Automatic Speech Annotation and Transcription in a Broadcast News task

Hugo Meinedo, João Neto

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

Abstract

This paper describes our work on the development of a system for automatic speech transcription applied to a Broadcast News (BN) task for the European Portuguese language. We developed audio segmentation modules including a scheme for tagging certain speaker clusters (anchors). We developed a speech recognition system for the broadcast news task using appropriate models. The tests were conducted using large quantities of BN data and show good results in terms of word error rate and processing time. This system is currently integrated in a prototype audio indexing and document retrieval system that is daily processing the main news show of the national Portuguese broadcaster.

1. Automatic transcription of BN data

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

During the last two years we have been working on the IST-HLT European programme project ALERT. 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/text streams, using continuous speech recognition, audio segmentation techniques and topic detection techniques [1, 2, 3].

This paper describes in more detail the development of our speech recognition engine namely the acoustic models alignment and training procedures, the vocabulary, lexicon and language model building and the decoding algorithm used. We will also describe the audio segmentation, classification and clustering module used to pre-process the audio. Finally we will present our BN speech recognition results and evaluate the impact of the automatic segmentation modules in the recognition.

2. Audio segmentation, classification and clustering

Audio segmentation is used in order to deliver to the user only the relevant information and to generate a set of acoustic cues to the speech recognition system and the topic detection algorithms. This work results in the segmentation of audio into homogeneous regions according to background conditions, speaker gender and special speaker id (anchors). This segmentation can provide useful information such as division into speaker turns and speaker identities, allowing for automatic indexing and retrieval of all occurrences of a particular speaker. If we group together all segments produced by the same speaker we can perform an automatic online adaptation of the speech recognition acoustic models to improve overall system performance.

We use several modules for segmentation, classification and clustering of each news show before proceeding to the speech recognition system. The purpose of the segmentation module is to generate homogeneous acoustic audio segments. The segmentation algorithm tries to detect changes in the acoustic conditions and marks those time instants as segment boundaries. Each homogeneous audio segment is then passed through the first classification stage in order to tag non-speech segments. All audio segments go through the second classification stage where they are classified according to background status. Segments that were marked as containing speech are also classified according to gender and are subdivided into sentences by an endpoint detector. All labeled speech segments are clustered separately by gender in order to produce homogeneous clusters according to speaker and background conditions. In the last stage an anchor detection is done, attempting to identify those speaker clusters that were produced by one of the pre-defined news anchors.

2.1. Audio Segmentation

The main goal for the segmentation is to divide the input audio stream into acoustically homogeneous segments. This is accomplished by evaluating, in the cepstral domain, the similarity between two contiguous windows of fixed length that are shifted in time every 10ms. In our system we introduce three distinct time analysis window pairs of 0.5, 1.0 and 4.0 seconds. Test results were obtained by two segmentation modules in a news program with total duration of 1 hour. When comparing our scheme of analysis windows with different sizes (0.5, 1.0 and 4.0 sec) against a system [4] with a single window pair of 0.5 sec, we were able to reduce significantly the number of missed boundaries, from 22% to 14%, although at the cost of increasing slightly the insertion rate, from 17% to 18%.

2.2. Speech / Non-Speech discrimination

After the acoustic segmentation stage each segment is classified using a speech / non-speech discriminator [5, 6], tagging audio portions without speech, with too much noise or pure music. This stage is very important for the rest of the processing since we are not interested in wasting time trying to recognize audio segments that do not contain “useful” speech. Classification tests were obtained for a subset of the ALERT development test set consisting of 4 different news programs with a total of 2 hours. We achieved with this classifier a very low error rate of 4.4% for tagging speech segments as non-speech. It is the worst error since these segments had useful speech and will not be recognized.

2.3. Gender and Background Classification

In our framework, gender classification is used as a mean to improve speaker clustering. By separately clustering each gender class we will have a smaller distance matrix when evaluating cluster distances which effectively reduces the search space. It also avoids short segments having opposite gender tags being erroneously clustered together. Background classification can be used to switch between tuned acoustic models in recognition and can help to detect better special situations like anchor filler sections with background music. The classification module [6], uses two MLP estimating posterior probabilities. One for the gender classification and the other for background status. In both cases, the output class is chosen through maximum likelihood calculation over the audio segment. Gender classification is very precise with a 7.1% misclassification error rate. The background classifier has a very difficult task because in the training material there are many overlapping, especially music plus noise.

