Audio data indexing : use of second-order statistics for speaker-based segmentation *

Perrine Delacourt and Christian Wellekens
Institut EURECOM, 2229 route des Crêtes,
BP 193, 06904 Sophia Antipolis Cedex, France
delacour@eurecom.fr

Abstract

The content-based indexing task considered in this paper consists in recognizing from their voice, speakers involved in a conversation. A new approach for speaker-based segmentation, which is the first necessary step for this indexing task, is described. Our study is done under the assumptions that no prior information on speakers is available, that the number of speakers is unknown and that people do not speak simultaneously. Audio data indexing is commonly divided in two parts : audio data is first segmented with respect to speakers utterances and then resulting segments associated with a given speaker are merged together. In this work, we focus on the first part and we propose a new segmentation method based on second order statistics. The practical significance of this study is illustrated by applying our new technique to real data to show its efficiency.

1 Introduction

The use of multimedia databases is now common. However, the question of data accessibility still remains. The retrieval of an audio document in a database should be easy and quick. Further, in the case where only a small portion of signal is of interest, it should be possible to access it directly without listening to the entire document. Therefore, content-based indexing systems are needed. Several indexing/retrieval keys can be considered : word, topic and so on. In this paper, the indexing/retrieval key we examine is the identity of the speakers present in the audio document. In other words, our speaker-based indexing task consists in the recognition of the sequence of speakers involved in the conversation (who speaks and when). In our context, we make no assumptions about the number of speakers, nor about prior knowledge of their characteristics (no model, no training phase). But we make the constraining assumption of only one person speaking at a given time.

In addition to applications to audio databases for easier and quicker retrieval, the speaker-based indexing task is also used in transcribing systems. Indeed, the recognition rate of transcribing systems is improved when the speech models are adapted to speakers. Therefore, our content-based indexing task can be used as a preprocessing step. The data corresponding to each speaker involved in the audio document are then used to adapt the speech models to each speaker. Finally, the speech recognition step can be initiated by using the speech models, which best describe the data and the speaker.

The process of our content-based indexing task is divided in two major parts. The audio data is first segmented in order to obtain speech segments containing one speaker only. Resulting segments related to the same speaker are then merged together by using hierarchical clustering (see e.g. [1]). These two parts can be considered separately and studied independently. The segmentation step can be used as a preliminary step for speaker tracking [2] and the hierarchical clustering may be used to group the messages belonging to a same speaker [3]. In this paper, we focus on the problem of speaker segmentation, which has received attention only recently in the literature. Segmentation algorithms based on a distance between two consecutive parts of the speech signal have been investigated in [4, 5, 6]. These segmentation algorithms suffer from a lack of stability since they rely on thresholding distance values. A segmentation algorithm based on the Bayesian Information Criterion (BIC) is presented in [7], but proves to require long speech segments. In this paper, we propose an algorithm which takes advantages of the two latter types of segmentation techniques. A first pass is operated, which essentially consists in a distance-based segmentation to detect the most likely speaker changing points. Several distances or similarity measures have been investigated. Then, a second pass validates or otherwise the previously detected changing points by using the Bayesian Information Criterion.

*This work was supported by the Centre National d'Etudes des Télécommunications (CNET), France under grant n^o 98 1 B.

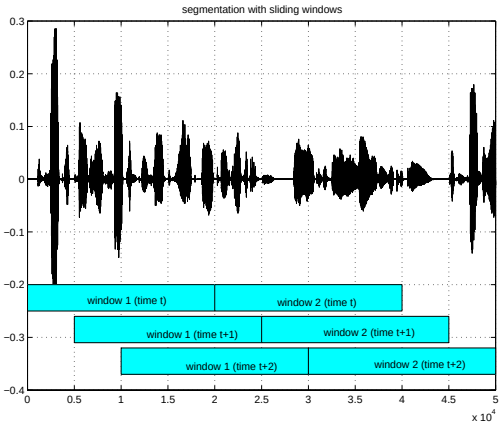

Figure 1. Principle of sliding windows

The paper is organized as follows. Section 2 presents the new segmentation algorithm based on second order statistics. Section 3 presents results of our segmentation technique when applied to real data. The quality of results are evaluated using criteria also presented in section 3. Section 4 details how the segmentation step presented here is embedded into the complete speaker-based indexing task. Finally, section 5 presents major conclusions and perspectives as to the use of such a technique in a speaker recognition process.

