Multidialectal Acoustic Modeling: a Comparative Study

Mónica Caballero, Asunción Moreno, Albino Nogueiras

Talp Research Center
Department of Signal Theory and Communications
Universitat Politecnica de Catalunya, Spain
{monica,asuncion,albino}@gps.tsc.upc.edu

Abstract

In this paper, multidialectal acoustic modeling based on sharing data across dialects is addressed. A comparative study of different methods of combining data based on decision tree clustering algorithms is presented. Approaches evolved differ in the way of evaluating the similarity of sounds between dialects, and the decision tree structure applied. Proposed systems are tested with Spanish dialects across Spain and Latin America. All multidialectal proposed systems improve monodialectal performance using data from another dialect but it is shown that the way to share data is critical. The best combination between similarity measure and tree structure achieves an improvement of 7% over the results obtained with monodialectal systems.

1. Introduction

Dialectal variability is an important degrading factor in Automatic Speech Recognition (ASR) performance. A dialect mismatch between training and testing speakers significantly influences the recognition accuracy. The availability of dialectal-specific language resources is a key factor to solve dialectal mismatches, but often, there is not enough data to develop such specific recognizers. A common approach to cope with lack of dialectal data is sharing available data from different dialects to build a multidialectal ASR system [1]. The result is a single set of models able to be used in all dialects.

Definition of a multidialectal set of acoustic models can be solved with analogous techniques as those used in multilingual acoustic modeling research (i.e. each dialect is assimilated to a different language).

Similarity between sounds of different languages (dialects) either can be defined by an expert or can be data-driven found. Expert methods use linguistic knowledge. The most common approach is based on IPA or SAMPA alphabets [1, 2, 3]: phones of different languages (dialects) are similar if they share the same IPA or SAMPA symbol. In data-driven methods, similarity between phones is commonly estimated by evaluating the distance of their acoustics models (i.e. HMMs), applying agglomerative [3, 4] or decision tree based [5] clustering algorithms. Other data-driven approaches find similarity between phones by means of a confusion matrix [2].

Detecting similarities between sounds at phone level has been the most common option in the last years. Recently, multilingual acoustic modeling evaluates similarity on contextual models. In [5] two systems that evaluate the similarity at a phone and at a contextual level, respectively, are presented. In both systems a multilingual phone set is defined based on IPA alphabet. They found that evaluating similarity at a contextual level leads to better performance in recognition results. A comparison between application of different clustering algorithms in defining a multilingual set of triphone units is shown in [4]. They evaluate similarity with contextual models applying both, agglomerative and decision tree based clustering algorithms. In their approach, no global phone set is defined, but broad phonetic groups. Although agglomerative clustering algorithm gives a reduced number of clusters, decision tree method is found to give better recognition results and they solve modeling units not seen in the training data.

Concerning the structure of the decision tree in context modeling, typically, a different decision tree is grown for each unit (or each state of each unit) of the phone set. A single global decision tree was used in [6] to evaluate a novel node splitting criterion. This tree structure allows to share parameters between different phonetic units. A single global decision tree was also presented in [7] in order to model pronunciation variation from an acoustic modeling point of view. Significant improvement in the system performance was achieved. In [8] authors applied this tree structure for multidialectal acoustic modeling for three Spanish dialects with encouraging results.

This paper shows a comparative study of different methods of combining training data to obtain a robust multidialectal set of acoustic models. Approaches differ in the way of evaluating the similarity of sounds, SAMPA alphabet and HMM parameters, and in applied decision tree structure, i.e. multiple roots and one single global decision tree.

Methods are tested on Spanish dialects. Spanish as spoken in Argentina, Caribbean area, Colombia, Mexico and Spain are the considered dialects in this work.

This paper is organized as follows: Section 2 deals with Spanish language and its phonetic transcription. Section 3 describes the recognition system used in this study. Section 4 is dedicated to describe the multidialectal data sharing approaches presented. Section 5 describes the experiments carried out and the obtained results. Finally, conclusions of this work are presented in section 6.

2. Spanish dialects - phonetic transcription

For each considered dialect, a canonical phonetic transcription in SAMPA symbols [9] is obtained. Transcriptions are obtained automatically by means of a set of rules and a dictionary of exceptions. Canonical transcription rules for Spanish dialects, as classified in this work, were proposed in [10]. The phonetic transcription of Latin American variants they proposed is based on the rules for transcribing Spanish as spoken in Spain. This initial set of rules is modified according to the specific phonetics of every dialect. To represent dialectal pronunciation accurately, standard SAMPA symbol set for Spanish is extended symbols /h/ and /Z/ to cope with all the Latin American dialects. /R/ is also added to represent post-vocalic [r].

