SPEECH RECOGNITION OF BROADCAST NEWS FOR THE EUROPEAN PORTUGUESE LANGUAGE

Hugo Meinedo, Nuno Souto, João P. Neto

L^{2}F - Spoken Language Systems Laboratory, INESC ID Lisboa / IST
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://l2f.inesc-id.pt/
{Hugo.Meinedo,Nuno.Souto,Joao.Neto}@inesc-id.pt

ABSTRACT

This paper describes our work on the development of a large vocabulary continuous speech recognition system applied to a Broadcast News task for the European Portuguese language in the scope of the ALERT project. We start by presenting the baseline recogniser AUDIMUS, which was originally developed with a corpus of read newspaper text. This is a hybrid system that uses a combination of phone probabilities generated by several MLPs trained on distinct feature sets. The paper details the modifications introduced in this system, namely in the development of a new language model, the vocabulary and pronunciation lexicon and the training on new data from the ALERT BN corpus currently available. The system trained with this BN corpus achieved 18.4% WER when tested with the F0 focus condition (studio, planed, native, clean), and 35.2% when tested in all focus conditions.

1. INTRODUCTION

In the last few years we have seen a large development of speech recognition systems associated to Broadcast News (BN) tasks. These developments open the possibility for new applications where the use of speech recognition is a major attribute.

During the last year and half we have been working on the IST-HLT European programme project ALERT [1]. The main goal of ALERT is to develop a system for selective dissemination of multimedia information. The idea is to build a system capable of identifying specific information in multimedia data consisting of audio/video/text streams, using continuous speech recognition, video processing techniques, audio/video segmentation techniques and topic detection techniques for the three languages of the project (French, German and Portuguese). For more details about the project please see the main web page [1].

In this framework we have a system that continuously monitors a pre-selected TV channel and according to a pre-defined database of shows searches for the start of each show. This is done by a jingle detection module presently based only on audio signals. On our actual system a model of each program's jingle was built. The audio stream is processed and this module will trigger when the beginning of the corresponding show is broadcasted and gives an instruction to start recording the show. The same model is responsible to find the final jingle of the show and stop the recording process. When we start recording a segmentation module is responsible to select the portions of the show to transcribe and for the splitting in small coherent segments. After this segmentation phase the speech recognition system transcribes each segment. The automatic transcriptions are used for a topic detection module that outputs a set of alert topics to be sent to end users through e-mail or fax, according to stored user profiles. The alert topic set is automatically generated, including metadata for each topic, transcripts or summaries out of transcripts, or both and a link to source files (video and sound) as URL-address number.

The work described in this paper concerns the development of a large vocabulary continuous speech recognition system for the Portuguese language to be used in the ALERT framework.

Within the ALERT project we have been working together with RTP (national Broadcast News Television in Portugal) to collect a Portuguese BN database comprising two main corpus. The first with approximately 80 hours to be used as training and test sets for the speech recognition systems. The second with more than 300 hours for the development of automatic topic detection algorithms. The collection of both corpus is completed. The first corpus was automatically transcribed by our baseline speech recognition system *AUDIMUS* [2] and manually corrected. Actually this manual annotation process is not yet finished.

During the last few years we have been developing *AUDIMUS* a continuous speech recognition system for the European Portuguese language. This system was first trained and evaluated using the BD-PUBLICO database [3], where the speakers were asked to read a set of sentences extracted in paragraph blocks from newspaper texts, a corpus of similar characteristics to Wall Street Journal database.

The availability of a BN corpus opened the possibility to an evolution of our system. It was possible to build a system with a new vocabulary with near 64k words, generate a 4-gram language model based on large amounts of text data (more than 384 million words) from newspaper texts and interpolated with the transcriptions texts. After the training process the evaluation results were good enough to derive automatic transcriptions that can be later used for topic detection in a media monitoring system.

In section 2 we will present our *AUDIMUS* system. The system for the BN data is presented in section 3. We start with a brief description of the data available followed by the description of the development of the language model and the vocabulary and pronunciation lexicon. Finally a set of results is presented. Section 4 presents some conclusions along with some issues for future work.