2 Speaker-based segmentation

The aim of segmentation is to obtain speech segments where only one speaker is talking. A distance-based segmentation is applied. First, the general principle of this procedure is explained: the distance measures we use and our technique for detecting speaker changing points are detailed. Due to a lack of stability in the detection of speaker changing points, a second pass using the Bayesian Information Criterion (BIC) is sometimes needed (2.2).

2.1 Distance-based segmentation

The principle behind speaker change detection is to measure a distance (or dissimilarity) value between two consecutive parts of the signal (called windows), assuming that each of these parts is related to one speaker only. A high value of this distance indicates a change of speakers at the common boundary of the two windows, whereas a low value signifies that the two portions of the signal in question correspond to the same speaker. The process is repeated along all the audio signal, as shown in figure 1 and, for each pair of windows, the distance value is stored.

Distance measures The distance measure should reflect how similar two segments of a signal are. A segment is a sequence of acoustic vectors $\mathcal{X} = \{x_1 ... x_{N_X}\}$ and is assumed to be related to one speaker only and to be generated by a multi-Gaussian process $\mathcal{X} \sim \mathcal{N}(\mu_X, \Sigma_X)$ where μ_X is the mean vector, Σ_X the covariance matrix and p the dimension of the acoustic vectors. In the following paragraphs, the measuring functions are referred as "distances" although they may not satisfy the three properties for being a distance function.

The *Generalized Likelihood Ratio (GLR)* was presented in [8, 9]. The following hypothesis test is considered :
• H_0 : both segments $\mathcal{X}$ and $\mathcal{Y}$ are generated by the same speaker $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y} \sim \mathcal{N}(\mu_Z, \Sigma_Z)$
• H_1 : segments $\mathcal{X}$ and $\mathcal{Y}$ are generated by different speakers $\mathcal{X} \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $\mathcal{Y} \sim \mathcal{N}(\mu_Y, \Sigma_Y)$
The corresponding likelihood ratio is defined by:

$$\lambda = \frac{L(\mathcal{Z}, \mathcal{N}(\mu_Z, \Sigma_Z))}{L(\mathcal{X}, \mathcal{N}(\mu_X, \Sigma_X)) L(\mathcal{Y}, \mathcal{N}(\mu_Y, \Sigma_Y))}$$

where $L(\mathcal{X}, \mathcal{N}(\mu_X, \Sigma_X))$ denotes the likelihood of the sequence of vectors $\mathcal{X}$ with respect to the Gaussian process $\mathcal{N}(\mu_X, \Sigma_X)$. To obtain a distance measure between segments, the negative of the logarithm is taken: $d_{\mathrm{GLR}} = -\log \lambda$.

The *Kullbach-Leibler distance* is defined as follows $\mathrm{KL}(\mathcal{X}, \mathcal{Y}) = \mathrm{E}_X \langle \log(P_X) - \log(P_Y) \rangle$ where $E_X < . >$ is the expectation operation performed with respect to the probability distribution function P_X of $\mathcal{X}$. The symmetrization given in [4] is $\mathrm{KL}_S(\mathcal{X}, \mathcal{Y}) = \mathrm{KL}(\mathcal{X}, \mathcal{Y}) + \mathrm{KL}(\mathcal{Y}, \mathcal{X})$.

