is defined as the most common variant, in number of speakers, of a given country or area. For the Caribbean variant, the variant spoken in Caracas was chosen; for Colombia, the dialectal variant spoken in Bogota was considered as canonical, and for the Argentinean variant, the canonical variant was chosen as the Spanish as spoken in Buenos Aires.

Transcriptions are obtained in an automatic form, each dialect with its specific transcription. The rule-based phonetic transcription is based on a previous work (Moreno, 1998) and modifies the canonical phonetic transcription of the Spanish as spoken in Spain (Mariño, 1993).

Table 1 shows the set of SAMPA symbols used in this project. The dark column indicates the specific symbols for one or more dialects but not common in the complete Spanish set.

SPAIN	a B b D d e f G g i j J k l m n N o p R r rr s t tS u w	jj x T
COLOMBIA	a B b D d e f G g i j J k l m n N o p R r rr s t tS u w	jj h
VENEZUELA	a B b D d e f G g i j J k l m n N o p R r rr s t tS u w	jj h
ARGENTINA	a B b D d e f G g i j J k l m n N o p R r rr s t tS u w	Z x h

Table 1. Phonetic inventories of the four considered dialects. Dark column shows the dialect-specific phones

4. Recognition System

4.1. System Description

This work was implemented in a recognition system developed at Universitat Politècnica de Catalunya, Spain, called RAMSES (Mariño, 2000).

The system is based on Semicontinous Hidden Marcov Models (SCHMM). Speech signals are parameterized with Mel-cestrum and each frame is represented by their Cepstrum C, their derivatives ΔC, $\Delta\Delta C$ and the derivative of the Energy. The three first features are represented by 512 gaussians and the Energy derivative by 128 gaussians. The phonetic units for this task are demiphones. Each phonetic unit is modeled by a 2 states left to right model.

4.2. Acoustic Modeling

Two different kinds of systems are modeled, monodialectal systems and a multidialectal system. Monodialectal systems were trained with data either from Spain, Colombia or Venezuela with its own phonetic inventory and phonetic transcription. Some previous experiments were performed to choose a suited set of phonemes and allophones. The phonetic inventory for each monodialectal is described in Table 1.

Two different systems are created for Spanish as spoken in Spain, one is trained with approximately the same amount of data that the other dialects and another using all the available data. The purpose is to distinguish the effect of increasing the number of speakers when comparing the results of the monodialectal systems. Latin-American systems are trained with the available data for each dialect.

The multidialectal system was created with a single phonetic inventory, that is, all the dialects share the same acoustic models. The idea behind is to include similar sounds in a single model having all the possible realizations in it. In this way the number of utterances to train a given phoneme is increased since they share all the available data. Moreover the system is more robust to transcription errors possibly due to generalization of rules over a big number of speakers.

Models are created from training material from all the dialects and the maximum number of speakers and utterances. The mapping of the phonetic units is done directly from their SAMPA representation, that is, sounds of different dialects which are represented by the same SAMPA symbol share common phoneme in the final inventory. These models form the biggest part of the inventory. Dialectal specific phones are trained with their dialect specific database only and are added to the global inventory. In terms of phones, the multidialectal phonetic inventory is composed by 32 phones. 30 phones are trained with data from all the dialects, the phone /h/ is trained with data from Colombia and Venezuela and the phone /T/ from Spanish as spoken in Spain is trained from data from that country only.

Information about cross-dialectals phoneme mappings can be found in (Nogueiras, 2002).

The phonetic unit considered for acoustic modeling is the demiphone, a contextual phonetic unit that models a half of a phoneme.

In a first approach demiphones defined by threshold are used. In this case a context dependent demiphone model is created if the number of training realizations in the training set for that particular unit is higher than a threshold. Otherwise, a context independent demiphone trained with all possible contexts is used. The threshold in the recognition system was set to 100 realizations.

A clustering procedure is applied in order to optimize the multidialectal acoustic models, obtaining a total contextual coverage, even for unseen units in train data and a smoothed training procedure. In order to obtain clustered demiphones a top-down bottom-up hybrid clustering algorithm is used. This procedure works in two steps. Firstly, for every phoneme a decision tree is learnt to classify the demiphones. During the recognition phase, this tree will be used to group the unseen units along with the units available during training. The clustering algorithm goes on by the agglomerative algorithm that gives the final cluster configuration. These clusters give a better representation for the units that in the previous experiments were modeled by a context independent demiphone.

