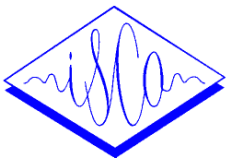

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

# ACOUSTIC MODELING OF CONTEXT DEPENDENT UNITS, FOR LARGE VOCABULARY SPEECH RECOGNITION IN SPANISH

*J. Alvarez-Cercadillo[1], C.-H. Lee, and L.A. Hernández-Gómez[2]*

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974, U.S.A.

## ABSTRACT

We present a study on acoustic modeling of Spanish phonetic units. Bootstrap with a set of English phonetic models, we first obtain context-independent unit models for Spanish. We then compare context-dependent modeling techniques based on the conventional *maximum likelihood* (ML) and the *maximum a posteriori* (MAP) criteria. We found the MAP-based context adaptation approach produces a better result than the ML approach when a large number of units need to be modeled but the amount of training data is limited.

10.21437/Eurospeech.1995-180

## 1. INTRODUCTION

In the past few years many acoustic modeling techniques have been studied for modeling phonetic units of different languages. The most prevailing approach is the use of the hidden Markov model (HMM) framework. In this paper, we conduct a comprehensive study of acoustic modeling of phonetic units for continuous speech recognition of Spanish.

Both context-independent units (CIU's) and context-dependent units (CDU's) are modeled. Each unit is modeled by a continuous density HMM (CDHMM) with mixture Gaussian state observation density [2]. The CDHMM parameters are estimated from a set of spoken Spanish utterances using first a ML criterion. In addition to ML training of CDU's, we also study the use of *maximum a posteriori* (MAP) technique [1] to estimate the parameters of a CDU using the corresponding CIU model as the seed model and the corresponding CDU data as adaptation data. Such an adaptation scheme is similar to MAP based *speaker adaptation* where a speaker-independent (SI) model is used as the seed model and the set of speaker-specific data is used as adaptation data to produce the speaker-adaptive models. We will refer to the new adaptation scenario as MAP-based *context adaptation* in this study. We compare recognition performance on a 2451-word speaker-independent, continuous speech recognition task. As expected, context-dependent modeling performs better than context-independent modelling. We also found that the recognition performance improved when more context-dependent units were modeled. The best performance was obtained when MAP-based context adaptation was used to model a large number of context-dependent units.

[1] J. Alvarez-Cercadillo is now with Speech Technology Group, at Telefonica I+D, and with Alfonso X University, Madrid.

[2] E.T.S.I. Telecomunication. Polytechnic University of Madrid.

## 2. TASK AND DATABASES

A Spanish Headline speech database for continuous speech recognition research, has been designed and recorded. This database includes a big quantity of tokens of each sound in most relevant contexts. The *corpus* is composed by 650 sentences that were selected by the Linguistic Research Department in AT&T Bell Labs according to some phonological criteria like phoneme occurrence frequency, and sentence intonation. *Vocabulary size* is equal to 2451. *Perplexity* of the task is equal to 2451 using the null grammar (every word can follow every word with an equal probability) and equal to 3 using a word pair grammar. The database is composed of 25 male and 25 female speakers, all of them saying 200 sentences, and one speaker saying all the 650 sentences. Therefore there are 50 x 200 + 650 = 10650 utterances available. Besides this, we have used the models trained from the TIMIT database to obtain an initial segmentation of the Spanish Headline database as will be explained later, and for interlanguage-testing.

In order to automatically generate the *Lexical Dictionary* for the Spanish speech database corpus, a simple *Orthographic-to-Phonetic Transcriber* for Spanish [4] has been used which needs just a set of 39 basic transcription rules.

## 3. SELECTION OF FUNDAMENTAL SPEECH UNITS AND EXTENDED SPEECH UNITS

In this section we deal with the selection of a speech unit set which cover all the sounds in Spanish. Besides this, we present a method for training CI Spanish models. A key issue in this method is that neither a segmented Spanish database, neither initial Spanish models were available; so English models were initially used. to segment the Spanish database.

### 3.1 Acoustic phonetics of Spanish

There are few basic phonemes in Spanish. With just 24 phonemes we can obtain a basic phonetic transcription of all Spanish words. Following we present some interesting phonetic characteristics of Spanish language that are related to speech recognition:

- There are just five vowel sounds, and they are acoustically very well defined.
- More common sounds are the vowels /e/ (13,51%) and /a/ (13,4%).