[1] http://alert.uni-duisburg.de/

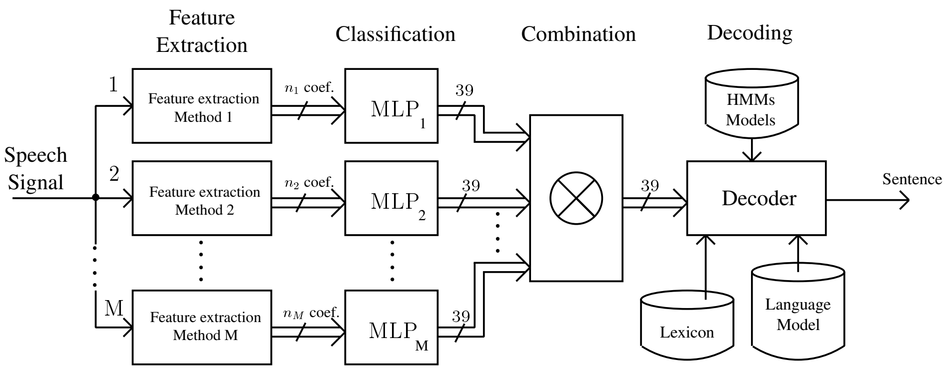

Fig. 1. *Acoustic modelling through the combination of several MLPs trained on distinct feature sets.*

2. *AUDIMUS* SYSTEM

AUDIMUS is a hybrid system [4] that combines the temporal modelling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative 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 estimating context-independent posterior phone probabilities given the acoustic data at each frame. This approach is attractive due to the discriminative capabilities and the need for fewer parameters of the connectionist component of the system. Also this component makes very few assumptions on the form of distribution of the input data. This approach differs from that used by most recognisers due to the estimation of the posterior probability of the word sequence given the acoustic data. Additionally this approach enables the use of posterior probability based pruning in decoding [5].

The acoustic modelling of *AUDIMUS* combines phone probabilities generated by several MLPs trained on distinct feature sets resulting from different feature extraction processes. These probabilities are taken at the output of each MLP classifier and combined using an appropriate algorithm [6]. The processing stages are represented in Figure 1. All MLPs use the same phone set constituted by 38 phones for the Portuguese language plus the silence. The combination algorithm merges together the probabilities associated to the same phone. If we use M MLPs the new probability value for phone y will result from the merging operation of M probability values, each one resulting from a different MLP.

The combination algorithm multiplies the probability values, which internally to the decoder corresponds to perform an average in the log-probability domain. The new *a posteriori* probability value for phone y given the acoustic vector $\mathbf{x}$ is defined as,

$$\hat{P}(y|\mathbf{x}) = \prod_{k=1}^{M} P_k(y|\mathbf{x}) \tag{1}$$

We are using three different feature extraction methods and MLPs with the same basic structure. The feature extraction methods are PLP [7], Log-RASTA [7] and MSG [8]. For the first two methods the baseline recognition systems use log-energy and PLP/Log-RASTA 12^{th} order cepstral coefficients and their first temporal derivatives summing up to 26 parameters per frame. The system based on the MSG method uses 28 coefficients generated by the process for each frame. The MLP classifiers incorporate local acoustic context via a multiframe input window of 7 frames (3 frames of left and right context around the central frame). The resulting network has a single hidden layer with 500 units and 39 output units corresponding to the context-independent phone classes.

We use a decoder with an efficient search strategy based on stacks, using posterior phone probability estimates, generated by the MLP, based pruning [5].

3. DEVELOPMENT OF A BN SYSTEM

In this section we will present the developments of *AUDIMUS* associated with a Broadcast News recognition task. We will start by presenting the Portuguese part of a BN database collected and presently available as a preliminary output of the ALERT project. Next we will discuss the building of robust language models and the vocabulary and pronunciation lexicon. A short review of the automatic segmentation and classification of BN data will be presented followed by the alignment and training process and finally by a set of evaluation results.

3.1. BN database

The database defined in the scope of the ALERT project was divided into 4 parts: a Pilot Corpus, a Speech Recognition Corpus, a Topic Detection Corpus and a Text Corpus. In this paper we are interested in the work done in terms of the Portuguese part of the database.

The basic idea associated with the Pilot Corpus was to serve as a testbed for assessing the adequacy of the methodologies for data collection, annotation and distribution. It should be representative of the type of data that would be collected under the other corpora. In this corpus we have approximately 5 hours of data. The Speech Recognition Corpus was the main part for training and evaluation of the speech recognition system with an amount of approximately 75 hours. The Topic Detection Corpus serves the goal of developing topic detection techniques with a minimum amount of 300 hours. Finally for the Text Corpus was established a minimum amount of 100 million words. In our case that amount was available previously to the project. The new corpora that are important for the present work are the Pilot Corpus and the Speech Recognition Corpus, that we will describe in more detail next.

3.1.1. Pilot Corpus

For this corpus we selected a set of shows, in a total of 11, presented in the two main channels of RTP (Channel 1 and Channel 2). The set of shows is representative of the programs that we want to monitorize under the ALERT project. The shows were manually segmented, annotated and transcribed. RTP as data provider was responsible for collecting the data at their premises and for making the annotation. The annotation was made through the Transcriber tool following the LDC Hub4 (Broadcast Speech) transcription conventions. INESC was responsible for helping training the annotator, validating the annotation and for packaging the data. For each show we have 4 files: the audio file, recorded at 44.1KHz mono with 16 bit/sample, a MPEG-1 video file, the transcription file and a worksheet file with a synthetic summary for each story.

