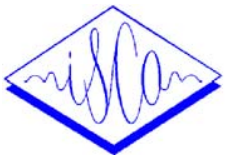
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Data Driven Multidialectal Phone Set for Spanish Dialects

Mónica Caballero, Asunción Moreno, Albino Nogueiras

Department of Signal Theory and Communications
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
Universitat Politècnica de Catalunya (UPC), Spain
{monica, asuncion, albino}@gps.tsc.upc.es

10.21437/Interspeech.2004-309

Abstract

This paper addresses the use of a data-driven approach to determine a multidialectal phone set for an automatic speech recognition system for Spanish dialects. This approach is based on a decision tree clustering algorithm that tries to cluster contextual units of different dialects. This procedure avoids the definition of a global phonetic inventory and the previous study of similarity of sounds.

The procedure is applied in Spanish as spoken in Spain, Colombia and Venezuela. Results show differences between phonemes that share the same SAMPA symbol in different dialects and also detect similarities between phonemes that are represented by different symbols in dialectal variants. Recognition results using this multidialectal approach overcome the monodialectal ones.

1. Introduction

Spanish is a global language spoken in many countries all over the world with several dialectal variants. A common approach to share and take the maximum advantage of the available data from all the considered dialects is to build a multidialectal automatic speech recognition system with a single set of models. Considering dialects as different languages, the problem of building a common acoustic unit set can be solved with multilingual acoustic modeling techniques, sharing data from similar sounds across dialects. Similarity between sounds can be found in an expert or data driven manner. Expert driven methods are based on linguistic knowledge. The main approach used is the IPA/SAMPA scheme, where sounds of different dialects or languages that share the same symbol are considered the same phone unit in the system [1] [2]. Data driven methods measure the similarity directly on the data and have been shown to give better results. An example can be found in [3] where there is a comparison between results obtained with phone mapping done by an expert based on the similarity of the phonological features and those obtained with an automatic process based in the creation of a confusion matrix to find the final phoneme inventory.

Working with contextual units, similarity can be defined at the phone level [2] or at the level of the contextual unit [4]. In the first case two steps are necessary: the definition of a multidialectal phone set and the extension to contextual units, which are usually modeled applying decision trees. When measuring similarity directly on contextual units, clustering algorithms are used to join contextual units across dialects.

This work aims to create a multidialectal system for the dialects of Spanish,considering only their phonetical differences. In this paper Spanish as spoken in Spain, Colombia and Venezuela are the dialects considered. In a previous work [6], a multidialectal phone set was defined based on SAMPA scheme. The resulting clusters were very conditioned for the definition of the phonetic set and the transcription rules for each dialect.

This paper presents a new approach to find the multidialectal models using a decision tree based clustering algorithm which evaluates similarity directly on contextual models using a entropy-based measure and adds questions about the dialect. Decision trees usually constrain which models can be joined, such as models sharing the same central phonetic unit [1] [2] [5] or models belonging to the same broad phonetic group (fricatives, stops, etc.) [4]. The proposed structure does not constrain which models can be clustered, avoiding the definition of a global phonetic set.

The resulting tree is analyzed from a linguistic point of view, trying to detect in a data-driven manner the main phonetic differences and similarities of the Spanish dialects.

This paper is organized as follows: section 2 describes the transcription system, section 3 describes the basis of the recognition system and the decision tree, section 4 explains the experiments performed and finally sections 5 and 6 show the results and the conclusions of this work.

2. Transcription of Spanish dialects

Spanish as spoken in Spain, Colombia and Venezuela are transcribed in SAMPA symbols. Standard SAMPA symbol set for Spanish [www.sampa.ucl.uk] is extended with /h/ [7] to cope with the Colombian and Venezuelan dialects. Transcriptions are obtained in an automatic form, each dialect with its specific transcription. Details can be found in [8]. Some transcription rules are specifically studied in this work: those that differ between dialects and transcript different graphemes into different phonemes in the Spanish dialect but into the same phoneme in the dialectal variants.

- Graphemes [j] and [g] followed by either [e] or [i] are transcribed in Spain as /x/; grapheme [s] in post-vocalic position is transcribed in Spain as /s/. In Venezuela both are transcribed as /h/.

- Graphemes [s] and [z] are transcribed in Spain as /s/ and /T/ respectively. Grapheme [c] followed by [e] or [i] is transcribed in Spain as /T/. In Colombia and Venezuela those are transcribed as /s/ (except [s] in post-vocalic position in Venezuela as was explained above).

