A LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION HYBRID SYSTEM FOR THE PORTUGUESE LANGUAGE

João P. Neto *Ciro Martins* *Luís B. Almeida*

Instituto de Engenharia de Sistemas e Computadores (INESC), Portugal
Instituto Superior Técnico (IST), Portugal
R. Alves Redol, 9, 1000 Lisboa Codex, Portugal
Phone: +351.1.3100315, Fax: +351.1.3145843
E-mails: jpn@inesc.pt cam@inesc.pt lba@inesc.pt

ABSTRACT

Due to the enormous development of large vocabulary, speaker-independent continuous speech recognition systems, which occur essentially for the US English language, there is a large demand of this kind of systems for other languages. In this paper we present the work done in the development of a large vocabulary, speaker-independent continuous speech recognition hybrid system for the European Portuguese language. This is a difficult task due to the basic development stage of this technology in the European Portuguese language. The development of a system of this kind for a new language depends on the availability of the appropriate source components, mainly a speech corpus and large amounts of texts. This work became possible due to the development of a new database (BD-PUBLICO), a large vocabulary speech corpus for the European Portuguese language developed by us over the last two years.

1. INTRODUCTION

In the last years we saw a significant development of large vocabulary speaker independent continuous speech recognition systems. Unfortunately these developments have been mainly done for the US English language. However this success in terms of both research and commercial systems has been pushing the development of this kind of systems for other languages.

The main goal of the work reported in this paper is to build a large vocabulary, speaker-independent continuous speech recognition hybrid system for the European Portuguese language.

Our group has a large experience in the development of hybrid systems [1] for continuous speech recognition for the English language with a major focus in the development of speaker-adaptation techniques [2] applied to the connectionist component of the hybrid system. Our task now is to port these systems to the European Portuguese language based on our previous experience in the development of this kind of systems. This is a difficult and ambitious task due to the primitive development stage of this technology in the European Portuguese language. We begun by developing and collecting an appropriate database (BD-PUBLICO) in both speech and text [3]. Meanwhile we implemented a baseline system based on an existing database (EUROM.1 SAM Portuguese database) where we developed and tested some techniques for automatic segmentation and labelling in parallel with the development of a basic lexicon and a small language model for the Portuguese language [4].

Based on these very important tools we started the overall process of developing a continuous speech recognition system for the European Portuguese language. Through the new database BD-PUBLICO becomes available a large amount of speech corpora. We developed lexicon and language models for this database. Starting from the baseline system we defined procedures for the initial alignments of the new database. After these developments was possible the training of the acoustic models which took us to a global continuous speech recognition system.

The state of the actual system is far way from being satisfactory still having possibility for further improvements.

In the next section we describe the main features of the BD-PUBLICO database and in section 3 we describe the hybrid system used by us. In section 4 we present the work done on the actual system. At the end we present some points to develop in a near future and some conclusions.

2. THE BD-PUBLICO DATABASE

The development of a large vocabulary, speaker-independent continuous speech recognition system needs an appropriate database of both text and speech. During the last two years we were involved in the design, development and recording of a new large vocabulary speech corpus for the European Portuguese language - the BD-PUBLICO database [3].

For this database we chose as corpus a newspaper's text. Our choice was the Portuguese PÚBLICO newspaper. PÚBLICO is one of the best daily newspapers in the European Portuguese language, with a broad coverage of subjects. In a first phase we collected 6 months of the newspaper text which gave us a total of approximately 11 million words. As we found later this amount of texts was insufficient for a robust estimation of language models for the Portuguese. At this moment we are collecting more texts trying to obtain a larger amount of words.

As recording population we selected students from the *Instituto Superior Técnico* (IST), a large engineering school from the Technical University of Lisbon, with undergraduate and graduate students. This is a young population but with a large variability of accents. The recordings took place in a soundproof room at INESC (Lisbon) and using a desk mounted microphone for the collection of the signal. The recordings were done in the period ranging from April to November 1997.

In this database the speakers were asked to read a set of sentences extracted in paragraph blocks from the newspaper text. This way we are imposing a dictation task but in a speaker-independent mode, given the number of speakers and the quantity of data that we collected for each speaker.

From the new collected BD-PUBLICO database results a training set with 100 speakers (50 male and 50 female) on a total of 8,089 utterances (approximately 22 hours of speech). A development and evaluation test sets were recorded for 10 speakers each (5 male and 5 female) with 40 sentences per speaker. These test sets are restricted to a 5,000 words (5K) vocabulary.

