A MULTIMEDIA APPROACH FOR AUDIO SEGMENTATION IN TV BROADCAST NEWS

Luis Perez-Freire, Carmen Garcia-Mateo

ETSI Telecomunicación
University of Vigo - SPAIN
e-mail: {lpfreire, carmen}@gts.tsc.uvigo.es

ABSTRACT

This paper deals with the task of audio segmentation in TV broadcast news. A multimedia approach for this purpose, by means of audio and video processing, is proposed. Thus, the segmentation system is composed by two differentiated parts: one analyzes the audio stream, and is based on the well-known Bayesian Information Criterion (BIC), whereas the other part extracts useful information from the video stream to improve the performance of BIC. An investigation of parameters involved in BIC formulation is also accomplished, in order to achieve the best results possible in our experimental framework: the database Transcrigal-DB. The final system provides significative improvements in both overall performance and robustness.

1. INTRODUCTION

Nowadays, automatic speech recognition (ASR) attracts the interest of many researchers, mainly due to the potential applications it represents. Automatic transcription of broadcast news adds several difficulties to those involved in a conventional ASR framework: coexistence of speech and non-speech fragments (music, background noise, etc) and appearance of multiple speakers.

A common solution adopted to get around these problems and improve speech recognition performance consists of partitioning the audio stream into segments according to speaker identity and acoustic nature of the audio (speech, music...). Different approaches for automatic audio segmentation have been investigated in recent years. The most elementary ones perform partitioning at silence locations using a speech recognition front-end [1]. Another set of techniques is based on statistical pattern modeling using Gaussian Mixture Models (GMM's) [2]: segmentation is achieved by means of a maximum likelihood classification. Another family of segmentation systems uses Bayesian Information Criterion (BIC) [3] based on model selection, placing boundaries at locations where acoustic changes occur. The latter approach has demonstrated its effectiveness, so its popularity is fully justified.

This paper deals with audio segmentation applied in a concrete scenario of TV broadcast news. In multimedia recordings like these, we have two related information sources, namely audio and video streams. Video and audio events are often synchronized: acoustic changes are more likely to occur in the neighborhood of video shot boundaries. With this consideration in mind, a multimedia approach (audio + video processing) for audio segmentation is proposed in this paper. Such an approach is based primarily on BIC, and in addition takes into account useful information extracted from the video stream to improve performance. With regard to BIC, we concentrate our efforts mainly on achieving the best results possible in our experimental framework, by adjusting the parameters involved in BIC formulation. The final system provides significative improvements in overall performance.

The remainder of the paper is organized as follows: in section 2, a brief revision of BIC theory is expounded, emphasizing the main parameters involved in BIC formulation; section 3 provides a description of the elements of the segmentation system along with a thorough description of the proposed integration between audio and video information; in section 4, the experimental framework is presented and results regarding performance and improvements are detailed; finally, conclusions are presented in section 5.

2. REVISION OF BIC THEORY

BIC is a model selection criterion by means of a hypothesis test. Applied to the segmentation of audio streams, it can be explained as follows. Let $\chi = \{x_i \in \Re^d, \quad i = 1, \ldots, N\}$ be the sequence of feature vectors (e.g. Mel-cepstrum) extracted from an audio stream. We assume that χ can be modeled as a multivariate Gaussian process:

$$x_i \sim N(\mu_i, \Sigma_i)$$

To decide whether an acoustic change occurs at time i or not, the following hypothesis test is considered:

$$H_0 : x_1 \ldots x_N \sim N(\mu, \Sigma)$$
$$H_1 : \begin{cases} x_1 \ldots x_i \sim N(\mu_1, \Sigma_1) \\ x_{i+1} \ldots x_N \sim N(\mu_2, \Sigma_2) \end{cases}$$