2.4. Sentence Division

Segments that were labeled as containing speech are divided into sentences by an energy endpoint detector. This is a crude and simple approximation that assumes a speech pause will correspond to an end-of-sentence. Unfortunately the news reporters and news anchors not always do a breath pause at the end-of-sentence points. This is the major source for incorrect sentence boundaries.

2.5. Speaker Clustering

The goal of speaker clustering is to identify and group together all speech segments that were produced by the same speaker. The clusters can then be used for an acoustic model adaptation in order to improve the speech recognition rate. Speaker cluster information can also be used by topic detection and story segmentation algorithms to determine speaker roles inside the news show allowing for easier story identification. Our speaker clustering algorithm makes use of gender detection. Speech segments with different gender classification are clustered separately. We used bottom-up hierarchical clustering [4]. We developed an efficient distance measure based on the Bayesian Information Criterion (BIC) [7, 8]. An adjacency term is used instead of the BIC threshold λ [6]. Empirically clusters having adjacent speech segments are closer in time and the probability of belonging to the same speaker must be higher. Using this we obtained a cluster purity greater than 97% with a mean number of clusters per speaker slightly over 3.1. The adjacency term in our modified BIC expression retained a high cluster purity and decreased significantly the number of clusters per speaker. The clustering algorithm proved to be sensitive not only to different speakers but also to different acoustic background conditions. This side-effect is responsible for the high number of clusters per speaker obtained in the test set results.

2.6. Anchor Detection

Anchors introduce the news and provide a synthetic summary for the story. Normally this is done in studio conditions (clean background) and with the anchor reading the news. Anchor speech segments convey all the story cues and are invaluable for automatic topic and summary generation algorithms. Also in these speech segments the recognition error rate is the lowest possible.

The news shows that our system is currently monitoring are presented by three anchor persons, two male and one female. We built individual speaker models for these anchors [6]. During the processing of a news show, after speaker sentence clustering, the resulting clusters are compared one by one against the special anchor cluster models to determine which of those belongs to one of the news anchors. We were able to achieve a percentage of deletions, that is, clusters not identified as belonging to the anchor around 9%, and percentage of insertions, that is clusters incorrectly labeled as anchor around 2%. These results are very promising especially due to the very low insertion rate.

3. AUDIMUS Recognition System

AUDIMUS is a hybrid speech recognition system [9, 10] that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs) [11]. 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.

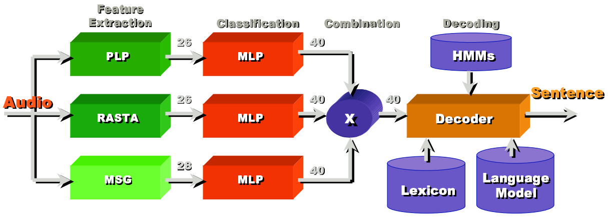

Figure 1: AUDIMUS - Recognition System.

The acoustic modeling of AUDIMUS combines phone probabilities generated by several MLPs trained on distinct feature sets resulting from different feature extraction processes [12]. These probabilities are taken at the output of each MLP classifier and combined using an average in the log-probability domain. The processing stages are represented in Figure 1. All MLPs use the same phone set constituted by 38 phones for the Portuguese language plus silence and breath noises. The combination algorithm merges together the probabilities associated to the same phone.

We are using three different feature extraction methods and MLPs with the same basic structure, that is, an input layer with 7 context frames, a non-linear hidden layer with over 1000 sigmoidal units and 40 softmax outputs. The feature extraction methods are PLP, Log-RASTA and MSG [13].

AUDIMUS presently uses a dynamic decoder that builds the search space as the composition of three Weighted Finite-State Transducers (WFSTs) [14, 15, 16, 17].

4. Language modeling

During the last years we have been collecting Portuguese newspapers from the web, which allowed us to start building a considerably large text corpus. Until the end of 2001 we have texts amounting to a total of 24.0M sentences with 434.4M words.