3.1.2. Speech Recognition Corpus

This corpus had the same thematic orientation as the Pilot Corpus. In this case we didn't collect the MPEG files and the audio was collected at 32KHz. The training set was collected on Oct./Nov. 2000, one full month with more than 61 hours, the development set in the first week of Dec. 2000, with more than 8 hours and the evaluation set in the second week of January 2001 with approximately 6 hours in a total of more than 75 hours of data. In this amount are included the initial jingles of the programs and some commercial breaks in the large duration shows. The annotation of this corpus follows the same rules as in the Pilot Corpus. However instead of starting from scratch we provided an initial automatic transcription of the shows. Based on our previous system [2, 6] and on the Pilot Corpus data we trained an initial system adapted to the BN data. With this system was possible to generate an automatic transcription. This transcription was then manually verified and augmented with the necessary annotations to complete the process. This process is not yet completed for the entire corpus. At present time we have available a total of approximately 17 hours. This training set was augmented with the Pilot Corpus resulting in a total of approximately 22 and half hours for training of the system. After removing the commercial breaks, the initial jingles and some parts with overlap speech we got a net total of approximately 18 and half hours of training data. For the system evaluation we are using the development test set presently available with a net duration of approximately 3 hours.

3.2. Language modelling

Our previous language models were based on "Público" newspaper editions available on the web in a total of approximately 46 million words. Recently a new corpus [9] containing the "Público" editions from 1991 to 1998 was made available. This new corpus containing 180 million words had some partial overlap with our previous texts. We augmented these texts with some other Portuguese newspapers available on the web till the end of 2000. Not using the overlapping texts we got a total of 335 million words, still a limited amount when compared with other languages. A large percentage (more than 66%) were from the "Público" newspaper. Very recently we augmented the texts even more by using the first six months of newspapers from 2001. We now have a total 384 million words.

Based on these texts we extracted a backed-off 4-gram language model. The models were trained using version 2 of the CMU-Cambridge SLM Toolkit [10].

We also used the BN transcription texts in order to model more realistically spontaneous speech. For the moment we have only a very limited amount of text, with only 191 thousand words. The language model made from these transcriptions yields a perplexity of 552 while the one made with newspaper texts obtains 150. A language model having a perplexity of 140 was created by interpolating both previous models with weight factors 0.825 (newspapers) and 0.175 (transcriptions).

3.3. Vocabulary and pronunciation lexicon

Generally the BN systems developed for other languages are based on 64K vocabulary sizes. In this work we followed the same initial approach. However we expect that to have a reasonable coverage of the Portuguese language we will need a larger vocabulary size. Nevertheless this problem can be circumvented depending on the specific application. Some applications can feed additional information which could help us to select/update the vocabulary. This is the case of the ALERT project where we could use previous information and some parallel data like newspaper texts to help us selecting new words and reducing the OOV rate.

From the 335 million words text set we extracted 427K different words. Around 100K occur more than 50 times in these texts. These words were selected and classified according to syntactic classes. From that set of words we selected a subset of 56K based on the weighted frequency of their occurrence in the text corpus according to the syntactic classes.

This set was augmented with basically all words from the transcripts training data giving a total of 57,564 words. The margin to a 64K vocabulary will be to incorporate the new words of the training data missing. Actually in the training set we had 12,812 different words in a total of 142,547 words.

For our present development test set corpus which has 5,426 different word in a total of 32,319 words, the number of OOVs using the 57K word vocabulary was 444 words representing a OOV word rate of 1.4%.

In order to build the pronunciation lexicon, we searched through different lexica available in our lab. For the words not available in those lexica (mostly proper names, foreign names and some verbal forms), we used an automatic grapheme-to-phone system to generate the corresponding pronunciations.

3.4. Alignment and training

The acoustic models were improved due to the availability of more transcribed BN data. All this BN training data was aligned using our previous system (*align 3*) and used for training the MLPs. Additionally, these new acoustic models, referred as *align 4*, were also trained to model some noises produced by the speaker. In this way we are trying to model more accurately the higher acoustic variability normally found in spontaneous speech.

Noise modelling was accomplished using an additional MLP output and a corresponding HMM. Initially the silence symbol was substituted in the desired outputs of *align 3* by the noise symbol for the locations where noise occurred. This proceeding was somewhat crude because it lacks accurate time modelling of the noise events but served us to bootstrap the initial training. Once this training was completed a new alignment was made (*align 4*) and the MLPs were re-trained.