As can readily be seen, the former hypothesis must be chosen when the analyzed stream contains no acoustic changes, whereas the latter must prevail when an acoustic change occurs at time i. Selection of the best fitting model is performed by computing the following statistic, according to BIC theory:

$$BIC(i) = R(i) - \lambda P \tag{1}$$

where

- $R(i)$ is a statistic from the maximum likelihood ratio

$$R(i) = N \log |\Sigma| - N_1 \log |\Sigma_1| - N_2 \log |\Sigma_2| \tag{2}$$

  with N, Σ being the number of vectors and covariance matrix respectively of the entire sequence, and $N_1, \Sigma_1, N_2, \Sigma_2$ the number of vectors and covariance matrices of partitions $\{x_1, \ldots, x_i\}$ and $\{x_{i+1}, \ldots, x_N\}$ respectively.
- P is the penalty, which corresponds to the number of free parameters of a multivariate Gaussian process in d dimensions:

$$P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N \tag{3}$$

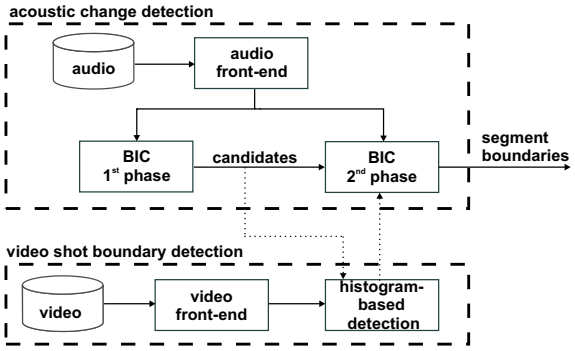


Fig. 1. Architecture showing all components of the segmentation system

- λ is the penalty weight, and can be viewed as an additional free penalty.

A positive value in (1) means that the model composed of two Gaussian processes best fits the data, thus an acoustic change at time i can be assumed. Detection of acoustic changes clearly depends on λ; as a matter of fact, performance of BIC-based systems is very sensitive to the selection of this parameter. Another parameter that requires special attention is N, i.e. the size of the analysis window, since reliability of Gaussian estimates depends directly on this value. Influence of both parameters, λ and N, will be analyzed in detail in section 4 in order to determine how they affect performance and to be able to select the most proper ones.

3. MULTIMEDIA APPROACH FOR AUDIO SEGMENTATION

The architecture of the segmentation system is depicted in figure 1. As can be seen, it performs two different tasks: detection of acoustic changes and detection of video shot boundaries. Each of these tasks is detailed in subsections 3.1 and 3.2. Section 3.3 explains how shot boundary detection can be taken into account to facilitate the detection of acoustic changes.

3.1. Acoustic change detection via BIC

The goal of the audio front-end in figure 1 is the parameterization of audio data: 13 dimension Mel-cepstrum vectors (including energy measure) are computed by segmenting the audio into 25 ms frames with overlapping of 15 ms. The other two elements are in charge of detecting acoustic changes by using a BIC-based approach.

Obviously, the scheme for acoustic change detection proposed in section 2 is applicable only to audio sequences containing at most one changing point, so it must be extended to the detection of multiple changes. Such an extension consists of applying BIC in two phases, in a manner similar to [4]:

1. In a first phase, BIC is evaluated in a sliding variable size window: BIC evaluation begins in a short window; if no acoustic change is found, window size is increased to include the subsequent audio fragment and BIC is re-evaluated. When an acoustic change is detected, the window size is re-set to its original value and located just after the detected point. This procedure is repeated until the end of the audio stream is reached. The aim of the current phase is to determine the approximate location of the acoustic changes, so a low resolution step (e.g. one second) when computing BIC values is adopted.
2. In a second phase, BIC is evaluated using a fixed size window (namely *refinement window*) centered in the candidates provided by the first phase. Resolution is increased (up to 0.1 seconds) in order to achieve precision. Furthermore, in this phase, some false detections generated by the first phase can be rejected by properly setting the window size, as will be demonstrated in the experimental results.

BIC value increases according to window size [3], and this may lead to false acoustic change detections using large analysis windows, so we consider that the penalty weight must show a dependence on the window size in order to achieve better robustness in the above algorithm. We propose the following correction for λ:

$$\lambda' = \lambda \left[\log_{10} \left(\frac{wsize}{ws_{ref}} \right) \right] \tag{4}$$

where $wsize$ is the window size in the first phase and ws_{ref} is the size of a reference window (both values given in seconds). The goal is to increase penalty weight smoothly (achieved by the log function) as long as the analysis window grows, so that false detection rate is reduced.