To study those special cases, transcription has been modified, identifying these Colombian and Venezuelan phones with a tag as is shown in Table 1. Additionally, phonemes /l/ and /r/ when they belong to a consonant group were specifically tagged with /_CG/. The rhotic phoneme /R/ is added to represent post-vocalic [r].

Graphemes	Spain	CO	VE	Tag. CO	Tag. VE
[j], [g]+[e], [g]+[i]	/x/	/h/	/h/	/h/	/h/
Post-vocalic [s]	/s/	/s/	/h/	/s/	/s_h/
[s]	/s/	/s/	/s/	/s/	/s/
[z],[c]+[e], [c]+[i]	/T/	/s/	/s/	/s_CV/	/s_CV/

Table 1: *Modification in the transcription of Colombian and Venezuelan dialects*

3. System description

This work was implemented in a recognition system developed at Universitat Politècnica de Catalunya, Spain. The system is based on Semicontinuous Hidden Marcov Models (SCHMM). Speech signals are parameterized with Mel-Cepstrum 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, a contextual unit that models the half of a phoneme. Each phonetic unit is modeled by a 2 states left to right model. Dialectal dependent acoustic models are created in order to allow dialectal models to be joined or kept separated. Those models are trained with material from its specific dialect and are tagged with its dialect to be able to distinguish models with the same name.

3.1. Decision tree

In the phonetic decision tree, each leaf node is a cluster of acoustic models and the branches represent questions relevant to the attributes of the models. Compound questions are formed by joining yes/no questions in an 'OR' manner. In every node the entropy is computed. According to answers to the questions, the different acoustic models that stay in a node are split into child nodes. The sum of the entropy of the two child nodes should be lower than the one computed for the parent node. Each question of the tree is selected in order to maximize the gain of entropy and gives the best split of the node.

The entropy of node A is calculated with the expression (1) where M is the number of models in the node, S is the number of states of each model, G is the number of gaussians in the codebook, $f(m)$ is the frequency of the model in the train data, $f(s|m)$ is the quotient between the frames of the state s and the total number of frames of the model the state belongs to and b's are the observation symbol probabilities.

$$H(A) = \sum_{m=1}^{M} f(m) \left[\sum_{s=1}^{S} f(s|m) \sum_{g=1}^{G} b_{sg} \log b_{sg} \right] \qquad (1)$$

Stopping splitting criteria is defined by a minimum increase of likelihood and/or a threshold of number of realizations in each cluster or leaf node.

Question set contains questions related to phonetic features of the models (i.e. is the context a vowel?) and one non-phonetic question, the dialect of the unit. This last question refers to the entire demiphone. Phonetic questions comprise questions about the type, the place and the manner of articulation. Also, questions related to special issues as if the phone is an aspiration or if the phonemes are part of a consonant group. In order to get the multidialectal phone set, the acoustic models that are used to build the tree are language dependent, trained with data from each dialect. Two different tree structures are proposed:

The first structure is the most common in the literature. A different tree is built for each unit of the phonetic inventory. The phonetic inventory is defined as the sum of the SAMPA phones necessary for the transcription of each dialect. Allophonic variations are not considered in root nodes. With this structure the tree is able to cluster allophones and contexts only. This approach is referred to as Tree_1.

The second structure is less restrictive. There's only one tree for all the units allowing clustering of contexts and central units over all dialects. This approach is referred to as Tree_2.

4. Experiments

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. Speakers were selected to have a broad coverage of ages and sexes in the country. In this work, 3500 speakers were selected for training and 200 for test. The databases of Spanish as spoken in the different dialectal variants of Latin America were created in the SALA [7] project. Each database consists of fixed network telephone recordings from 1000 different speakers, 800 speakers were selected for training and 200 for test. Phonetically rich words and sentences are used to train the systems. Tests are composed of application words, strings of digits and phonetically rich

words. The total data available for training and test is presented in Table 2.

DIALECT	#training utterances	#test utterances
SPAIN	24,330	1,472
COLOMBIA	4,246	1,273
VENEZUELA	6,649	1,822

Table 2: *Training and test material for each dialect*

4.1. Experimental systems

In order to study the broad phonetic classes achieved automatically and the similarity between same phones across dialects, a decision tree that clusters monophones is built. Question set contains linguistic questions and also ask about the dialect and no restrictions avoiding clusters are applied.