A language model generated only from newspaper texts becomes too much adapted to the type of language used in those texts. When this kind of language model is used in a continuous speech recognition system applied to a Broadcast News task it will not perform as good as one would expect because the sentences spoken in Broadcast news do not match the style of the sentences written in the newspaper. If we created a language model from Broadcast News transcriptions it would probably be more adequate for this kind of speech recognition task. The problem is that we do not have enough BN transcriptions to generate a satisfactory language model. However we can improve the adaptation of a language model to the BN task by combining a model created from the newspaper texts with a model created from BN transcriptions, using linear interpolation [18].

One of the models is always generated from the newspaper text corpus while the other is a backoff trigram model using absolute discounting and based on the training set transcriptions of our BN database. The optimal weights used in the interpolation are computed using the transcriptions from the evaluation set of our BN database. These are also used to estimate the perplexity of all models.

The final interpolated model has a perplexity of 139.5 and the newspapers model has 148.0. It is clear that even using a very small model based in BN transcriptions we can obtain some improvement in the perplexity of the interpolated model [19].

5. Vocabulary and pronunciation lexicon

When creating the vocabulary we have to consider that it needs to have a limited size to avoid expanding considerably the search space during a speech recognition task. Currently we are limiting our vocabulary size to 64k words. From the text corpus with 335 million words created from all the newspaper editions collected until the end of 2000, we extracted 427k different words. About 100k of these words occur more than 50 times in the text corpus. Using this smaller set, all the words were classified according to syntactic classes. Different weights were given to each class and a subset with 56k words was created based on the weighted frequencies of occurrence of the words. To this set we added basically all the new words present in the transcripts of the training data of our Broadcast News database still being developed, giving a total of 57,564 words. Currently the transcripts contain 12,812 different words from a total of 142,547.

From the vocabulary we were able to build the pronunciation lexicon. To obtain the pronunciations we used different lexica available in our lab. For the words not present in those lexica (mostly proper names, foreign names and some verbal forms) we used an automatic grapheme-phone system to generate corresponding pronunciations. Our final lexicon has a total of 65,895 different pronunciations.

For the ALERT 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%.

6. Weighted Finite-State Transducer based decoder

The integration of knowledge sources in large vocabulary continuous speech recognition using weighted finite state transducers (WFSTs) is spreading in the speech recognition community [20, 21]. The approach can be briefly described in the following way: all knowledge sources (such as the lexicon or the language model) are encoded as weighted finite-state transducers (WFSTs); the knowledge sources are combined using WFST composition, in a very large integrated static network, that is then optimized using well founded algorithms such as determinization, minimization and pushing. In this way, the search space is built by the dynamic composition of the HMM/MLP topology transducer, the lexicon transducer and the language model transducer. The flexibility of the approach comes from the uniformity of representation and combination of the knowledge sources, allowing the easy integration of novel sources. The efficiency comes from the optimization algorithms that make very good use of the sparsity of the integrated network. Furthermore the decoder [14] uses a specialized algorithm integrating linguistic components in run time computing the minimization, pushing and determinization of the composition between the lexicon and language model transducers [15, 16]. In our BN experiments we used the compact representation of WFSTs in memory, and where able to run in less than 512MB of RAM.

7. Acoustic models alignment and training

For training the acoustic models in our BN speech recognition system we started collecting broadcast news data in the scope of the ALERT project. The collected BN data was first automatically transcribed using our baseline speech recognition system AUDIMUS, that was originally developed for dictation tasks and was trained using a clean read speech corpus [9, 10] similar to WSJ0. The automatic transcriptions were then manually corrected. Each time that a significant amount of manually verified BN data was available, we realigned and retrained our acoustic models. So far we made 8 iterations. The first available data, about 5 hours from the ALERT Pilot corpus, was used to bootstrap the BN acoustic models by two iterations of forced alignment and MLP training. The third alignment and training iteration already used some data from the ALERT Speech Recognition corpus, about a full week of news programs. Together with the pilot corpus it consisted on 13 hours of BN data. For the acoustic models trained using alignment number 4, over 22.5 hours of BN data were used. In the alignment number 6 we were able to use the complete ALERT Pilot corpus and Speech recognition corpus consisting of 45 hours of training material.