3.2. Shot boundary detection

Digital video can be viewed as a collection of static images (the so-called *frames*) played in sequence. Video sequences are organized in *shots*. A shot is a sequence of frames defining a continuous action. Shot transitions or shot boundaries may be detected by means of a dissimilarity measure between adjacent frames. A wide variety of methods for detection of shot boundaries exists [5]. Histogram-based methods provide a good trade-off between accuracy and amount of computation time, so they seem to be a proper choice for our system. The histogram-based detector works as follows:

1. First, the luminance histogram is extracted from every frame (video front-end in figure 1). Luminance of an image is computed by means of the following expression:

   $$L = 0.3R + 0.59G + 0.11B$$

   with R, G, B being the color components of the considered frame.
2. A distance is computed between histograms corresponding to every two adjacent frames:

   $$d(f, f') = \sum_{j=0}^{N} \frac{|H(f,j) - H(f',j)|}{H(f,j) + H(f',j)} \tag{5}$$

   where $H(f,j)$ is the luminance histogram for frame f and luminance level j, and N is the number of bins each histogram is divided into. In our case, N is equal to 256, i.e. the number of discrete luminance values a pixel can have. If (5) is above a certain threshold, a shot boundary is assumed, as is sketched in figure 2.

   The preceding equation is a normalized difference: it takes values in the range [0,1], which facilitates threshold selection. The closer the value of (5) to 0, the more similar the frames compared.

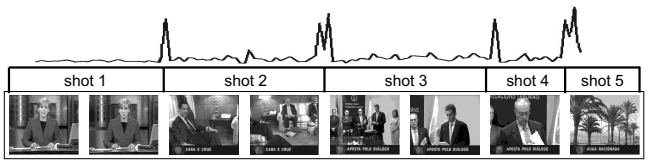


Fig. 2. Shot boundary detection

Transitions between shots may be either abrupt or gradual. In order to avoid difficulties derived from the existence of gradual transitions (which often yield to small values in (5)), we compare non-adjacent frames, i.e. histogram distance is computed every three or four frames, depending on the video frame-rate: the underlying idea is that by comparing frames more separated in time, an accentuation of differences between them is achieved. Moreover, in this way, a reduction in processing time is achieved.

3.3. Integration of acoustic change and shot boundary detection

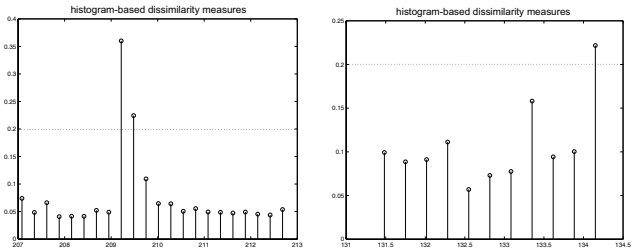
We want segmentation to be based mainly on BIC, so we decided to use the information provided by the shot boundary detector only to adjust the penalty weight λ dynamically in the BIC equation (1). As has been previously mentioned, acoustic changes are more likely to occur in the vicinity of video shot transitions. Information provided by the shot boundary detector may therefore be used as a cue to apply a selective penalty in BIC equation (1). The integration is accomplished as follows.

Shot boundary detection is evaluated in short windows (e.g. two seconds) centered in the candidates provided by BIC first phase. If shot boundary is detected, penalty weight (λ) in BIC second phase is reduced, otherwise its value remains unaffected. By reducing penalty weight, we are being more permissive in the admission of candidates as acoustic changes. Penalty weight is reduced by a factor depending on the reliability of the detected shot boundary. A certain point in a video stream is marked as a shot boundary when (5) is above a certain threshold but the points in its neighborhood did not reach that threshold. Thus, we define the reliability for a detected shot boundary as a function of the ratio between the value of the point identified as a boundary and the average of the rest of the points in its neighborhood (those ones contained within the limits of the analysis window):

