AUDIO SEGMENTATION, CLASSIFICATION AND CLUSTERING IN A BROADCAST NEWS TASK

Hugo Meinedo, João Neto

L^{2}F - Spoken Language Systems Laboratory
INESC-ID Lisboa / Instituto Superior Técnico
{hugo.meinedo, joao.neto}@inesc-id.pt

ABSTRACT

This paper describes our work on the development of an audio segmentation, classification and clustering system applied to a Broadcast News task for the European Portuguese language.

We developed a new algorithm for audio segmentation that is both accurate and uses less computational resources than other approaches. Our speaker clustering module uses a modified BIC algorithm which performs substantially better than the standard KL2 and is much faster than the full BIC. Finally, we developed a scheme for tagging certain speaker clusters (anchors) using trained cluster models. A series of tests were conducted showing the advantage of the new algorithms. This system is part of a prototype system that is daily processing the main news show of the national Portuguese broadcaster.

1. INTRODUCTION

The last years show a large demand for the monitoring of broadcast news programs with a large variety of applications. We have been developing a system for selective dissemination of multimedia information in the scope of the ALERT project where the user is able to specify which kind of contents he wants to access. To accomplish that goal we have been working in the development of a broadcast news speech recognition system associated with automatic topic detection algorithms. 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 we have been working on audio segmentation, classification and clustering.

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. Some of these features are implemented on our system.

We use several modules for segmentation, classification and clustering of each news show before proceeding to the speech recognition system. The architecture is shown in Figure 1.
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

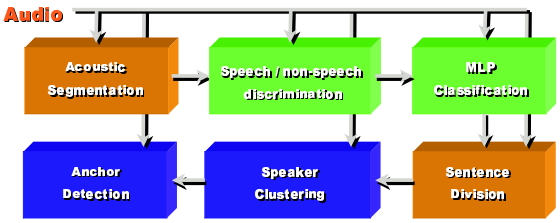

Fig. 1. Segmentation and Classification system overview.

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 labelled 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.

This paper is organized as follows: section 2 describes the audio segmentation module and section 3 presents the speech / non-speech discrimination module. Section 4 details the gender and background classification. The sentence division algorithm is presented in section 5 and the speaker clustering module is described in section 6. Section 7 details the anchor detection module and finally in section 8 some conclusions are drawn.

2. 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. We used the symmetric Kullback-Liebler, KL2 [1], as the distance measure to evaluate acoustic similarity. Each window is modelled by a gaussian distribution. Large values for the KL2 imply that the distributions of the windows are more dissimilar. 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 diagram of our audio segmentation module is shown in Figure 2.

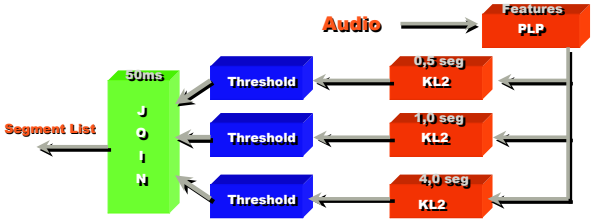

Fig. 2. KL2 audio segmentation.

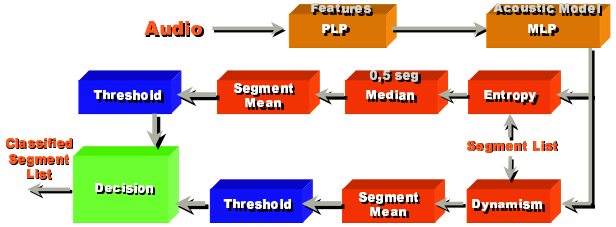

Fig. 3. Entropy + Dynamism Classification.

In our system we introduce three distinct time analysis window pairs of 0.5, 1.0 and 4.0 seconds. Small analysis windows obtain a higher degree of time accuracy. Larger windows have less time accuracy but were able to detect slower audio transitions. The final segment transition list is a weighted sum of the three transition lists evaluated inside a 50 msec window. To the more time accurate segmentation systems, i.e., using smaller windows, were given more importance and can overrule the others that use larger windows.

Table 1 highlights the results obtained by two segmentation modules in a news program with total duration of 1 hour. Errors are presented in standard form, percentage of deleted and inserted boundaries. The first KL2 segmentation module uses a standard single window pair of 0.5 sec and the second KL2 segmentation module uses our scheme with three different window sizes.