Silence and a noise models are also used. In the multidialectal system, language-dependent silence models also are introduced. Experiments showed that the use of four models of silence, three of them trained with data from each dialect, improved results.

5. Experiments

5.1. Training and testing data

The systems are trained with a set of phonetically rich sentences and phonetically rich words. Only clean data was used, avoiding material with intermittent noise and mispronounced words and truncations.

The recognition systems were evaluated in two different tasks. One consists of isolated application words (i.e. delete, enter, etc.) plus isolated digits (from zero to nine). The other consists of the recognition of six connected digits.

Specific training and testing data is exposed below:

- *Train*. Five different training sets are defined. Two of them correspond to the dialect spoken in Spain, one including only utterances from 800 speakers and other with all the available data. Two more training sets are designed for dialects spoken in Colombia and Venezuela. Finally a combined set of all of the dialects is created.

- *Test*. Four tests, one for every considered dialect, are created for each task. In the first task, named application words and isolated digits (AW & ID), there are 40 possible different words to be recognized (30 application words plus the ten digits). Even thought application words are not totally equal across the dialects, most of them are very similar and there is no complexity added in any of them. The second task, connected digits (CD), consists of six connected digits.

The total number of utterances for each training set and test set is summarized in Table 3. The combined training set is composed by the sum of all the training utterances.

DIALECT	#training utterance	# test AW & D utterances	# test CD utterances
SPAIN 800 sp.	6.698	-----	-----
SPAIN	24.330	625	270
COLOMBIA	4.246	751	149
VENEZUELA	6.649	1.048	247
ARGENTINA	--------	1.160	304

Table 3. Training and testing material for the systems implemented

5.2. Experimental Systems

Six different systems were created. Four of them are monodialectal, that is, each one belongs only to a single dialect (Spanish as spoken in Spain, Colombian and Venezuelan). The remaining are the two multidialectal system approaches, trained simultaneously with these three dialects. Argentinean dialect is not included at all in the training process.

Table 4 shows the number of models finally created and how many of these models are context dependent or context independent. Multidialectal (1) refers to the first approximation using demiphones defined by a threshold. Multidialectal (2) refers to the system where the clustering algorithm is used. It can be observed that this last system reduces the number of models to train, decreasing as well the number of parameters to estimate.

SYSTEM	#models	#CD	#CI
SPAIN 800 l.	607	545	62
SPAIN	786	724	62
COLOMBIA	476	414	62
VENEZUELA	568	510	58
MULTID. (1)	862	798	64
MULTID. (2)	825	825	0

Table 4. Number of models in all the ASR systems created

Table 5 shows the percentage of context independent models for each one of the tests and tasks using the multidialectal demiphones defined by threshold. Even there are no high percentages, it points out that even in simple tasks there are situations not well modeled. This fact justifies the application of the clustering algorithm.

Task / Dialectal tests	Application Word Isolated Digits	Connected Digits
SPAIN	0,32	0,00
COLOMBIA	0,84	0,00
VENEZUELA	0,91	2,21
ARGENTINA	0,46	0,00

Table 5. Percentages of context dependent models in the tests used to evaluate the systems

Experiments have been designed taken into account three main points of interest:

- Performance of monodialectal systems in order to compare them with multidialectal approaches.
- Cross-dialectal recognitions in order to observe similarities between dialects and databases. Specific interest is in recognizing Latin-American dialects by means of the Spanish monodialectal recognizers. Those recognitions show separately the effect of the similarity of the dialects and the effect of increasing the number of speakers in the training process.
- Performance of the multidialectal systems recognizing all the dialects.

6. Results

Results of the experiments are summarized in Tables 6, 7 and 8. Table 6 shows the results of the task of application words and digits. In Table 7 the results of the connected digits task are represented, and finally, Table 8 shows the results for the improved multidialectal system for both tasks. All the results presented in this section are expressed in terms of % WER (Word Error Rate).