8. Speech recognition results

Speech recognition results were conducted in the ALERT development test set which has over 6 hours of BN data. The experiments were made using a Pentium III 1GHz computer running Linux with 1Gb RAM. Table 1 summarizes the word error rate (% WER) evaluation obtained by the recognition system. The lines in Table 1 show the increase in performance by each successive improvement to the recognition system.

The first column of results refers to the F0 focus condition where the sentences contain only prepared speech with low background noise and good quality audio. The second results column comprises the WER obtained in all test sentences, including noise, music, spontaneous speech, telephone speech, non-native accents and also including the F0 focus condition sentences.

		% WER		
MLPs	Decoder	F0	All	xRT
1000-a	stack	18.3	33.6	30.0
1000-a	WFST	18.7	32.0	6.4
1000-b	WFST	18.8	31.6	4.8
2000	+ min det L	18.0	30.7	4.3
4000	“	16.9	29.1	3.7

Table 1: BN speech recognition evaluation using the complete development test set.

The first line of results in Table 1 were obtained using MLPs with 1000 hidden units and our old stack decoder algorithm. Compared with the second line of results we see that there was a significant increase in performance obtained when we switched to the new WFST dynamic decoder, especially in decoding time, expressed in the last column as real-time speed. The third line of results refers to the recognition results after alignment number 6 which was the first to use the complete BN training corpus totalizing 45 hours of speech data. The fourth line of results shows the improvement obtained by substituting the determinized lexicon transducer by one that was also minimized. The last line shows 8% relative improvement obtained from increasing the hidden layers to 4000 units. This increase was necessary because the acoustic models MLPs were no longer coping with all the variability present in the ALERT BN corpora that were used as training data.

9. Impact of segmentation in recognition

Our automatic segmentation modules pre-process the audio stream and tag the segments that are fed to the speech recognition system for transcription. Since this automatic segmentation is not perfect we wanted to evaluate its impact on speech recognition in terms of word errors and processing time. We conducted a series of tests using a news program with 29 minutes from the ALERT development test set. For this news program the automatic transcription was produced considering three different audio pre-processing situations. In the first situation, the sentences for transcription and their divisions were chosen using the manual segmentation information. The program was divided into 241 useful sentences. In the second situation, no pre-processing was done and the whole program was considered a single sentence. In the third situation the program was processed by our automatic audio segmentation system, by which the news show was divided into 366 sentences.

Table 2 summarizes the word error rate results obtained and the processing time expressed as multiples of real-time speed.

			xRT	
Segmentation	sent.	% WER	front-end	decod
manual	241	26.9	1.3	2.7
without	1	27.1	1.4	18.0
automatic	366	29.0	1.9	2.9

Table 2: Audio segmentation impact on BN speech recognition.

The first conclusion drawn from the inspection of the results presented in Table 2 is that our automatic audio segmentation module produced more sentences than it should when compared with the manual segmentation. Sentences are shorter and more subdivided. This is mainly due to incorrect sentence divisions in places where there are speaker breath pauses. These subdivided sentences originate language modeling errors near the incorrect boundaries. Without automatic segmentation, that is, when we considered the whole program as a single sentence, the opposite effect occurs although not as problematic, because the language model only introduces article words to connect different sentences. That is why the WER for this case is closer to the manual segmentation.

Additionally, from Table 2 we see that in the situation without segmentation the decoding is very time consuming because of the search space dimension (the whole news program is considered a single sentence). Our automatic audio segmentation takes slightly more processing time than the manual segmentation, as it should due to the segmentation and classification modules.

10. Concluding remarks

Broadcast News speech recognition is a difficult and very resource consuming task. Our recognition engine evolved substantially through the accumulation of relatively small improvements. We are still far from ideal recognition results, the ultimate goal, nevertheless this technology is able to drive very useful applications, including audio indexing and spoken document retrieval.

Our automatic audio segmentation, classification and clustering modules proved fairly accurate while consuming relatively less computational resources than other approaches. Additionally we developed a scheme for tagging anchor speaker clusters using trained models which aids the story segmentation module of our document retrieval system.