	Errors	
Segmenter	Deletions %	Insertions %
single KL2, 0.5 sec	22	17
three KL2, 0.5, 1.0, 4.0 sec	14	18

Table 1. *KL2 Segmentation evaluation.*

Using our scheme of analysis windows with different sizes we reduced significantly the number of missed boundaries although at the cost of increasing slightly the insertion rate. For evaluation purposes, we used a tolerance window of 0.5 sec around the true boundary. This small tolerance window has a significant impact on errors because if a detected boundary is somewhat displaced and is outside the tolerance window two errors will occur: a deletion and an insertion. This segmentation system is sufficiently accurate and at the same time much less computational intensive than for instance the more used BIC [2, 3] that evaluates three full covariance matrices at each time frame.

3. SPEECH / NON-SPEECH DISCRIMINATION

After the acoustic segmentation stage each segment is classified using a speech / non-speech discriminator, 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 recognise audio segments that do not contain "useful" speech.

Figure 3 represents the speech / non-speech classification module. 12^{th} order PLP coefficients are extracted from the audio signal. These feature vectors are input into a Multi-Layer Perceptron (MLP) that was trained to estimate context-independent phone posterior probabilities. This MLP is the same used as acoustic model by our hybrid HMM/MLP recognition system. It was trained using 22 hours of BN data and has an architecture with 7 context input frames of 26 features (12^{th} order PLP coefficients plus energy and delta features), an hidden layer with 1000 sigmoidal units and 40 softmax output units, representing the 38 phones of the European Portuguese plus silence and breath noises.

Local posterior probabilities estimated by the MLP are used to calculate two acoustic confidence measures: instantaneous per-frame entropy and the probability dynamism [4]. The entropy of the K posterior probability estimates associated with HMM states q_k is defined as,

$$H(n) = -\sum_{k=1}^{K} P(q_k|x^n)\log(P(q_k|x^n))$$

where x_n is the acoustic vector at time n and $P(q_k|x^n)$ the posterior probability of phone q_k given x_n at the input. Low values for the entropy indicate regions where the acoustic model provides a good match to the observed input data, since the distribution of phone posteriors will be dominated by a single class phone. High values of entropy represent more uniformly distributed probability values and indicate regions of poorly modelled audio by the acoustic model and are likely candidates to be regions of music, noise or very degraded speech.

This instantaneous entropy measure is inherently noisy due to phone transitions during normal speech and a median filter with a 0.5 sec window was used to smooth the output. Finally, we calculate the average value for the segment.

The probability dynamism measures the rate of change in phone probability estimates and is given by

$$D(n) = \sum_{k=1}^{K}(P(q_k|x^{n-1}) - P(q_k|x^n))^2$$

The value for dynamism in normal speech is high because probability estimates for well modelled speech segments change abruptly and frequently. Non-speech signal are less varying and consequently will receive lower dynamism values. Again, the average for one audio segment is calculated.

Both acoustic confidence measures are threshold and serve as input into a finite-state machine that serves as an hard-decision rule classifier.

The confusion matrix in Table 2 shows the results obtained for a test set that consists of 4 different news programs with a total of 2 hours. We can see that this classifier has 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 recognised.

Speech / Non-Speech	Hypothesis	
Reference	Speech	Non-Speech
Speech	95.6 %	4.4 %
Non-Speech	10.3 %	89.7 %

Table 2. *Confusion matrix for the speech / non-speech classifier.*

4. 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.

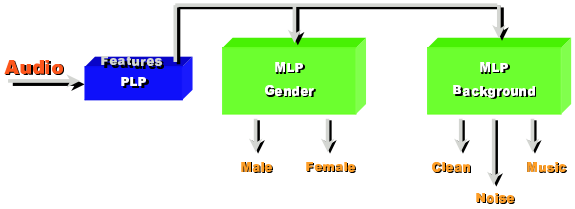

Fig. 4. MLP classification.

The classification module, shown in Figure 4, uses two MLP estimating posterior probabilities. One for the gender classification and the other for background status. Both classifiers use a MLP with 9 input context frames of 12^th order PLP features and a hidden layer with 250 sigmoidal units. The gender MLP has two output classes, male and female, and the Background MLP has three output classes, clean, noise and music. In both cases, the output class is chosen through maximum likelihood calculation over the audio segment.

These classifiers were trained using a subset of our speech recognition training corpus. This subset consists of 11 different news programs with a total of 6 hours. Table 3 summarizes the results obtained by the classification stage when evaluated in a test set with 2308 audio segments.

Gender	Hypothesis	
Reference	Male	Female
Male	95.2 %	4.8 %
Female	2.3 %	97.8 %

Table 3. *Confusion matrix for then gender classifier.*

As we can see from Table 3, the gender classification is very precise showing low misclassification error rates. In Table 4 we show the results obtained testing the background classification module. The background classifier has a very difficult task because in the training material there are many overlapping, especially music plus noise. Furthermore, we found that many hand annotated segments have dubious classifications when certain noises corrupt a normal clean background.

