A Stream-based Audio Segmentation, Classification and Clustering Pre-processing System for Broadcast News using ANN Models

Hugo Meinedo, João Neto

L^{2}F - Spoken Language Systems Laboratory
INESC-ID Lisboa / IST
Rua Alves Redol, 9 1000-029 Lisboa, Portugal
{hugo.meinedo, joao.neto}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract

This paper describes our work on the development of a low latency stream-based audio pre-processing system for broadcast news using model-based techniques. It performs speech/non-speech classification, speaker segmentation, speaker clustering, gender and background conditions classification. As a way to increase the modelling accuracy our algorithms make extensive use of Artificial Neural Networks (ANN) thus avoiding the rough assumptions normally made about the audio signal distribution. Experiments were conducted on the COST278 multilingual TV broadcast news database and compared with current state of the art algorithms using standard evaluation tools. Additionally we investigated the impact of automatic audio pre-processing system within the recognition using a large broadcast news test database for the European Portuguese. These tests show a small degradation in recognition performance when compared with hand labelled audio segmentation. Our system is part of a prototype close-captioning system that is daily processing the main news show of two Portuguese Broadcasters.

1. Introduction

Broadcast News media monitoring is an important technology but poses a number of difficulties and challenges for speech processing. We have been building a media monitoring system for TV broadcast news. In this kind of application the speech signal not only has to be transcribed but also characterized in terms of acoustic content.

To accomplish this characterization the first stage in our media monitoring system is an audio pre-processor. This pre-processor is responsible for i) the segmentation of the signal into acoustically homogeneous regions, ii) for classification of those segments according to background conditions, speaker gender and iii) for identifying all segments uttered by the same speaker. The segmentation provides information regarding speaker turns and identities allowing for automatic retrieval of all occurrences of a particular speaker. The segmentation can also be used to improve performance through adaptation of the speech recognition acoustic models. Additionally the final transcriptions enriched by the pre-processing information are somewhat more human readable. A near-term objective of our media monitoring system is to be able to function in real-time in order to provide automatic close-captioning. To accomplish this not only the pre-processor modules must have been designed for online processing (capable of working in stream-based mode) but also the system needs to have faster than real-time processing time. It is also a constraint that this must be achieved with low constant latency so not to broaden the gap between input signal and output transcriptions. Furthermore, our system makes extensive use of Artificial Neural Network models in an attempt to model more precisely the audio signal without increasing significantly the complexity.

This paper describes our work on the development of a low latency stream-based audio pre-processing system for broadcast news. The paper is organized as follows: section 2 reviews the databases used for training and evaluation purposes. The next four sections introduce the audio pre-processor in detail. Section 7 presents an brief overview of our speech recognition and an assessment of errors introduced by automatic segmentation. The paper ends with some concluding remarks.

2. Broadcast News corpora

For the training and evaluation of the system two different databases were used. One has appropriate size for training complex models and for speech recognition. The other database was used for evaluation of speech segmentation and classification algorithms since it facilitates comparisons with state of the art algorithms being developed by our partners in the European COST278 action [1, 2]. Additionally the evaluation tools used were also developed within the COST278 action and are common to all partners.

2.1. Portuguese-BN corpus

The Portuguese-BN corpus [3] was collected in close cooperation with RTP the Portuguese public broadcast company. Its primary goals were all the news programs, national and regional, from morning to late evening, including both normal broadcasts and specific ones dedicated to sports and financial news. Given its broader scope and larger audience, the 8 o'clock PM news program was selected as the prime target. The Speech Recognition database of the Portuguese-BN corpus is composed by three sets: Training, Development and Evaluation. The Training set is composed of 99 different broadcast news shows with over 46h total time. We used this set to develop the acoustic pre-processor modules. The Evaluation set which has 6h duration was used to assess speech recognition performance using manual and automatic segmentation.

2.2. COST278-BN corpus

The pan-European COST278-BN database was used for evaluation of audio segmentation. At present it consists of 30 hours of broadcast news recordings, divided into ten equally large

national data sets. Each national set contains some complete news shows broadcasted by TV stations in one country or region. The transcription was performed according to a protocol described in [4]. The database covers nine European languages: Belgian Dutch (BE), Portuguese (PT), Galician (GA), Czech (CZ), Slovenian (SI and SI2), Slovak (SK), Greek (GR), Croatian (HR) and Hungarian (HU).