Different recognition systems are created. First of all, monodialectal systems are created in order to compare their results with the multidialectal approaches. A decision tree based clustering is applied in order to smooth the final acoustic models. The total number of models for each system can be found in Table 3.

	SPAIN	COLOMBIA	VENEZUELA
No Models	757	518	595

Table 3: *Number of models for each dialectal ASR*

Four multidialectal systems are created. For each tree structure, i.e. Tree_1 and Tree_2, two sets of acoustic models are trained. One set is obtained pruning the tree to get a reduction of models of approximately 36 % over the sum of the number of models of each monodialectal system. That percentage is achieved with 800 models. The second set is composed by 1200 models, which means a reduction of 58%. The threshold of number of realizations in each leaf is set to 100.

5. Results

5.1. Linguistic results

Linguistic results are obtained analyzing the monophone tree and the Tree_2. The monophone tree separates in big groups the Spanish phonemes, according to the linguistic theory. The resultant structure is shown in figure 1. Dialect question normally appears at last positions. The first split is made between vowels (including semivowels) and consonants. In the consonant branch, the groups detected are: nasals, lateral liquids, voiced and unvoiced stops, rhotics, approximants, and fricatives. Building the tree with models trained with the same amount of data from each dialect, Venezuelan phones are separated from the rest. This can be due to the different recording conditions. When models are trained with the total available data, Spanish dialect is separated, possibly due to

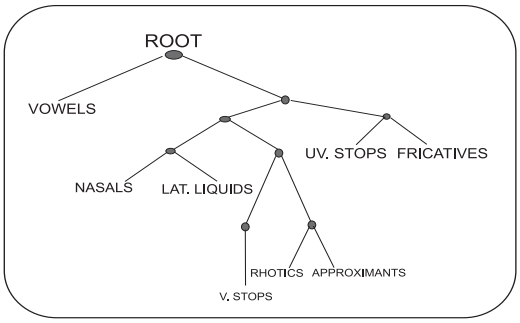


Figure 1: *Monophone tree structure.*

the large amount of data from this specific dialect.

When asking for contexts, the tree structure mostly remains the same, but there are exceptions. Every group will be commented separately:

Nasals. Nasal phonemes of Venezuelan dialect are separated from the rest immediately. This indicates that those realizations are really different from the ones in Spain and Colombia. Specific dialectal contexts (/h/) produce links between Venezuelan and Colombian models. The different nasals phonemes are kept separated, except when the contexts define different transcription over the dialects i.e. clusters having models *VE_N+b*,*ES_m+b*,*CO_m+b* can be found in the tree. Possibly the sounds are very similar, but transcription rules give a different symbol for Venezuelan dialect.

Liquids. Liquid phonemes are divided in two groups, laterals and rhotics. In both groups the phones /l/ and /r/ are separated from /l_GC/ and /r_GC/ at higher level. That shows that liquid phones in consonant group position appear to be nearly different allophones. Phoneme /R/ is separated immediately from the others phones. That shows the necessity of using one exclusive symbol to represent post-vocalic [r].

Stops. A main division is made between the voiced and unvoiced stops phones. Unvoiced stops are the first to be separated from the rest of phonemes. Voiced stops are first divided depending if they are approximant or not instead of dividing them taking into account the place of articulation (bilabial, velar or dental). Colombian approximants are separated very early from the Spanish and Venezuelan ones. This fact can be explained for the different transcription of graphemes [b], [d] and [g] in Colombia.

Fricatives. The monophone tree allows to distinguish three big groups: the alveolar phoneme /s/; the velar and glottal phonemes /h/, /x/ and Venezuelan /s_h/; and the last group with labiodental /f/, interdental /T/ and /s_CV/. Dialect question appears after separating the different units. On the contrary, in the demiphone tree, models are nearly always splitted depending on the dialect and after that, phonemes are separated. Spanish models are the first to be separated. Looking with more detail:
The /s/ of the Spanish dialect gets apart from the Colombian and Venezuelan ones very early. That behaviour agrees with the linguistic definition of the apical realization of the

phoneme /s/ in Spain with respect to the sibilant realization in Latin-American countries.
The Spanish velar phoneme /x/ and the American glottal /h/ are kept together most of the time, showing that they are not so different. Venezuelan phone /s_h/ is in the group of velar/glottal fricatives, not in the branch of /s/, but stays in a different cluster.
Phonemes /f/, /T/ and /s_CV/ constitute the last group. /f/ is separated from the other two and it is clear that it has to be considered different. In monophone tree, both Colombian and Venezuelan phones /s_CV/ are associated with the Spanish /T/. To interpret this result it has to be taken into account that there is a possibility of transcription errors with speakers of the South of Spain, who most of the times pronounce /T/ as Latin-American speakers. In the demiphone tree, as a consequence of the early apparition of the dialect question, /s_CV/ is in the same branch as /s/ for each Latin-American dialect.