3. HYBRID SYSTEM

In our work we use an hybrid system that combines the temporal modelling capabilities of hidden Markov models (HMMs) with the pattern classification capabilities of multilayer perceptrons (MLPs). 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.

A PLP feature extraction is applied to the speech signal. From this pre-processing phase results a frame with the log energy and PLP-12 cepstral coefficients and their first temporal derivatives. Therefore the feature vector has a total of 26 coefficients. The MLP incorporate local acoustic context via a multiframe input window. We are using a context window of 7 frames (3 frames of left and right context around the central frame). Due to these frames which are appended in the MLP input we have a total of 182 inputs. The resulting network has a single hidden layer with 1,000 units and 39 output context-independent phone classes (about 222,000 weights). The phone classes were the same as in SAM_PA.

4. ACTUAL SYSTEM

The work reported on this section concern the different steps that we made up to date on the development of a continuous speech recognition hybrid system for the European Portuguese language.

Porting a system of this kind to a new language involves identifying the different stages in the process that are language dependent. The various components such as the database, the set of phones, the pronunciation lexicon, the language modelling and the initial segmentation and labelling of the training database are significantly language dependent. Obviously the training and performance of our system will depend on the quality and availability of each of these components.

4.1. Lexicon

After the selection from the BD-PUBLICO database of the several sets (training and test) results a list with 27,833 different words. This list of words was phonetically transcribed by a rule system generating an initial set of pronunciations [4]. In this work we are using the 39 phone classes from the SAM_PA phone set. This rule system has some know difficulties. Due to these problems the lexicon was hand revised by a specialised linguist generating a multipronunciation lexicon. In Table 1 we present the size and the number of pronunciations associated with the training and development test sets.

	Vocabulary dimension	Number of pronunciations
Training Set	15,852	18,320
Development Set	5,000	5,706

Table 1: *Vocabulary dimension and the number of pronunciations associated with the training and development test sets.*

These pronunciations were defined based on linguistic knowledge of the correct form of pronunciation of the language. However in the speech corpus we found a large variability of accents which made the matching between these pronunciations and the speech signals very difficult. One of the improvements that we expect to introduce on the system is the use of smoothing techniques to generate alternative pronunciations based on actual pronunciations and on the likelihood of the acoustic-phonetic models.

4.2. Automatic segmentation and labelling

When developing a system for a new language the initial alignment and labelling of the database is very important. This problem was overcome for the US English through the development of the hand labelled TIMIT database [5].

In a previous work [4] we developed a baseline system based on the EUROM.1 SAM Portuguese database where we tested techniques for automatic segmentation and labelling using the TIMIT database.

The same kind of approach was initially tested for the development of the actual system, but without success. The training of the acoustic models based on the segmentation and labelling resulting from the TIMIT presented frame level classification errors greater than 50%. With this result it was not possible to apply the iterative alignment/training process.

We tested another approach were we begun from our previous baseline system. This approach was successful resulting on an initial classification rate at the frame level of 65%. From this rate it was possible to implement the iterative alignment/training process.

Obviously the resulting alignment is not perfect but was generated in an automatic way achieving more than 70% frame level classification rates. This is a good result given the large phonetic variability present on the speech signals.

4.3. Language modelling

The initial text on the BD-PUBLICO database contain approximately 11 million words. From these 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.

As we have about 9 million words for the language model training and due to the large variability of subjects present on the PÚBLICO newspaper it was not possible to estimate trigram language models as desired. Also for the bigram language model we had just 221,956 pairs for a 5K vocabulary size which is an insuficient number.

The evaluation results proved to us that for the Portuguese, or at least in the case of the BD-PUBLICO database, we will need large amounts of text to be able to robustly estimate stochastic language models. In that sense we begun collecting more texts from the PÚBLICO newspaper and we expect them to be a valuable help in a near future, improving substantially the language models and thus reducing the actual system's word error rate.

4.4. Training and evaluation of the system

The system that we are developing should be speaker independent. However due to the large amount of speech data supplied by the BD-PUBLICO database and our limited computational resources we had to find small systems to bootstrap this development.

On the first approach we tried a small number of speakers (10 male and 10 female) using all the utterances available for that speakers resulting in a total of 1,616 utterances. However the results obtained on the training were very discouraging due to the large variability present on the speech files compared with the small amount of training data.