Background	Hypothesis		
Reference	Noise	Music	Clean
Noise	64.6 %	1.0 %	34.4 %
Music	30.2 %	55.4 %	14.4 %
Clean	11.3 %	0.3 %	88.4 %

Table 4. *Confusion matrix for the background classifier.*

5. SENTENCE DIVISION

Segments that were labelled 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.

6. 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 [1]. In this approach, speech segments are modelled in the cepstral domain by a gaussian distribution. Initially each segment is considered a cluster. The algorithm computes a distance matrix for all clusters and the two closer ones are considered for joining in a new cluster. Clusters are linked together until the distances exceed a pre-defined value. At that point the clustering ends. Several appropriate distance measures can be used, namely the KL2 [1], the generalized likelihood ratio or the BIC [2, 3].

Our first experiments were conducted using the KL2 metrics to evaluate cluster distances. Latter on, we developed a more efficient distance measure based on the BIC.

The distance measure when comparing two clusters using the BIC can be stated as a model selection criterion where one model is represented by two separated clusters C_1 and C_2 and the other model represents the clusters joined together $C = \{C1, C2\}$. The BIC expression is given by,

$$BIC = nlog|\Sigma| - n_1log|\Sigma_1| - n_2log|\Sigma_2| - \lambda P$$

where $n = n_1 + n_2$ gives the data size, Σ is the covariance matrix, P is a penalty factor related with the number of parameters in the model and λ is a penalty weight. If $BIC < 0$ the two clusters are joined together.

We made two modifications to this criterion. First, we considered that the gaussian distributions had diagonal covariance matrices, that is, we considered that the features were uncorrelated. Our speaker clustering tests showed that this modified BIC performs

better than the KL2 and at the same time is much less computationally intensive than the full BIC. Second, an adjacency term is used instead of the BIC threshold λ. The new penalty weight is now given by $\lambda = 1 + 1.25k$, where k represents the number of adjacent speech segments between both clusters C_1 and C_2. If the clusters do not have adjacent segments, $\lambda = 1$. When they have adjacent segments, $\lambda > 1$, and the model favouring a single cluster will be less penalized. Empirically clusters having adjacent speech segments are closer in time and the probability of belonging to the same speaker must be higher.

Table 5 illustrates the results for the speaker clustering module using different distance metric criterions in a test set with 3 news shows of over 2 hours. Results are shown in terms of: *mean cluster purity*, defined as the ratio between the number of sentences from the dominating speaker and the total number of sentences in the cluster, and the *mean number of clusters per speaker*.

	Clusters	
Distance metrics	Purity %	Per speaker
KL2	96.7	5.40
modified-BIC	97.8	4.88
modified-BIC + adjacency	97.8	3.15

Table 5. *Speaker cluster purity and mean number of clusters per speaker.*

Normally, a higher cluster purity is more desirable and less costly for subsequent processing than a smaller number of clusters per speaker. Looking at the results, we see that the adjacency term in the 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.

7. 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. Each model is composed of sentence clusters representing speech from the anchor in different background conditions. Normally a model does not have more than nine clusters. Each anchor model was built using sentences from 2 news shows of over 1 hour.

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. This cluster comparison uses the KL2 distance metrics to measure cluster similarity. If the KL2 value is lower than a specified threshold the cluster is tagged as an anchor cluster.

Table 6 shows the results obtained in a test set having 3 news shows with total duration over 2 hours. Results are given in terms of percentage of deletions, that is, clusters not identified as belonging to the anchor, and percentage of insertions, that is clusters incorrectly labelled as anchor. The results are very promising especially due to the very low insertion rate.

	Errors	
Anchor	Deletions %	Insertions %
Male 1	1.3	1.1
Male 2	9.3	1.2
Female	5.3	2.2

Table 6. *Results for anchor cluster detection.*

8. CONCLUSIONS

This paper reports our work on the development of an audio segmentation, classification, speaker clustering and anchor identification system applied to a Broadcast News task for the European Portuguese language.

We presented a new algorithm for audio segmentation that is both accurate and uses less computational resources than other approaches. Our speaker clustering module uses a modified BIC algorithm which performs substantially better than the standard KL2 and is much faster than the full BIC. Finally, we presented a scheme for tagging certain speaker clusters (anchors) using trained cluster models.

The results obtained using test sets with several news shows reveal the performance gains introduced by the new algorithms.

9. 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).

10. REFERENCES

[1] 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.

[2] 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.

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

[4] G. Williams and D. Ellis, "Speech/music discrimination based on posterior probability features," in *Proc. Eurospeech '99*, Budapest, Hungary, September 1999.