Affricates. This case is special due to the system that has been used, so the group where the phoneme /tS/ can be found depends if it is the right or left part of the phoneme. /tS/ is combination of two manners of articulation, meaning that it consists of a stop immediately followed by a fricative at the same place of articulation. This linguistic definition of the phoneme is what has been observed in the clusters obtained. When looking at the left part of the phoneme, /tS/ stays with the phoneme /t/. When looking at the right part, it stays in the group of the phonemes /f/, /T/ and /s_CV/. Even linguistic knowledge says that /tS/ is pronounced more occlusive in Spain, there is no an early separation of the dialectal phonemes in this case.

5.2. Recognition results

Table 4 shows recognition results. The table summarizes the results of the three monodialectal systems, and the results obtained for each of the multidialectal models sets for both tree structures approaches shown in section 3. Results obtained in the previous work with the system based in SAMPA scheme and no model smoothing applied are also presented (Multi SAMPA S.).

This multidialectal approach outperforms previous results mapping SAMPA scheme [6]. All systems give similar results. Models defined by the structure 2 - no restrictions are applied to the tree-, give the better results, improving the recognition rate in all the dialects.

WER (%)	**SP**	**CO**	**VE**
Monodialectal S.	97.55	96.06	96.62
Multi SAMPA S.	97.16	95.44	94.96
Tree_1 800 S.	97.20	96.35	95.92
Tree_1 1200 S.	97.39	96.00	96.29
Tree_2 800 S.	97.40	96.35	96.61
Tree_2 1200 S.	**97.66**	**96.70**	**96.78**

Table 4: *Recognition rates*

6. Conclusions

This data-driven approach allows analyzing and detecting similarities and differences between phones over the considered Spanish dialects. The system is based on a measure of entropy on the monodialectal models. Apparently, to cope with all the Spanish dialects, the standard SAMPA symbol set for Spanish needs an extension.

Recognition results with the phonetic set of models defined by the less restrictive decision tree overcomes the monodialectal system, the SAMPA based clustering and the classical tree structure systems. This performance is important since it allows clustering avoiding the previous definition of a multidialectal phonetic set, solving errors due to the linguistic classification of sounds and gives information of similarity between sounds automatically. This technique can be very useful in cases where no linguistic information is available for some variant.

Currently we are applying this procedure to the complete set of Spanish dialects and to Arab language.

7. Acknowledgements

This work was granted by Spanish Government TIC 2002-04447-C02

8. References

[1] J. Köhler, "Multilingual phoneme recognition exploiting acoustic phonetic similarities of sounds", Proc. of Int. Conf. on Spoken Language Processing, ICSLP 96.

[2] T. Schultz, A. Waibel. "Multilingual and Crosslingual Speech Recognition", Proceedings of the DARPA Broadcast News Transcription and Understanding, 1998.

[3] W. Byrne, P. Beyerlien et al., "Towards Language Independent Acoustic Modeling", Proc. Int. Conf. On Acoustics, Speech and Signal Processing, ICASSP'00.

[4] A. Zgank, B. Imperl et al. "Crosslingual speech recognition with multilingual acoustic models based on agglomerative and tree-based triphone clustering". Proc. European Conf. on Speech Communication and Technology, 2001.

[5] U. Uebler, "Multilingual speech recognition in 7 languages", Speech Communication n°35, pp. 53-69. 2001

[6] A. Nogueiras, M. Caballero, A. Moreno. "Multidialectal Spanish Speech Recognition", Proc. Int. Conf. On Acoustics, Speech and Signal Processing, ICASSP'02.

[7] URL http://www.sala2.org

[8] A. Moreno, J.B. Mariño, "Spanish dialects: Phonetic transcription".Proc. of Int. Conf. on Spoken Language Processing ICSLP '98.