The next step was to try a different approach where we use all the speakers but with just a subset of the utterances available for each speaker. In this approach we divide the speakers in two sets according to their gender. We selected 30 utterances per speaker ending with 1,500 total utterances. The experiences that we made were using just the male speakers.

This approach was successful in terms of training of the acoustic models. We use a network with a structure of 182 input units (7 input frames with 26 coefficients each resulting from a PLP feature extraction and their corresponding derivatives) a single hidden layer with 1,000 units and 39 output units corresponding to the phonetic context independent classes. We obtained rates of about 70% at the frame level classification on the training. However when we evaluate the system in the 5K development test set using just the male speakers (5 speakers in a total of 202 utterances) we obtained a 63% word error rate. This is a very high error rate. We know that our problem is on the very poor estimation of the language models. We created an alternative language model based on the source texts for the development test set. Obviously this is not a legal language model because we are using the texts from the test set. In this case we obtain a 42% word error rate. As we can see the language model is having a large influence on the final performance of the system and we expect that the word error rate can be further reduced with a more robust language model.

We still tried another approach in terms of training of the acoustic models. We use all the utterances available for each speaker maintaining the division by gender. In this case we had 4,046 sentences from the 50 male speakers. The alignment was based on the previous network and we maintained for computational reasons the same network structure. We are increasing the number of feature vectors maintaining the number of parameters of the MLP. Obviously the next step will be increasing the number of parameters. The evaluation of this system in the 5K development test set results in a 54.5% word error rate with the normal language model, and 35% with the language model extracted from the texts of the development test set.

As we can see this new experience brought a redution of the word error rates and we expect that future trainings increasing the number of parameters of the MLP will bring further reductions. Also with a more robust language model we will certainly achieve a better performance for the system.

5. FUTURE WORK

Due to the initial stage of our system there still is a large scope for improvements. These improvements will be on the different components and possibly some of them will affect our hybrid system's global structure.

In terms of acoustic models we are developing a similar system for the female speakers in much the same way that we developed for the male speakers. Our main problem results from the inability to use large amounts of speech training data. To deal with this problem we plan to use at least two different approaches: a parallel training of different acoustic models followed by their combination and the application of “boosting” methods. Other possible improvement results from the fact that we are using the MLP to estimate context-independent posterior phone probabilities. In the same way as the classical systems we must extend our system to a context-dependent posterior phone probabilities estimation as done before for the hybrid RNN/HMM based on Recurrent Neural Networks [6]. Still on the acoustic models the incorporation of incremental online speaker-adaptation methods will result on a substantial improvement obtained without any extra demands on the speaker, i.e. without an enrolment phase [2].

Where we expect to achieve a substantial improvement is on a more robust language models estimation. Since we are colleting more new texts we will be able to generate more robust N-gram language models with the possibility to use larger N orders.

6. CONCLUSIONS

In this paper we presented the first steps done in the development of a large vocabulary, speaker-independent continuous speech recognition hybrid system for the European Portuguese language. This is a difficult task due to the basic development stage of this technology for the European Portuguese language.

The results obtained so far show some improvements but are far way from being satisfactory. Despite the difficulty of the task associated with the BD-PUBLICO database we certainly will be able to reduce the error rate produced by our system.

The development of a system of this kind must be an iterative process where in each step we have to improve some of the components resulting in the improvement of the overall system. As a consequence there is a new possibility for further improvements on the individual components repeating again the process.

7. ACKNOWLEDGEMENTS

This work was partially funded by the ESPRIT Long Term Research project 20077 SPRACH and PRAXIS XXI project 1654. An acknowledgement goes to the PÚBLICO newspaper for making its texts available for this work.

8. REFERENCES

1. H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994.
2. J. Neto, C. Martins and L. Almeida, *An Incremental Speaker-Adaptation Technique for Hybrid HMM-MLP Recognizer*, Proceedings ICSLP 96, pp. 1289–1292, 1996.
3. J. Neto, C. Martins, H. Meinedo and L. Almeida, *The Design of a Large Vocabulary Speech Corpus for Portuguese*, Proceedings of EUROSPEECH 97, 1997.
4. J. Neto, C. Martins and L. Almeida, *The Development of a Speaker Independent Continuous Speech recognizer for Portuguese*, Proceedings EUROSPEECH 97, 1997.
5. W. M. Fisher, V. Zue, J. Bernstein and D. Pallett, *An acoustic-phonetic database*, 113th Meeting of the Acoustical Society of America, 1987.
6. D. Kershaw, *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System*, Phd Thesis, Cambridge University Engineering Department, 1996.