Table 1 shows the list of SAMPA symbols used for the transcriptions of the Spanish dialects. Each row shows the SAMPA symbols used for each dialectal variant. For sake of simplicity, the center column groups symbols that are shared across all dialects.

Table 1: SAMPA symbols used for phonetic transcription of Spanish dialects.

DIALECT	Shared Ph.	Non-Shared Ph.
ARGENTINA	a b B d D f g G i j J k l m n N o p r rr R s t tS u w z	Z x h
CARIBBEAN		jj h
COLOMBIA		jj h
MEXICO		jj x
SPAIN		jj x T

3. Recognition system

This work was developed in an in-house ASR system. The system uses Semicontinuous Hidden Markov 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. C, ΔC, and $\Delta\Delta$C are represented by 512 Gaussians and the Energy derivative is represented by 128 Gaussians.

The phonetic units are demiphones [11], a contextual unit that models the half of a phoneme taking into account its immediate context. A phone is modeled by two demiphones: ’$l-ph$’ ’$ph+r$’, where l and r stay for the left and the right phone context, respectively, and ph is the phone. Each demiphone is modeled by a 2 states left to right model.

3.1. Decision tree based clustering algorithm

In the phonetic decision tree, each node is a cluster of acoustic models and the branches represent questions relevant to the attributes of the phonetic units the model represent. To grow the tree, the entropy of each node is computed; according to answers to the questions, the acoustic models that stay in a parent node are split into two child nodes; for each possible question, the decrease of entropy is calculated as the entropy of the parent node minus the sum of the entropy of the two child nodes. The question that maximizes the decrease of entropy defines a new branch of the tree.

In every tree node, discrete approximation is used to evaluate partitions. With this approximation, 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 training 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. Using semicontinuous HMM, b’s are the mixture weights for each of the Gaussians of the codebook.

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

Stopping splitting criteria is defined by a minimum decrease of entropy and/or a threshold in the minimum number of realizations in the training data for each final cluster (leaf node).

Question set inquires about phonetic features of the phonetic unit the model represents (type, place and manner) and, optionally, non-phonetic questions (i.e. the position in the word, if the phone is an aspiration, if the phone belongs to a consonant group), or the dialect of the unit. Questions inquire about one single attribute (e.g. manner of articulation) and multiple questions about the same attribute are allowed using a logical OR link (e.g. is manner of articulation nasal OR fricative?). Question set is fully dependent on the approach and it will be exposed in the next sections.

4. Acoustic modeling - data sharing

4.1. Measures of similarity

We apply and compare two ways of evaluating the similarity, the first one is based on the SAMPA alphabet and the second one is a HMM based measure embedded in a decision tree based clustering algorithm.

4.1.1. SAMPA based

Sounds of different dialects that have the same representation in the SAMPA alphabet are considered the same phone. A global phone set is composed by all the SAMPA symbols necessary to represent the full set of dialects. Similarity measured in this way is done at a phone level. It is the most common approach used in the literature, and seems very useful if different languages/dialects share quite a lot of symbols, as in Spanish dialects.

4.1.2. HMM based

A set of context-dependent acoustics models (HMM) are trained for each dialect. A decision tree driven by the entropy measured over the dialect-dependent HMMs are used to define which sounds are similar enough to share training data. Multiple questions in the tree allow clusters of dialects. In order to be able to separate realizations of the same unit across dialects, dialect-dependent models are marked with a dialect tag. HMM based similarity measure allows to detect similar context-dependent acoustic units.

4.2. Tree structures

Two tree structures are studied: Multi-Root structure that applies SAMPA restrictions to the clustering algorithm, and One-Root structure without SAMPA constraints.

4.2.1. Multi-Root structure

A different tree (root) is created for each unit of the phone set. In each root all the context-dependent acoustic models belonging to the same phone are found. It is the typical structure used for context modeling in monolingual systems. Parameter sharing is not allowed across different units of the phone set. For multidialectal purposes, a previous step defining a global phone set is necessary. This previous definition is based on SAMPA alphabet. A similar tree structure is used in [5].

4.2.2. One-Root structure

A single tree is built for all the units in the phone set. Its root contains all the context-dependent acoustic models of all the

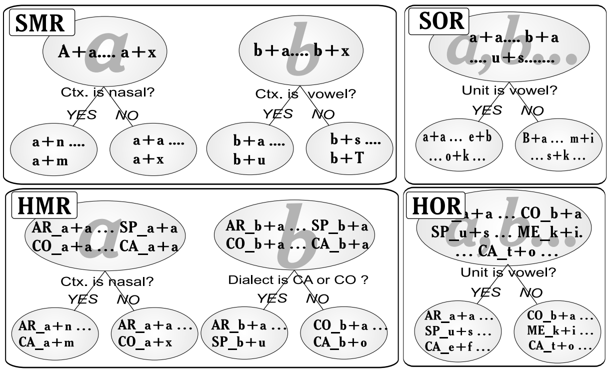

Figure 1: Multidialectal acoustic modeling approaches

units. This structure gives flexibility to share data across different phones.