$$R = 1 - \frac{V^{avg}_{noboundary}}{V_{boundary}} \tag{6}$$

where $V_{boundary}$ and $V^{avg}_{noboundary}$ are the value of the point at the boundary and the average of the rest of the points, respectively. R takes values in [0, 1] interval. In figure 3, two examples of shot boundary detection are shown, representing the dissimilarity measure between frames. In both cases, threshold is set to 0.2. Figure 3 (a) corresponds to a correctly detected boundary, whereas 3 (b) corresponds to a false detection, caused by camera panning. In the former case, the value obtained for R is high; on the contrary, the latter case provides poor reliability. Relying on the value of R, the new value for λ is computed as follows:

$$\lambda'' = \lambda \left[\log_{10} \left(\frac{10 \max(R, 0.4)}{0.4} \right) \right]^{-1} \tag{7}$$



(a) (b)

Fig. 3. Correct (a) and false (b) shot boundary detections

Thus, the closer to 1 the reliability factor R, the stronger the reduction of λ. The frontier between good and poor reliability has been empirically established at 0.4. On the other hand, function "max" avoid increments in λ when reliability falls below 0.4. Equation (7) is envisaged to alleviate pernicious effects of possible false shot boundary detections in system performance. In a histogram-based method, false detections may be caused by camera zooming, panning, tilting, etc.

4. EXPERIMENTAL RESULTS

4.1. Experimental framework

The database used for assessment of the proposed segmentation system is named *Transcrigal-DB* and consists of about 13 hours of audio and video material collected from TV broadcast news transmitted by "Televisión de Galicia" TV station, so audio is in Galician, which is a romance language similar to Spanish and Portuguese. The audio is sampled at a rate of 16 KHz and stored in WAV format, whereas video is captured at a rate of 15 frames/s and stored in AVI format. Prior to experiments, audio material is manually transcribed and labeled using the Transcriber [6] program.

4.2. Assessment of performance

Results are given in terms of false detection (FD) and false rejection (FR) rates:

$$\%FD = \frac{\#\ false\ detections}{\#\ total\ detections} \tag{8}$$

$$\%FR = \frac{\#\ missed\ detections}{\#\ total\ true\ changing\ points} \tag{9}$$

A first set of experiments was intended to analyze BIC-based segmentation without taking into account video information. First we compared performance provided by the system using the modified penalty proposed in equation (4) with performance achieved by the same system using a fixed value for λ. Experimental results confirmed that the use of the adaptive penalty yields to a reduction in FD rate, as was expected. Comparing FD rates achieved by both systems for a given FR rate, table 1 was obtained. The value for ws_{ref} in equation (4) was set to 4, since this is the initial window size (in seconds) in BIC first phase.