Train / Test	MULTI	SPAIN	SPAIN 800 sp.	CO	VE
SPAIN	1,1	1,4	1,4	2,5	2.1
CO	1,6	2,2	2,3	2,4	2,3
VE	1,3	2,3	2,5	2,4	2,2
AR	1,6	1,8	2,3	2,8	2,3

Table 6. Results for the task "Application Words and Isolated Digits"

Train / Test	MULTI	SPAIN	SPAIN 800 sp.	CO	VE
SPAIN	0,5	0,7	1	2,5	1,5
CO	2,1	3,3	3,6	3,9	3,1
VE	2,8	3	4,3	5,6	3,4
AR	1	2,1	2,5	3,2	3,1

Table 7. Results for the task "Connected Digits"

Task / Test	Application W & I. Digits	Connected Digits
SPAIN	1,1	0,5
COLOMBIA	1,5	2,1
VENEZUELA	1,1	2,1
ARGENTINA	1,6	1,1

Table 8. Results obtained with the multidialectal system using clustered demiphones

Looking the performance of monodialectal systems it's clear that the best results are given by the system from Spain, even when it's trained with the same number of speakers that the others dialects. Colombian system is affected by the poor training data available. Reasons for higher error rates with the Venezuelan system comparing with the Spanish have been investigated, measures have been taken in order to check that all the databases have the same quality in terms of signal to noise ratio, but at present no justification have been found.

Cross-dialects results show that the systems are nearly equivalent with little differences in performance and with better results depending on the number of speakers. Increasing the number of speakers in the Spanish system improves the results obtained with only 800 speakers. Apparently, there are no deviations due to the different recording system used in Venezuela (analogue lines, μ law). Looking at the recognition results of the system trained with data from "Spain", it seems that Argentinean is closer to the Spanish spoken in Spain than to other Latin American countries. This agrees with the classical separation of the Latin American dialects.

Multidialectal results outperforms all the monodialectal and cross-dialectal results in both tasks for all the dialects. In the cases that tests contains words modeled by with context dependent models, multidialectal system using clustered demiphones improves the recognition rate. That reduction in the WER is higher for that tests with higher percentage of context dependent models, as it happens with the Venezuelan tests. Table 9 shows percentages of improvement of monodialectal performance using the final multidialectal system.

Task / Test	Application W & I. Digits	Connected Digits
SPAIN	22 %	20 %
COLOMBIA	38,7 %	45,5 %
VENEZUELA	47,7 %	39,2 %
ARGENTINA	9,9 %	46 %

Table 9. Percentages of improvement from monodialectal to multidialectal systems.

7. Conclusions

Cross-dialect experiments show the necessity of joining efforts among different databases to improve the recognition results.

Multidialectal system can work as a single system to cope with all the Spanish dialects considered and provides better performance in all the evaluated cases. Clustering algorithm gives a better modeling and improves the results obtained with demiphones defined by threshold.

The performed experiments validate the automatic rule-based Spanish phonetic transcription. Currently authors are working in improving the automatic transcription taken into account aspects like sounds that can differ between dialects and are represented by the same SAMPA symbol, position of a sound in the word and stress.

8. Acknolowgedments

This work was supported by the Spanish Government grant number TIC2000-1005-C03-01.

9. References

Höge H. et al. (1997) European Speech Databases for Telephone Applications *Proc. Int. Conf. On Acoustics, Speech and Signal Processing*. ICASSP'97.

Llisterri J., J.B. Mariño (1993) Spanish adaptation of SAMPA and automatic phonetic transcription, *Report SAM-A/UPC/001/V1*.

Mariño J.B., A. Nogueiras, P. Pachès, A. Bonafonte (2000). The demiphone: an efficient contextual subword unit for continuous speech recognition. *Speech Communication*, Vol. 32, No. 3, (pp. 187-197).

Moreno A., H. Höge, J. Köhler, J. B. Mariño (1998). SpeechDat Across Latin America Project SALA. *Proc. First Int. Conf. On Language Resources & Evaluation,* ICLR'98.

Moreno A., J.B. Mariño, Spanish dialects: Phonetic transcription. *In Proceedings of International Conference on Spoken Language Processing,* ICSLP'98. Sidney, Australia.

Nogueiras A., M. Caballero, A. Moreno (2002). Multidialectal Spanish Speech Recognition. To Appear in *Proceedings of International Conference on Acoustics Speech and Signal Processing*, ICASPP 2002, Orlando, USA.

Strick H., C. Cucchiarini (1999). Modeling pronuntiation variation for ASR: A survey of the literature. *Speech Communication,* Vol 29,(pp 225-246).