3.5. Automatic Segmentation and Classification

The segmentation of BN speech data was accomplished using the KL2 distance metrics [11] evaluated over the PLP coefficients, extracted from audio signal, using two bordering windows of 0.5 seconds. We considered a segment boundary when the KL2 distance reached a maximum. These maxima were selected using a threshold detector. By using a small analysis window a high degree of time accuracy is obtained.

After this segmentation stage the classification algorithm calculates the mean entropy value for each audio segment. The entropy was calculated using the PLP acoustic model probabilities averaged for every frame within the audio segment. High value of entropy indicate a failure of the acoustic model and are likely candidates to be regions of music, non-speech or very degraded speech.

This automatic segmentation and classification module has been used along with the recognition system to produce an automatic transcription that is later manually hand corrected.

An evaluation using BN data almost constituted by spontaneous speech obtained a degradation of 8% relative in WER for the same decoding time when compared to the manual segmentation and classification. This degradation was due to incorrect sentence segmentations points and is difficult to correct in spontaneous speech without higher level knowledge.

3.6. Experiments

Table 1 summarizes the evaluation results for the different tests conducted. We see that the new acoustic models (*align 4*) trained using more BN data achieves a relative improvement of 14%. The language model interpolated using both newspapers text and transcriptions text provided an additional gain of 5% relative.

		% WER	
		57K	
System	LM	F0	All
align 3	newspapers	21.4	41.5
align 4	newspapers	19.4	35.6
align 4	transcriptions	34.3	49.2
align 4	interpolated	18.4	35.2

Table 1. *Evaluation of the successive systems over the test set.*

4. CONCLUSIONS

This paper reports our work on the development of a large vocabulary continuous speech recognition system applied to a Broadcast News task for the European Portuguese language in the scope of the ALERT project.

We started by presenting the baseline recogniser AUDIMUS, which was originally trained and tested with a corpus of read newspaper text. This system was modified to encompass a 4-gram language model and a new 57K vocabulary and trained on new data from the ALERT BN corpus.

Using the subset of the ALERT BN corpus currently available we have built a training corpus of 18 and half hours and a test corpus of 3 hours. The system trained with this BN corpus achieved an 18.4% WER when tested with the F0 focus condition (studio, planned, native, clean), and 35.2% when tested in all focus conditions.

Given the amount of the BN data available and used for training, the improvements that can be expected merely through training with a large annotated corpus of BN could not therefore be fully accomplished yet. The availability of more training data will result on the upgrade of the language model, based on the texts from the training transcriptions, that together with further training of the system, will bring a substantial improvement.

5. ACKNOWLEDGMENTS

This work was partially funded by IST-HLT European programme project ALERT and by FCT project POSI/33846/PLP/2000. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

6. REFERENCES

[1] R. Amaral, T. Langlois, H. Meinedo, J. Neto, N. Souto and I. Trancoso, "The development of a Portuguese version of a media watch system", In Proceedings EUROSPEECH 2001, Aalborg, Denmark, 2001.

[2] J. Neto, C. Martins and L. Almeida, "A large vocabulary continuous speech recognition hybrid system for the Portuguese language", In Proceedings ICSLP 98, Sydney, Australia, 1998.

[3] J. Neto, C. Martins, H. Meinedo and L. Almeida, "The Design of a Large Vocabulary Speech Corpus for Portuguese", In Proceedings of EUROSPEECH 97, Rhodes, Greece, 1997.

[4] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Press, 1994.

[5] S. Renals and M. Hochberg, "Start-synchronous search for large vocabulary continuous speech recognition", IEEE Trans. Speech and Audio Processing 7, pp. 542-553, 1999.

[6] H. Meinedo and J. Neto, "Combination of acoustic models in continuous speech recognition hybrid systems", In Proceedings ICSLP 2000, Beijing, China, 2000.

[7] H. Hermansky, N. Morgan, A. Baya and P. Kohn, "RASTA-PLP Speech Analysis Technique", In Proceedings ICASSP 92, San Francisco, USA, 1992.

[8] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram", Speech Communication, 25:117–132, 1998.

[9] Paulo Rocha and Diana Santos, "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa", In Proceedings PROPOR'2000, Brasil, 2000 (in Portuguese) [http://cgi.portugues.mct.pt/cetempublico/].

[10] P. R. Clarkson and R. Rosenfeld, "Statistical Language Modelling Using the CMU-Cambridge Toolkit", in Proceedings of EUROSPEECH 97, Rhodes, Greece, 1997.

[11] M. A. Sliegler, U. Bain, B. Raj and M. Stern, "Automatic Segmentation, Classification and clustering of Broadcast News", In DARPA Proc. Speech Recognition Workshop, Morgan Kaufmann, 1997.