# THE DEVELOPMENT OF A SPEAKER INDEPENDENT CONTINUOUS SPEECH RECOGNIZER FOR PORTUGUESE

João P. Neto, Ciro A. Martins and Luís B. Almeida

*INESC - IST*
*R. Alves Redol, 9*
*1000 Lisboa - Portugal*
*E-Mails: jpn@inesc.pt, cam@inesc.pt, lba@inesc.pt*

## ABSTRACT

The development and evaluation of large vocabulary, speaker-independent continuous speech recognition systems are mainly done for the American English language. In this paper we present the work done to date in the development of an hybrid large vocabulary, speaker-independent continuous speech recognition system for the European Portuguese language. Due to the lack of a large appropriate speech and text database to be used in the development of that system we started collecting a large database and at the same time began developing a baseline system based on a smaller database. On this baseline system we applied techniques for automatic segmentation and labeling, in parallel with the development of a basic lexicon and language model for Portuguese. In the last part of this paper we also present the first steps of our work over the new database.

## 1. INTRODUCTION

The last years show a great improvement in large vocabulary, speaker-independent continuous speech recognition systems. Unfortunately the development and evaluation of these systems are mainly done for the American English language, through the support of ARPA, and little work has been done on evaluating state of the art recognizers on other languages. One good exception comes from the EU LRE project SQALE [1] and from the LTR project SPRACH under which the present work has been done[1].

The goal of the work reported in this paper was to build a hybrid speaker independent, large vocabulary continuous speech recognition system for the European Portuguese language. Our group has a large experience in the development of hybrid systems [2] for continuous speech recognition for the English language with a major focus in the development of speaker-adaptation techniques [3] applied to the connectionist component of the hybrid system.

However, the situation of the Portuguese language is different from those of the English and some other European languages. There is no large appropriate speech and text database that can be used in the development of that system. The EUROM.1 SAM Portuguese database which is the only available database with continuous speech has only a very limited number of sentences per speaker [4]. For that reason we started collecting a large database of both text and speech [5]. In the development of this new Portuguese database our aim was to create a corpus equivalent in size to the WSJ0 database [6]. We also chose as corpus a newspaper's text and our choice was the Portuguese "PÚBLICO" newspaper. This database is being collected at INESC and is expected to be available in CD format in September 97. The database will have three sets: the *training set* with approximately 8,000 sentences (around 80 sentences for each of the 100 speakers) and a development test set and an evaluation test set with 400 sentences each (around 40 sentences for each of the 10 speakers). On both development and evaluation test sets each speaker will additionally have 15 speaker-adaptation sentences. For more details on this database see [5] on these Proceedings.

The other main difficulty, for Portuguese, was the fact that there was no segmented and labeled database. This is a common problem for new languages and was overcome for American English through the development of a hand labeled database (TIMIT [7]). In this paper we will present a solution for the Portuguese language based on the TIMIT database.

In our opinion the first attempt to build a large vocabulary, speaker-independent continuous speech recognition system for a new language must take an iterative approach. In our case this was not just a question of choice but we were forced to use it, due to the unavailability of appropriate databases. In that sense we began developing a baseline system for the Portuguese language, based on the EUROM.1 SAM Portuguese database, where we tested techniques for automatic segmentation and labeling, in parallel with the development of a basic lexicon and language model for Portuguese. During this period we have developed and started collecting the PÚBLICO database. As soon as the first recordings of this new database were available we started devel-

[1] One of the goals of the SPRACH project is the porting of hybrid systems to new languages: French and Portuguese.

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

oping our current system (starting from the baseline system). The new system was trained and tested over the first, already collected speech material of PÚBLICO database. With this system we started developing large vocabularies and stochastic language models.

In the next section we describe the architecture of our hybrid system and in sections 3 and 4 we present the baseline system and our current system respectively. At the end we present some conclusions and further work to be done.

## 2. HYBRID SYSTEM

The hybrid approach that we use combines the temporal modeling capabilities of hidden Markov models (HMMs) with the pattern classification capabilities of multilayer perceptrons. In this hybrid HMM/MLP system, a Markov process is used to model the basic temporal nature of the speech signal. The MLP is used as the acoustic model within the HMM framework. The MLP estimates context-independent posterior phone probabilities to be used in the Markov process. We use an HMM/ANN hybrid system where the connectionist part is based on a multilayer perceptron (MLP), with a single hidden layer and incorporating local acoustic context via a multiframe input window.

## 3. BASELINE SYSTEM

### 3.1. Database description

The baseline system was based on the EUROM.1 SAM Portuguese database which is currently available through ELRA. This database consists of read speech with three different sets of speakers and with different recording material [4]. The database has no phonetic labeling, no dictionary and no language model. We selected as training set, to be used during the development of the system, the *passages* of the so called Many Talker Set (60 speakers with 3 passages each, giving a total of 179 passages). The test set consists of the passages of the so called Few Talker Set (10 speakers with 3 passages each, giving a total of 30 passages). Each passage is composed of 5 thematically connected sentences. There were 10 different passages given a total of 50 different sentences. Each speech file contained the all passage. The database contains a total of 3,408 words, of which 1,314 were different words. As can be concluded this is a small database with a medium size vocabulary.