We have to remark that frequency of vowels is almost 50%, so a good training of the five vowel models will be important in order to get a high recognition rate. Another key issue in Spanish speech recognition is the abundance of short words: almost 50% of Spanish words have less than three sounds, so lot of keywords will be misrecognized.

### 3.2 Selection and modeling of context independent units

For our research we have selected the phone-like units (PLU's) as fundamental speech units [2]. In order to cover the 24 basic Spanish phonemes, and the silence, a set of 25 CIU's has been selected. Next we explain the procedure for obtaining the initial segmentation of the Spanish database using English models. Using this segmentation, Spanish CI models will be trained.

#### *3.2.1 Bootstrapping from English Phone Models*

For training the PLU models, a variation of segmental K-means training procedure [5] has been implemented. In this method each sentence is represented as concatenation of PLU models according with its phonetic transcription, including an optional silence between words, and at the beginning and end of each sentence. Phoneme segments are extracted over this network via Viterbi decoding. For this decoding process some *initial subword models are necessary.*

In some approaches where this initial models are not available, an initial linear segmentation has been used [2], obtaining good results only when a large quantity of data is available. Our approach consists in using some well trained *English phonemes for initially segmenting the Spanish database*, and obtaining the initial (and not precise) boundaries of the Spanish phonemes. These boundaries will be the seed for the K-means training procedure. In order to choose the most appropriate English seed model set, we have established a correspondence between each Spanish phoneme and its correspondent English one, following some phonological criteria as *soundness*, *place* and *manner* of *articulation.* Using this criterion we found that 17 of the 24 Spanish phonemes exactly matched with 17 English ones, and just 7 Spanish phonemes were not present in English.

| Spanish Phoneme | Phonetic Classification | Phonetic Classification |
|---|---|---|
| t | Stop Dental Unvoiced | Stop Alveolar Unvoiced |
| d | Stop Dental Voiced | Stop Alveolar Voiced |
| j | Fricative Velar Unvoiced | Fricative Glotal Unvoiced |
| r | Flap | Stop Alveolar Voiced |
| rr | Trill | Glode Retroflex Voiced |
| ll | Lateral Palatal Voiced | Affricate Palatal Voiced |
| ñ | Nasal Palatal Voiced | Nasal Alveolar Voiced |

Table 1: Spanish Phonemes not present in English, and phonetic differences with the English phonemes that were used as initialization.

For these 7 Spanish phonemes we chose an approximate English phoneme for its initialization. In table 1 we show those Spanish phonemes that were initialized with an approximate English one.

#### *3.2.2 Segmentation-Training Loop*

Once we have decided to use English phonemes as initial models, the complete procedure for segmentation and estimation is well defined. It consists in the following six steps:

1. Identification of the complete set of subword Spanish units based in phonemes (PLU). We have consider 25 CIU, as said in subsection 3.1.
2. Establishment of a correspondence between all Spanish sounds and its more alike English one, as discused before.
3. Select the 25 English phone models that will be the seed for segmenting the Spanish database.
4. Segment the Spanish database using previous models.
5. Train new models using segments obtained in step 4.
6. Repeat steps 4-5 till convergence.

We have repeated this training-resegmentation loop just 5 times, because no improvement were observed from the third iteration. The model improvement at each iteration can be seen in figure 1, where classification rate is plotted at each iteration. These results will be discussed in the next section.

#### *3.2.3 Phonetic Classification Results*

In order to test models obtained at different steps of the above training algorithm, phonetic classification over the test database was made after each iteration. Results using 4, 8, 16 and 32 mixtures are shown in figure 1. At first iteration 32 mixtures models classifies correctly 78% of phoneme segments. This good performance is explained because seed English models were initially well trained, and because the established correspondence between Spanish and English phonemes is almost perfect 80% of times. An increment of 4% and 1% is obtained in the

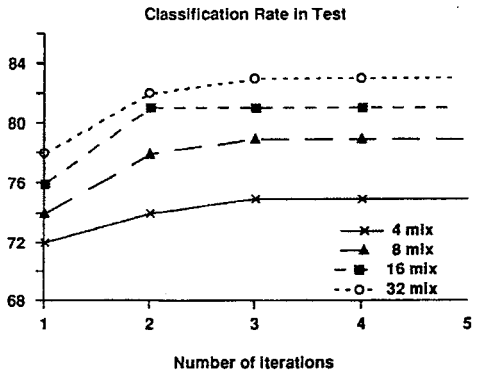


Fig1. Phonetic classification rate at each Iteration.

second and third iteration, and no improvement was observed in fourth and fifth iteration, so the training algorithm was stopped. The classification rate at this point was equal to 83%. These results probe that the followed training method converges toward optimal models.