4.3. Multidialectal acoustic modeling approaches

Four approaches for multidialectal acoustic modeling are obtained combining the types of similarities measures and decision tree structures presented above. Graphical representation of these approaches are shown in Figure 1.

4.3.1. SAMPA based measure, Multi-Root tree structure (SMR)

In this approach, a multidialectal phone set is defined through SAMPA symbols. Context modeling is achieved applying a decision tree clustering algorithm. Question set only inquires about the unit context. This is the most immediate approach and the most intuitive.

4.3.2. SAMPA based measure, One-Root tree structure (SOR)

This approach defines a global phone set based on SAMPA. The application of the One-Root tree structure allows joining different phones and contexts. Question set contains questions about the phone itself as well as the context. This is a total novel approach in multidialectal/multilingual research.

4.3.3. HMM based measure, Multi-Root tree structure (HMR)

Dialect-dependent models are created for each contextual unit. Similarity is only evaluated across phones with the same SAMPA representation. Question set ask for the context unit and the dialect.

4.3.4. HMM based measure, One-Root tree structure (HOR)

A single tree is grown with all the dialect-dependent models in the root node. This approach gives maximum freedom to the clustering algorithm. Dialect information is given to the tree and no SAMPA restrictions are applied. As in HMR approach, models with the same SAMPA representation can be distinguished. One-Root tree structure allows models with distinct SAMPA representation be joined if they are similar enough. This approach makes the system totally automatic and independent of the prior phonetic knowledge. This system was first presented by the authors in [8].

5. 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 4,000 different speakers. In this work, 3,500 speakers were selected for training and 200 for test. The databases of Spanish as spoken in Latin America were created in the SALA [12] project. Each database consists of fixed network telephone recordings from 1000 different speakers, 800 speakers were selected for training and 200 for test. The systems are trained with a set of phonetically rich words and sentences. The recognition test is composed of phonetically rich words with a vocabulary size of 4,500 words, identical for all dialects. Table 2 shows the total amount of training and testing data for each dialect considered.

Table 2: *Training and testing utterances for dialects considered in this study.*

DIALECT	AR	CA	CO	ME	SP
train utt.	9,568	9,303	8,874	11,506	40,936
test utt.	2,575	2,411	2,358	2,022	3,632

5.1. Monodialectal systems - baseline recognizers

One baseline recognizer was built for each dialect. The purpose is to compare results with the multidialectal approaches proposed in this work. A decision tree based clustering algorithm was applied for context modeling. For each unit of the dialect phone set a different tree was created.

Table 3 shows the number of models of each system. The system of Spain reaches the largest set of models, due to the larger amount of available data. The total number of models needed for recognizing all dialects is 3,596. Table 4 shows the percentage of Word Error Rate (WER%) for the baseline recognizers and their average value. The system for Spain gives the best result. Caribbean and Argentinean systems have similar performance. Colombian and Mexican systems give the worst rates. A possible explanation for this worse performance is the higher noise in the speech signals in these databases compared with the other ones.

Table 3: *Number of models for Spanish monodialectal systems.*

DIALECT	AR	CA	CO	ME	SP
N. HMM	662	688	683	716	847

5.2. Multidialectal acoustic modeling

Results of these experiments are summarized in Table 4. All systems improve the baseline average WER. Multilingual approaches slightly degrade the performance of the baseline of Spain. This result is not surprising since we are adding variability to a well-trained system. We consider that this degradation is acceptable for the sake of the system.

Approaches SMR and SOR, which define the phone set based on SAMPA alphabet, uses the smallest number of models (988 and 981 models, respectively). These figures are comparable to the number of monodialectal models. Results are similar with both systems. SOR system decreases the average WER over dialects to 7.02%. The improvement of baseline results in both cases is caused by the reduction of WER in Colombian, Mexican and Caribbean variants.

Dialect querying (approaches HMR and HOR) grew the decision tree to 3,600 leaf nodes. Experiments were done in order to determine the optimal size of the acoustic model set. Best results were obtained with 2,000 models in both cases. HMR approach improves the performance achieved with SMR and SOR approaches. Using One-Root tree structure (HOR approach) leads to the best system, reducing average WER in nearly 7% over baseline results. This system outperforms all Latin American baseline results. WER for Spanish as spoken in Spain is nearly as good as in its dialect-specific system.