### 3.2. Lexicon development

The lexicon was built from the phonetic transcription of the sentences, which was included in the database. The transcriptions take into account the co-articulation effects linguistically expected from the sequence of words in the sentences. For that reason we obtained several pronunciations for some of the words in the lexicon. In the end we obtained 1,437 different word pronunciations generating a multipronunciation lexicon.

### 3.3. Automatic segmentation and labeling

Because there is no segmented and labeled speech database for the Portuguese language we tried to label the SAM database through an automatic process. The basic idea of this automatic process is to use the TIMIT database to create acoustic-phonetic models for English. Then through a linguistic approximation mapping from the TIMIT phonemes and phones to the Portuguese phonemes, we create acoustic-phonetic models for Portuguese. In this process the following steps were involved:

1. We trained first the acoustic-phonetic models for English over the TIMIT database (we trained an MLP over that database to perform phoneme classification).

2. We created two conversion tables: one from the TIMIT phonemes to the IPA symbols [7] and another one from the Portuguese SAM phonemes to the IPA symbols. Putting both tables together we created a mapping from TIMIT phonemes to Portuguese phonemes. Obviously not all the phonemes have correspondence. The TIMIT phonemes without correspondence are mapped to the closest Portuguese phonemes. There are just a few phonemes in this situation. Conversely, there are a few Portuguese phonemes which do not have correspondence in TIMIT. For these we defined a small fixed value for their probabilities.

3. The next step was to feed these TIMIT acoustic-phonetic models (the MLP) with the SAM database. The output of these acoustic-phonetic models (the output of the MLP) are estimates of context-independent posterior phone probabilities.

4. The resulting probabilities were transformed according to the previously defined mapping table, becoming a new set of probabilities corresponding to the Portuguese phonemes.

5. These probabilities were used in the decoder to perform a forced alignment. As input to this forced alignment process we also used the above mentioned Portuguese baseline lexicon and the correct sentence text from the SAM database. The outputs of this process were the labels (Portuguese phonemes) corresponding to the frame division of the SAM database.

6. After completing the previous forced alignment we trained a new MLP over the SAM database (with the labels from that alignment) to perform phoneme classification.

7. With this new network (completely obtained from Portuguese components) we generated a new set of probabilities.

8. We iterated several times the forced alignment and the acoustic-phonetic model training described in steps 5, 6 and 7.

Table 1 shows the percentage of correct frames in the different iterations of this process. The evaluation was made during the training, with the labels from the previous alignment.

| Iteration | % correct frames on training set | % correct frames on test set |
|---|---|---|
| 1 | 54.86 | 53.15 |
| 2 | 63.40 | 62.26 |
| 3 | 65.41 | 63.91 |

Table 1. Evolution of the training and alignment process.

The results show the improvement made by the MLP training on the phoneme classification (at the frame level). This process of training/alignment proved to be effective in decreasing the classification error.

### 3.4. Language modeling

To create the language models for the baseline system we were limited to the SAM texts. The sentences of these texts were generated artificially and explicitly for this database. That means the sentences were not directly extracted from any texts. As previously mentioned we have 50 different sentences with a total of 3,408 words of which 1,314 are different from one another. In the total we had 3,158 word pairs, of which 2,782 were different and 2,560 pairs just occurred once. From these numbers, as expected, we were unable to make a bigram model and we decided to use a simple word pair language model.

We built two different word pair language models. For the first one we collected the pairs from the SAM text for the sentences isolated from one another. In this case we got 250 separate sentences. For the second language model we considered as our unit the passage by itself, and we got 50 passages. In this case there was no division between sentences in the same passage. This second model describes better what we really have on the database.

### 3.5. Evaluation

After the training/alignment process we evaluated the hybrid system in terms of word recognition. In the training phase we used 179 files from the Many Talker Set. For evaluation we picked the 10 speakers of the Few Talker Set, choosing just three passages from each speaker. The results are presented in the Table 2.

| Grammar | test set word error % |
|---|---|
| word pair 1 | 50.1% |
| word pair 2 | 15.6% |

Table 2. Percentage of word error in the test set depending of the language model.

These results show the great influence of the language model in such a small task. In the first case (word pair 1) there were no connections between sentences in the language model. When we introduced those connections in the language model (word pair 2) there was a great improvement in the system performance.

The baseline system proved to be very useful as a start for the overall recognition process for the Portuguese language. The major step was on the automatic alignment where we started from the TIMIT alignment and ended with acoustic-phonetic models for Portuguese. However this task had some severe limitations: a very limited vocabulary, a very limited amount of text to generate good language models and very few training and test data.

## 4. CURRENT SYSTEM

As mentioned in the Introduction our final goal is to build a speaker independent, large vocabulary continuous speech recognition system for the European Portuguese language. There is a long way from the baseline system to our final goal. To achieve it we designed a new task based on the PÚBLICO newspaper, where we started collecting a large speech database. The speech database is being collected from April to June 1997. In the work reported in this section we are using a small initial part of the database, collected to validate and test the overall process.