In order to validate the obtained Spanish phone models, and compare their performance with well trained English ones, an analysis of classification for English language was also made. For a better comparison of phonetic classification results, we have divided English and Spanish sounds into five classes: vowels, semivowels, stops, nasals and fricatives. The results are shown in Table 2. Total classification rate was equal to 86% in Spanish using 25 CIU's and equal to 72% using 48 English models. As can be seen, much better results were obtained for Spanish. This better performance is due to the following main reasons:

- First, using just 25 Spanish units the number of units for classification is lower than that in English;
- Second, we think that Spanish vowels are better modeled due to their stable acoustic realization (there are just five vowels in Spanish and they are well defined).

The set of English vowels and semivowels obtained a classification rate of 66% and 78% respectively, while Spanish vowels obtain globally 90%. If we notice that almost 50% of Spanish sounds are vowels, we could better understand that a better classification rate is obtained for Spanish.

### 3.3 Selection of context dependent units

CIU's are not able to model coarticulation effects between two consecutive sounds. For this reason we have modelled a set of context dependent units too. The collection of units which forms this set have been selected from the occurrence frequency of triphones in the corpus, by fixing a *threshold* T which decides that a triphone is going to be included, if it appears at least T times in the training set [2]. Following this criterion, we have initially fixed the threshold equal to 2500 tokens. Secondly we fixed the threshold equal to both 70 and 20

| Database | Headlines (Spanish) | TIMIT (English) |
|---|---|---|
| Vowels | 90% | 66% |
| Semivowels | --% | 78% |
| Stops | 84% | 80% |
| Nasals | 77% | 69% |
| Fricatives | 84% | 79% |
| Total | 86% | 72% |

Table 2: Comparative results for English and Spanish phonetic classification

tokens, and obtained a set of 1037 and 2087 CDU's respectively.

## 4. ACOUSTIC MODELING OF CONTEXT DEPENDENT UNITS

In this section we present two different methods for obtaining triphone models. In the first approach triphone models are estimated using maximum likelihood criteria, and context dependent units (CDU's) are obtained. The second approach uses maximum a posteriori criteria (MAP) for triphone modelling, and context adapted units (CAU's) are obtained.

Word recognition results using both unit sets are presented for continuous speech recognition. Experiments were made under several conditions: i.e. number of units, different grammars, etc... The *word accuracy rate* has been obtained without using grammar (perplexity 2451). Results are presented versus number of modelled units.

### 4.1 Recognition Using CD units

For modelling the coarticulation effect, three different sets of 56, 1037 and 2085 triphones were selected as explained in section 3.3. In the first approach these models were trained following the ML criterion, and CD units were obtained. Recognition was made, and the obtained results using these CDU's are shown in figure 2, plotted in full squares. This figure shows word accuracy results using a null grammar (perplexity 2451) and 16 mixtures. It can be seen that word accuracy increases progressively as number of units is increased, going form 56% with 25 units towards 70% using 2085 units.

Results using 4 mixtures CDU models are plotted in dashed lines. A similar improvement with the number of units is observed, going from 49% using 56 units, toward 68% using 2085 units.

From these results we can conclude that CDU's can be used for solving the coarticulation problem. Improvement will be larger as more units are used, taking care or having tokens enough of each CDU for getting a good training.

### 4.2 Recognition Using CA Units

When the 2085 unit set is used, 1048 units (2085-1037) have just between 70 and 20 tokens for training. When this quantity of tokens is around 70, good trained models can be obtained using ML criterion. But hen the number of tokens is around 20, ML method has no enough data to accurately estimate model parameters. For these cases, alternative training methods are necessary, and some of them have been successfully proposed like *deleted interpolation* [3].