%FR	6	7	8	9	10	11	12
%FD reduction	2.4	2.8	2.2	3	2.8	2.8	2

Table 1. Reduction in FD rate as a function of FR rate

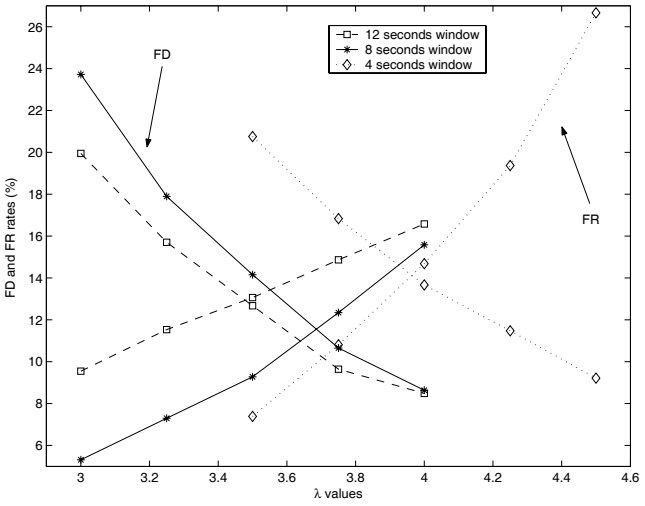


Fig. 4. Error rates as a function of refinement window size and λ

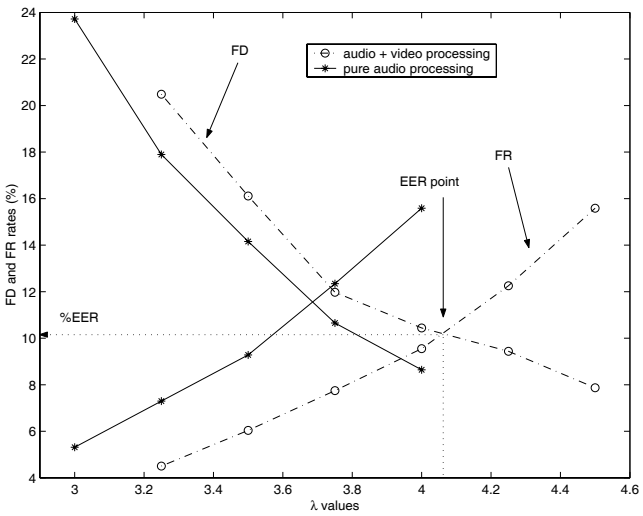


Fig. 5. Improvements achieved using video information (both curves were obtained using 8 seconds refinement windows)

The second aim of these experiments is to assess the influence of λ and refinement window size (used in BIC second phase) in performance. Figure 4 shows results as a function of these two parameters. Decreasing curves correspond to FD rate, whereas increasing curves correspond to FR rate. The point that leads to equal error rate in FD and FR for a given window size is the EER point and it is a representative measure of performance. It can be seen that λ may be adjusted to trade-off between FA and FR, but performance is highly dependent on the window size: with short-sized windows, Gaussian estimates are poor due to the lack of data so BIC decisions are often inaccurate. On the other hand, large windows increase reliability (thus allowing one to reject some false alarms generated in the first phase of the algorithm) but are more likely to include more than one acoustic change, clearly contradicting the hypothesis presented in 3.1. A compromise solution was achieved in the experimental results by choosing 8 seconds as window size. Figure 4 shows that for a given FR rate, 8 seconds window provides the lowest FD rates.

The second set of experiments is oriented to assess improvements in performance achieved by adding information about video shot boundaries. In figure 5, results using the combined segmentation system depicted in figure 1, are compared with the best results achieved without taking into account video information. As can be seen, inclusion of video information leads to a significative reduction of the overall error rates (for example, EER goes from 12% to 10%). Moreover, a flattening in both FD and FR curves can be noted, especially in the neighborhood of the EER point. Such a flattening yields to an improvement in terms of robustness in the selection of λ, since system performance is less sensitive to variations in this value.

5. CONCLUSIONS AND FURTHER WORK

An audio segmentation system that takes into account video information has been presented. Video shot boundary detection allows the use of a selective penalty in BIC evaluation; this way significant improvements were reached, in terms of performance and robustness in selection of λ parameter in BIC formulation. Furthermore, a qualitative study of window size-dependent BIC parameters has been accomplished, proposing an adaptive penalty weight and experimentally verifying the importance of a proper selection for the refinement window size. Regarding the multimedia segmentation system, the use of a more accurate video shot boundary detector could lead to even better results, so this could be a future track to explore.

6. ACKNOWLEDGEMENTS

This project has been partially supported by Spanish MCyT and Xunta de Galicia under the projects TIC2000-1104-C02-01 and PGIDT01PX132201PN respectively.

7. REFERENCES

[1] T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, S.J. Young, "Segment generation and clustering in the HTK Broadcast News Transcription System," *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, pp 133-137*, 1998.

[2] J.L Gauvain, L. Lamel, G. Adda, "Partitioning and transcription of broadcast news data," *Proc. ICSLP'98, Sydney, Australia, Vol 5, pp 1335-1338*, 2002.

[3] Scott Shaobing Chen, P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," *Proc. of the DARPA Broadcast News Workshop*, 1999.

[4] P. Sivakumaran, J. Fortuna, A.M. Ariyaeeinia, "On the use of the bayesian information criterion in multiple speaker detection," *Proc. EuroSpeech, Vol. 2, pp 795-798*, 2001.

[5] J.S. Borezcky, L.A. Rowe, "Comparison of video shot boundary detection techniques," *Proc. SPIE 2664, pp 170-179*, 1996.

[6] C. Barras, E. Geoffrois, Z. Wu, M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication, Vol 33, No 1-2*, 2000.