### 4.1. Current database

The PÚBLICO database is intended to have approximately 8,000 training utterances (equivalent in size to WSJ0), spoken by 100 speakers. This database is described in detail in [5]. Presently we are using a small setup set with 10 speakers (5 male and 5 female) and 812 utterances. This set was selected from the training part and used in the validation of the overall process of sentence selection and recording. This set represents approximately 10% of the final training set. It is still a very small database, but larger than SAM, and is useful for the development of the alignment and training process of the current system.

### 4.2. Lexicon

From 188 editions of the PÚBLICO newspaper, corresponding to 24,287 articles and 10,976,009 total words, we computed a Word Frequency List (WFL) with a total of 155,867 different words. After selection of the different sets (training, development test and evaluation test) we ended with a total of 27,833 different words. This list of words was phonetically transcribed by a rule system [2]. This system has some know difficulties. Due to these problems this lexicon is being hand revised by a specialized linguist. The lexicon will be further perfectioned through the use of smoothing techniques to generate alternative pronunciations based on actual pronunciations and on the likelihood of the acoustic-phonetic models.

### 4.3. Automatic segmentation and labeling

In the database labeling process we started from the baseline system. Again we used an iterative alignment/training process as described in Section 3.3. A few iterations of this process have already been made with the results presented in Table 3. In the total we had 10 speakers (5 male and 5 female) with 812 utterances. We chose 652 for training (corresponding to 8 speakers) and 160 for validation (corresponding to 2 speakers). In the following table we present results on the training and validation sets.

| Iteration | % correct frames on training set | % correct frames on validation set |
|---|---|---|
| 1 | 61.61 | 63.60 |
| 2 | 66.30 | 66.07 |
| 3 | 67.58 | 67.93 |
| 4 | 68.00 | 68.14 |

Table 3. Evolution of the alignment/training process with the first 10 speakers of the PÚBLICO database.

This alignment/training process is still in progress.

### 4.4. Language modeling

From the total PÚBLICO texts we selected 80% as the training part, 10% as development part and 10% as evaluation part. From the training part, bigram backoff closed language models were computed. The 5K development test set language model yielded a perplexity of 231. This represents a large perplexity task. For more details see [5].

### 4.5. Future Work

We are in the middle of the alignment/training process and the evaluation of the system is still producing a large word error. Some causes of this error are known. The speech database is still rather small. The alignment/training process is not complete, too. Furthermore, the lexicon is not yet fully corrected. As these limitations progressively get eliminated, we expect to be able to approach the performances that are obtained for American English with recognizers of the same kind.

## 5. CONCLUSIONS

In the first part of this work we present a baseline system for continuous speech recognition based on the SAM database. The development of this baseline system proved to be a very useful step in the implementation of the overall recognition process for the Portuguese language, with a major focus on the automatic alignment of the speech database.

Due to the lack of a large appropriate speech and text database we decided to design and collect a large database based on newspaper text. This is described in [5] on these Proceedings.

The second part of this work is based on the first 10 speakers of a setup set extracted from the new database. We started to build a new system based on a new lexicon and a new bigram language model. With the further development of this system and its extension to the complete database, we expect to obtain a good recognizer for the European Portuguese language.

## 6. ACKNOWLEDGMENTS

This work was partially funded by the ESPRIT Long Term Research project 20077 SPRACH and PRAXIS XXI project 1654. An acknowledgment goes to the PÚBLICO newspaper for making its texts available for this work. An acknowledgment is also given to Prof. Isabel Trancoso to make available the SAM database and to provide the rule based phonetic transcriptions.

## REFERENCES

[1] H. Steeneken and D. Van Leeuwen, *Multi-lingual Assessment of Speaker Independent Large Vocabulary Speech Recognition Systems: the SQALE Project*, in Proceedings EUROSPEECH 95, pp. 1271-1274, Madrid, Spain, 1995.

[2] H. Bourlard, N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer, 1994.

[3] J. Neto, C. Martins and L. Almeida, *An Incremental Speaker-Adaptation Technique for Hybrid HMM-MLP Recognizer*, in Proceedings ICSLP '96, Philadelphia, USA, pp. 1289-1292, 1996.

[4] C. Ribeiro, I. Trancoso and M. Viana, *EUROM.1 Portuguese Database*, Report of ESPRIT Project 6819 SAM-A, Nov. 93.

[5] J. Neto, C. Martins, H. Meinedo and L. Almeida, *The design of a Large Vocabulary Speech Corpus for Portuguese*, in Proceedings EUROSPEECH '97, Rhodes, Greece, 1997.

[6] D. Paul and J. Baker, *The Design for the Wall Street Journal-based CSR Corpus*, in Proceedings ICSLP '92, pp. 899-902, 1992.

[7] *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-5050651996, October 1990.

[2] This system was developed at INESC in collaboration with CLUL, in the group of Prof. Isabel Trancoso.