In our approach we have used the *MAP technique* for modelling triphones [6]. We will refer them as context adapted (CA) models. The "a priori" information necessary for each CA model, has been obtained from that phone model which match the central phone of the triphone. After this, the training of each triphone has been made using those phones embedded in the same context. Using this seed as a priori information not so many tokens of each unit are necessary for training.

Results in continuous speech recognition using 16 mixtures CA models are shown in figure 2. A 56% *word accuracy* was obtained using 56 CAU's, which is worse than the one obtained using CDU's. However CAU's overcomes CDU's when 1037 units are used. This improvement is also observed using 2085 units where 72% *word accuracy* is obtained using CAU's, and 70% using CDU's.

From these results we can see that CAU models fix non-frequent sounds better than CDU models. This behavior is explained because MAP criterion uses better the information embedded in the CIU models, wich is specially interesting for training less frequent units.

## 5. SUMMARY AND DISCUSSION

In this paper we have established a procedure for designing a Spanish continuous speech recognition system based in subword units which gives good performance.

A basic unit set has been defined, using 24 Spanish phonemes plus a silence unit.

A Spanish continuous speech database has been recorded. This database includes 50 speakers saying 200 sentences each, from a 650 sentences corpus. The Spanish database has been automatically segmented with a training-resegmentation loop, using initial English models.

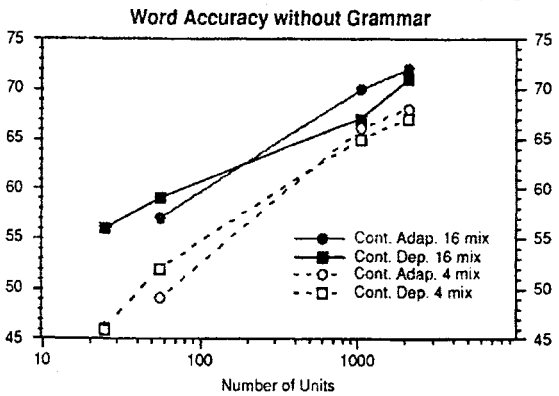


Fig. 2 Word accuracy versus number of CAU and CDU.

Some comparative studies between English and Spanish languages have been made. In recognition tests we have found that phonetic classification in Spanish was 86%. This rate is much better than the one obtained in English (72%), because of the less number of basic unit set, and the more acoustic stability of Spanish vowels.

Continuous speech recognition results using 25 context independent phone models were equal to 59% of word accuracy with 32 mixtures models and using a null grammar (perplexity 2451).

Finally we have studied the context dependent unit modeling. Three different units set of 56, 1037 and 2085 units were defined, by forcing those units to appear at least 2500, 70 and 20 times in the database respectively. These units have been trained using both the ML and MAP methods obtaining context dependent units (CDU) and context adapted units (CAU). Results show that CAU's model better than CDU's when few training data for the unit is available.

From these results we propose the use of a mixed set of CDU's and CAU's. In this set, CDU's are chosen to model most frequent units in training database, and CAU's are used to model those sounds with very few training data. Preliminary experiments in this line show a global improvement in word accuracy.

## ACKNOWLEDGMENTS

We thank Joseph Olive for allowing us to use the Spanish Headline Corpus. We also thank Pilar Prieto for her expert advises about the phonetics of the Spanish language, and to Javier Poyatos and J.J. Molina for their help in organizing the Spanish database.

## REFERENCES

[1] Gauvain, J.L. et al. MAP Estimation for Multivariate Gaussian Observations of Markov Chains. IEEE Trans. SAP. April 1994.

[2] Lee, C.-H., et al. "Acoustic Modelling for Large Vocabulary Speech Recognition". Computer, Speech and Language, 4,. 1990.

[3] Lee, K.F. Automatic Speech Recognition. "The Development of the SPHINX system". Kluwer Academic Publishers. 1989.

[4] Lopez, E. Ph. D. Thesis. "Estudio de Técnicas de Procesado Lingüístico y Acústico para sistemas de conversión texto-voz en español, basados en concatenación de unidades". Polytechnic Univ. Madrid. 1993.

[5] Rabiner, L. et al. An Introduction to HMM. IEEE ASSP Magazine, vol 3 (1986). 4-16.

[6] Alvarez-Cercadillo, Jorge. Ph. D. Thesis. "Modelado acústico y teorías de adaptación MAP para reconocimiento de habla continua en castellano". Polytechnic Univ. Madrid. 1995.