Our recognition system evolved substantially with the availability of large quantities of BN training data. The MLPs with 4000 units trained on all of this BN data permitted a significant decrease in WER. Furthermore, the use of interpolated language models showed that is possible to improve a newspaper texts based language model using a small BN transcriptions based model. The use of dynamic WFST decoder allowed us to improve the speed and also reduce WER from over 30 to as little as 4 times real-time speed.

Our system is currently integrated in a prototype audio indexing and document retrieval system that is daily processing the main news show of the national Portuguese broadcaster [2].

11. 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". Hugo Meinedo is sponsored by a FCT scholarship (SFRH/BD/6125/2001).

12. 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 *Proc. EUROSPEECH 2001*, Aalborg, Denmark, 2001.

[2] J. Neto, H. Meinedo, R. Amaral, and I. Trancoso, "A system for selective dissemination of multimedia information resulting from the alert project," in *Proc. ISCA MSDR 2003*, Hong-Kong, China, April 2003.

[3] R. Amaral and I. Trancoso, "Segmentation and indexation of broadcast news," in *Proc. ISCA MSDR 2003*, Hong-Kong, China, April 2003.

[4] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news," in *DARPA Proc. Speech Recognition Workshop*, 1997.

[5] G. Williams and D. Ellis, "Speech/music discrimination based on posterior probability features," in *Proc. EUROSPEECH 1999*, Budapest, Hungary, September 1999.

[6] H. Meinedo and J. Neto, "Audio segmentation, classification and clustering in a broadcast news task," in *Proc. ICASSP 2003*, Hong-Kong, China, April 2003.

[7] S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA Proc. Speech Recognition Workshop*, 1998.

[8] B. Zhou and J. Hansen, "Unsupervised audio stream segmentation and clustering via the bayesian information criterion," in *Proc. ISCLP 2000*, Beijing, China, October 2000.

[9] J. Neto, C. Martins, and L. Almeida, "The development of a speaker independent continuous speech recognizer for portuguese," in *Proc. EUROSPEECH 1997*, Rhodes, Greece, 1997.

[10] J. Neto, C. Martins, and L. Almeida, "A large vocabulary continuous speech recognition hybrid system for the portuguese language," in *Proc. ICSLP 1998*, Sydney, Australia, 1998.

[11] H. Bourlard and N. Morgan, *Connectionist Speech Recognition - A Hybrid Approach*, Kluwer Academic Publishers, Massachusetts, EUA, 1994.

[12] H. Meinedo and J. Neto, "Combination of acoustic models in continuous speech recognition," in *Proc. ICSLP 2000*, Beijing, China, 2000.

[13] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Comunication*, vol. 25, pp. 117–132, 1998.

[14] D. Caseiro and I. Trancoso, "A decoder for finite-state structured search spaces," in *ASR 2000 Workshop*, Paris, France, Sept. 2000.

[15] D. Caseiro and I. Trancoso, "On integrating the lexicon with the language model," in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, September 2001.

[16] D. Caseiro and I. Trancoso, "Transducer composition for "on-the-fly" lexicon and language model integration," in *Proc. ASRU 2001 Workshop*, Madonna di Campiglio, Trento, Italy, December 2001.

[17] D. Caseiro and I. Trancoso, "A tail-sharing wfst composition for large vocabulary speech recognition," in *Proc. ICASSP 2003*, Hong-Kong, China, April 2003.

[18] R. Rosenfeld, *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*, Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Apr. 1994, also appears as technical report CMU-CS-94-138.

[19] H. Meinedo, N. Souto, and J. Neto, "Speech recognition of broadcast news for the european portuguese language," in *Proc. ASRU 2001 Workshop*, Madonna di Campiglio, Trento, Italy, December 2001.

[20] M. Mohri, "Finite-state transducers in language and speech processing," *Computational Linguistics*, vol. 23, no. 2, pp. 269–311, June 1997.

[21] M. Mohri, F. Pereira, and M. Riley, "Weighted automata in text and speech processing," in *ECAI 96 Workshop*. Budapest, Hungary, August 1996.