3. Pre-processing System Overview

Our system, shown in Figure 1 is composed by five modules: three for classification (Speech / Non-speech, Gender and Background), one for speaker clustering and one for acoustic change detection. All five modules are model-based meaning that they incorporate algorithms trained using a priori information (from databases). As a way to increase the modelling accuracy our algorithms make extensive use of Artificial Neural Networks (ANN) thus avoiding the rough assumptions normally made about the audio signal distribution, namelly the assumption that fairly large blocks of audio (2 to 3 seconds) can be approximated by Gaussian distribuitions. All ANN used are of the type feed-forward fully connected Multi-Layer Perceptron (MLP) and were trained with backpropagation algorithm.

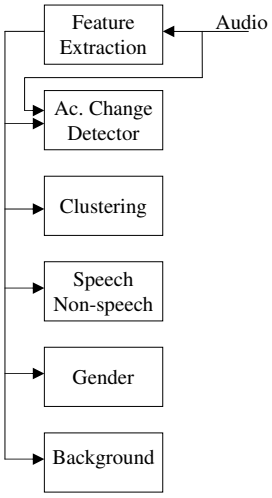

Figure 1: *Pre-processing system overview.*

4. Classification modules

The Speech / Non-speech module is responsible for identifying audio portions that do not contain speech, with too much noise or pure music. This serves two purposes: first, no time will be wasted trying to recognize audio portions that do not contain speech; second, reduces the probability for speaker clustering mistakes.

Gender classification is used to improve speaker clustering. By clustering separately each gender class we 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 could be used to switch between tuned acoustic models (trained separately for clean speech, with noise or in the presence of music). So far it is only being used to enrich the metadata in the final transcription XML file.

All classifiers have the same base architecture. It is composed of a MLP with 9 input context frames of acoustic features each one with 26 coefficients. That is, 12^{th} order PLP features plus log energy plus deltas. The MLP has an hidden layer with 300 sigmoidal units. The output unit of the Speech / Non-speech MLP can be viewed as giving a probabilistic estimate of the input frame being speech or non-speech.

When the Acoustic Change Detector hypothesized the start of a new segment, the first N_{class} frames of that segment are used to calculate the speech/non-speech, gender and background classification. The estimate is chosen through maximum likelihood calculation. After initial experiments in the training set, N_{class} was set to 300 frames. This relatively short interval is a trade-off between performance and the desire for a very low latency time. These MLP classifiers were trained using the whole training set of the Portuguese-BN database. The reference frame classifications were derived from the STM manual reference files.

	Sp %	Non-sp %	Acc %
Sp / Non-sp	97,5	70,6	95,6
	Male %	Female %	Acc %
Gender	96,7	90,2	94,5

Table 1: Frame classification percentages.

Evaluation was conducted using the complete COST278-BN database [4] and the standard tools developed by the COST278-BN SIG. Table 1 sumarizes classification results for the percentage of correctly classified speech frames, non-speech frames and accuracy. The same for Gender, percentage of correctly classified male frames, female frames and accuracy.

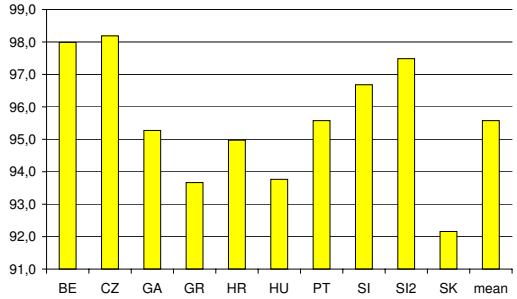

Figure 2: *Accuracy results for Speech / Non-speech.*

Figure 2 represents the graphic for accuracy obtained in each language and the mean value for all languages. We can observe a large variability among different language sets.

When compared against the best algorithms evaluated in [5] using the same database our system achieved similiar results.

5. Acoustic Change Detector

The main goal for the Acoustic Change Detector is to divide the input audio stream into acoustically homogeneous segments. This module uses a hybrid two stage algorithm combining energy, metric and model based techniques and is represented in Figure 3. During the first stage a large set of candidate change points are generated. In the second stage these candidate change points are evaluated again and some that do not correspond to true speaker change boundaries are eliminated.

The first stage uses two algorithms to generate the set of candidate boundaries. The first one is metric-based. This is accomplished by evaluating, in the feature cepstral domain, the similarity between two contiguous windows of fixed length that are shifted in time every 10ms. We used the symmetric Kullback-Liebler, KL2 [6], as the distance measure to evaluate acoustic similarity. The KL2 is calculated over 12^{th} order PLP coefficients extracted from the audio signal. We considered a segment boundary when the KL2 distance reached a maximum. The maxima values are selected using a pre-determined threshold detector. The second algorithm is energy-based. Instantaneous energy is calculated in a frame basis. From this energy values two measures are derived, the median filtered by a 50 frame window and a long term average of 250 frames. A threshold detects when the median signal drops bellow the long term average (small and big pauses in speech discourse). Of course these pauses may not correspond to speaker change (our main goal) but generally they do. These two algorithms complement themselves: energy is good on slow transitions (fade in/out) where KL2 is limited because of fixed length window. Energy tends to miss the detection of rapid speaker changes for situations with similar energy levels while KL2 does not.