Table 4: *WER% for baseline and multidialectal recognition systems.*

DIALECT	**Mono**	**SMR**	**SOR**	**HMR**	**HOR**
ARGENTINA	7.34	8.31	7.76	6.37	6.23
CARIBBEAN	6.71	6.27	6.27	6.41	6.41
COLOMBIA	9.22	8.28	8.28	7.97	7.81
MEXICO	10.10	8.01	8.17	9.62	8.65
SPAIN	3.62	4.74	4.6	4.46	4.04
AVERAGE	*7.40*	*7.12*	*7.02*	*6.97*	***6.63***

5.3. Discussion

Table 5 shows the percentage of full multidialectal (clusters containing data from all dialects) and semi-multidialectal (clusters containing data from more than one dialect) nodes. Percentages for HMM based systems are calculated for the 2,000 acoustic model set.

Maximum data sharing is given by approaches that define a phone set based on SAMPA alphabet. Total percentage of all multidialectal models for SMR and SOR approaches is similar. SOR approach slightly increases data sharing percentage as it allows joining dialect-specific models in the same cluster.

Systems based on HMM based measure allow to separate realizations of the same phone across different dialects. Opening the decision tree up up 2,000 clusters decreases full multidialectal nodes percentage. When One-Root structure is applied semi-multidialectal nodes percentage is increased substantially.

These percentages shows that better recognition performance is achieved sharing data between clusters of dialects than sharing data across all variants.

Table 5: *Data sharing percentages across dialects in multidialectal systems.*

Approach	**Sam**	**SOR**	**HMR**	**HOR**
Full multid. C.	69.23%	69.72%	6.70%	6.20%
Semi multid. C.	20.65%	21.61%	11.20%	14.85%

6. Conclusions

In this paper, approaches for building a robust multidialectal set of acoustic models have been presented. Our objective was to take the maximum advantage of sharing data across dialects to achieve higher recognition rates. All proposed systems improve monodialectal performance using data from other dialects but, as it has been shown along the paper, the way to share data plays an important role. Comparison between measures of similarity leads to the conclusion that it is better to determine similarities between sounds across dialects with a HMM based measure than based on SAMPA alphabet. Applying One-Root tree structure reduces WER substantially.

7. Acknowledgements

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

8. References

[1] Chengalvarayan, R., 2001. Accent-Independent universal HMM based speech recognizer for American, Australian and British English. In Proceedings Eurospeech , Aalborg, Denmark, 2001, pp. 2733–2736.

[2] Byrne, W., Beyerlein, P., Huerta, J. M., Khudanpur, S., Marthi B., Morgan J., Peterek N., Picone J., Vergyri D., Wang W., 2000. Towards language independent acoustic modeling. In Proceedings ICASSP, Istanbul, Turkey, 2000, Vol. 2., pp. 1029–1032.

[3] Köhler, J., 2001. Multilingual phone models for vocabulary-independent speech recognition tasks. In Speech Communication, Vol. 35, Issues 1-2, August 2001, pp 21–30.

[4] Imperl, B., Kačič, Z., Horvat, B., Žgank, B., 2003. Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones. In Speech Communication, Vol. 39, Issues 3-4, February 2003, pp. 353–366.

[5] Schultz, T., Waibel, A, 2001. Language independent and language adaptative acoustic modeling for speech recognition. In Speech Communication, Vol. 35, August 2001, pp 31–51.

[6] Duchateau, J., Demuynck, K., Van Compernolle, D., 1997. A novel node splitting criterion in decision tree construction for semi-continuous HMMs. In Proceedings Eurospeech, Rhodes, Greece, 1997, volume 3, pp. 1183–1186.

[7] Yu, H., Schultz, T., 2003. Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition. In Proceedings Eurospeech, Geneva, Switzerland, 2003, pp. 1869–1872.

[8] Caballero, M., Moreno, A., Nogueiras, A., 2004. Data driven multidialectal phone set for Spanish dialects. In Proceedings ICSLP, Jeju Island, Korea, 2004, pp. 837-840.

[9] Gibbon, D., Moore, R., Winski, R., 1997. Handbook of Standards and Resources for Spoken Language Resources. Mouton de Gruyter, New-York, 1997, ISBN 3-11-015366-1.

[10] Moreno, A., Mariño, J.B., 1998. Spanish dialects: Phonetic transcription. In Proceedings ICSLP, Sidney, Australia, 1998, paper 0598.

[11] Mariño, J.B., Pachés-Leal, P., Nogueiras A., 1998. The Demiphone versus the Triphone in a Decision-Tree State-Tying Framework. In Proceedings ICSLP, Sydney, Australia, 1998, Vol. I, pp. 477–480.

[12] Moreno, A., Höge, H., Köhler, J., 1998. SpeechDat Across Latin America. Project SALA. In Proceedings International Conference on Language Resources and Evaluation (LREC), Granada, Spain, 1998, Vol. I, pp. 367–370.