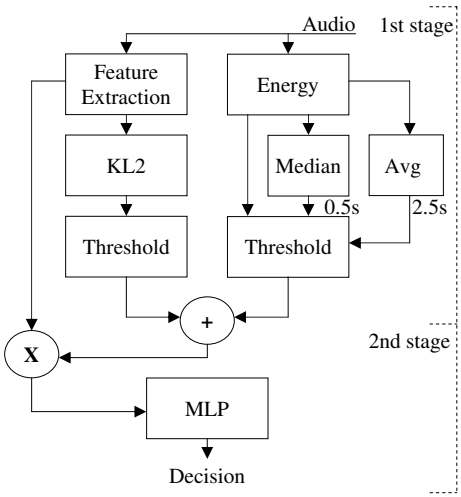

Figure 3: *Acoustic Change Detector.*

The second stage uses a MLP classifier to make a decision whether the candidate boundary should be removed or not. The input of the classifier uses a huge 300 frames context ($N_{acd} =$ 3 sec) of acoustic features with 12^{th} order PLP plus log energy and a hidden layer with 150 sigmoidal units.

The MLP classifier was trained by generating candidate boundaries for the whole training set using the first stage and then aligning those boundaries using the reference STM files. This procedure generated an appropriate boundary training set with balanced positive and negative examples. By using audio feature vectors directly into the MLP input we are using a more precise signal model and not assuming that the PDF of the features of the audio signal is gaussian in the N_{acd} window.

	Recall %	Precision %	F-measure %
Change Detector	78,9	65,5	70,9

Table 2: Evaluation of the Acoustic Change Detector.

Again evaluation was conducted using the COST278-BN database. We used standard measures Recall (% of detected speaker change points), Precision (% of detected points which are genuine change points) and F-measure (defined as $2RP/(R+P)$). Table 2 summarizes the results for the acoustic change detector. When compared with the best algorithms evaluated in [1, 2, 5] using the same database we got good results.

6. Speaker Clustering

The goal of speaker clustering is to identify and group together all speech segments that were uttered by the same speaker. Our algorithm works in the following way: after the acoustic change detector signals the existence of a new boundary and the classification modules determine that the new segment contains speech of male/female gender, the first N_{clus} frames of the segment are compared with all clusters found so far. The segment is merged with the cluster for which the lower distance was calculated if bellow a predefined threshold. The distance of a segment to a cluster is given once more in a model-based approach by a MLP. This classifier was trained to estimate the probability of a given segment of acoustic features belonging to a particular cluster also specified in terms of audio feature vectors.

Our speaker clustering algorithm makes use of gender detection. Speech segments with different gender classification are clustered separately. There are two MLP classifiers, one for each gender type. To represent the cluster several alternatives were tested: the feature vector of one of the cluster elements, an average of all elements feature vectors.

We made $N_{clus} = 300$ frames meaning that the worst case cenario for latency is $N_{clus} + N_{acd} = 600$ frames.

In order to evaluate the clustering, a bi-directional one-to-one mapping of reference speakers to clusters is computed (NIST rich text transcription evaluation script). It defines the correct speaker/cluster for a cluster/speaker. Obviously, unmapped clusters/speakers have no correct speaker/cluster. On the basis of this information, the Q-measure is defined as the geometrical mean of the percentage of cluster frames belonging to the correct speaker and the percentage of speaker frames labeled with the correct cluster. Since these percentages are zero for unmapped clusters/speakers, we have also cluster-speaker pairs. Another performance measure is the Diarization Error Rate (DER) which is defined as the percentage of frames with an incorrect cluster-speaker correspondence.

	Q %	Q_{map} %	DER %
Clustering	68,1	87,8	31,6

Table 3: Evaluation for Speaker Clustering.

After evaluation in the COST278-BN database and comparing with the best algorithms we achieve comparable Q percentages and DER worse must likely due to a higher number of cluster per speaker. Results are summarized in Table 3.

7. Speech Recognition

Our automatic acoustic segmentation system pre-processes the audio stream and tags the segments that are fed to the speech recognition system for transcription. Since the automatic segmentation and classification are not perfect we wanted to evaluate its impact on speech recognition in terms of word error rate (% WER).

AUDIMUS.MEDIA [3] is a hybrid speech recognition system that combines the temporal modelling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of MLPs. The acoustic modelling combines phone probabilities generated by several MLPs trained on distinct feature sets resulting from different feature extraction processes. Currently our vocabulary has 65k words associated to a multi-pronunciation lexicon. The corresponding OOV rate is 1.4%. We use an interpolated 4-gram language model combining a model created from newspaper texts with a model created from the Portuguese-BN training transcriptions (46 h). AUDIMUS.MEDIA presently uses a dynamic decoder that builds the search space as the composition of three Weighted Finite-State Transducers (WFSTs) [7], the HMM/MLP topology transducer, the lexicon transducer and the language model transducer

A series of recognition tests were conducted on the Portuguese-BN evaluation test set. These tests serve to evaluate the performance impact of using our automatic pre-processing acoustic segmentation system in the complete BN transcriptioning system. First, we recognized the test set using the reference segmentation, including the hand labelled sentence segmentation boundaries. In the second test, the recognition used the manual segmentation but without sentence boundaries, that is, segments were sent to the recognizer in blocks of utterances from the same speaker. Finally, the recognition used the segmentation from our automatic pre-processing system. Recognition results are presented in Table 4.

	% WER	
Segmentation	F0	All
Manual with sentence boundaries	11.6	27.3
Manual with speaker blocks	14.8	31.1
Auto	12.8	31.2

Table 4: Acoustic segmentation impact on BN speech recognition. F0 focus condition = planned speech, no background noise, high bandwidth channel, native speech. All = All other acoustic conditions.

The first conclusion drawn from the inspection of the results presented in Table 4 is that our automatic acoustic preprocessing system has a performance comparable to the manual segmentation when utterances are joined in speaker blocks. This is natural since it is exactly what the automatic system tries to achieve (speaker block segmentation). In that sense we have to conclude that these tests show only a small degradation in recognition performance when compared with hand labelled audio segmentation. Finally, we can see that sentence boundaries make a lot of difference in terms of WER. The problem seems to be in the language model which is introducing erroneous words (most article words) trying to connect different sentences.

These results confirm that a small degradation in performance of speech pre-processing is not very important to achieve good recognition results. Furthermore, in our case a sentence boundary segmentation module is crucial for recognition or the decoder must hypothesize the end/beginning of a new sentence in the middle of an input audio sentence.

8. Conclusions

In this paper we described the development of a low latency stream-based audio pre-processing system for broadcast news relying heavily on model-based techniques. We evaluated the performance of the system components using the COST278-BN database that is being used as test bed for comparison of different algorithms. This has the great advantage of simplifying direct comparisons using the same evaluation tools and protocols without the need for implementing all state of the art algorithms. Our system shows very good performance while maintaining a very low latency for stream-based operation. Additionally we investigated the impact of automatic audio preprocessing system within the recognition using the Portuguese-BN large broadcast news test database. The recognition tests show a very small degradation (0.1% absolute) in performance when compared with hand labelled audio segmentation under the same conditions, that is, recognition of utterance blocks from the same speaker.

9. Acknowledgments

The work presented in this paper was performed in the broadcast news Special Interest Group within the COST278 action on Spoken Language Interaction in Telecommunications. This work was partially funded by FCT project POSI/PLP/47175/2002. Hugo Meinedo is sponsored by a FCT scholarship (SFRH/BD/6125/2001).

10. References

[1] A. Vandecatseye and J. P. Martens, "A fast, accurate and stream-based speaker segmentation and clustering algorithm," in *Proc. Eurospeech 2003*, Geneva, Switzerland, September 2003.

[2] J. Zdansky, P. David, and J. Nouza, "An improved preprocessor for the automatic transcription of broadcast news audio stream," in *Proc. ISCLP 2004*, Jeju, Korea, October 2004.

[3] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "Audimus.media a broadcast news speech recognition system for the european portuguese language," in *Proc. PROPOR 2003*, Faro, Portugal, June 2003.

[4] A. Vandecatseye and et al., "The COST278 pan-european broadcast news database," in *Proc. LREC 2004*, Lisbon, Portugal, July 2004.

[5] J. Zibert and et al., "The COST278 broadcast news segmentation and speaker clustering evaluation - overview, methodology, systems, results," in *Proc. Interspeech 2005*, Lisbon, Portugal, September 2005, (Submitted).

[6] 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.

[7] D. Caseiro and I. Trancoso, "Transducer composition for "on-the-fly" lexicon and language model integration," in *Proc. ICASSP '2003*, Hong Kong, China, April 2003.