All the *similarity measures* presented in this section are described in more details in [10]. Two segments $\mathcal{X}$ and $\mathcal{Y}$ of a parameterized signal should have similar covariance matrices, respectively Σ_X and Σ_Y, if they are generated by the same speaker. More formally, to measure how similar $\mathcal{X}$ and $\mathcal{Y}$ are, we consider the matrix $\Gamma = \Sigma_X \Sigma_Y^{-1}$. If both segments arise from the same speaker then $\Sigma_X = \Sigma_Y$, so that Γ is the identity matrix. The first similarity measure is defined as:

$$\mu_G(\mathcal{X}, \mathcal{Y}) = a - \log g + \frac{1}{p}(\mu_X - \mu_Y)\Sigma_X^{-1}(\mu_X - \mu_Y)^T - 1$$

where a is the arithmetic mean of the eigenvalues λ_i of Γ and g is the geometric mean. Clearly, if $\Sigma_X = \Sigma_Y$ (i.e. $\mathcal{X} = \mathcal{Y}$), then $\mu_G = 0$, otherwise $\mu_G > 0$. A second similarity measure is deduced from the previous one. It is based on the fact that mean vectors can be affected by the transmission channel and should not be taken into account for the second measure: $\mu_{GC}(\mathcal{X}, \mathcal{Y}) = a - \log g - 1$ The third similarity measure is a sphericity test for the matrix Γ: $\mu_{SC}(\mathcal{X}, \mathcal{Y}) = \log \frac{a}{g}$ The last similarity measure μ_{DC} is based on the absolute deviation of the eigen values of Γ compared to 1. All these similarity measures do not satisfy the symmetry property of a distance. Therefore, they are symmetrized as follows $\mu_S(\mathcal{X}, \mathcal{Y}) = \mu(\mathcal{X}, \mathcal{Y}) + \mu(\mathcal{Y}, \mathcal{X})$ where μ is either μ_G, μ_{GC}, μ_{SC} or μ_{DC}.

Detection of the speaker changing points All the above distance measures result in similar characteristics: a high value indicates a speaker changing point. By contrast, a

low value means that both segments are likely to be related to the same speaker. Therefore, once all distance values have been calculated, local maxima are searched for to detect speaker changing points. To detect the local maxima, we first filter the distance values graph with an averaging filter. The graph on the left in figure 2 shows the distance graph for the Generalized Likelihood Ratio (GLR) and the graph on the right illustrates the same GLR distance graph after filtering. A local maximum is considered as significant if the differences between its value and those of the minima surrounding it are above a certain threshold. For example, there is a non-significant local maximum inside the circle in the graph on the right of figure 2. Moreover, if two maxima are too close from one another, we keep only the one corresponding to the highest value, as shown in figure 3 where the last peak of the μ_{SC} distance graph presents two local maxima and the last one is preferred to the penultimate one.

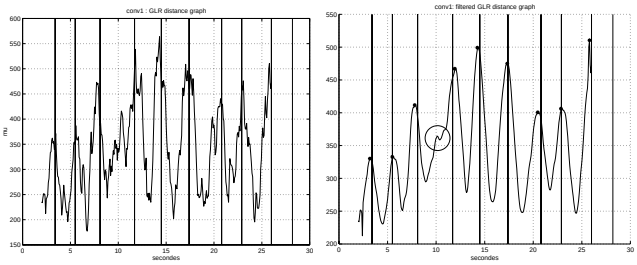

Figure 2. Detection of the speaker changing points : from left to right, GLR distance graphs before filtering and after filtering

Our detection method was designed to meet the following requirements:

• the same type of detection is required regardless of the distance measure used,

• we would rather have the utterances of a same speaker split in several segments than a segment containing utterances of several speakers,

thus making this procedure sub-optimal in general. Experience showed that by fine-tuning the parameters, an acceptable segmentation was obtained. However, the automation of this tuning for real data was not possible. In this context, we use a more formal approach. A second pass using the Bayesian Information Criterion (BIC) is considered to refine our results.

2.2 Refinement with BIC

We saw in the last section that our preliminary detection technique showed to be sub-optimal in general. It results in an over-segmentation and hence, short segments. A second pass is therefore introduced, which aims at refining this segmentation before performing hierarchical clustering. Consecutive segments will be merged if they satisfy the Bayesian Information Criterion (BIC) detailed in [7]. The BIC is a likelihood criterion penalized by the model complexity. Given $\mathcal{Z} = \{z_1, ..., z_{N_Z}\}$ a sequence of N_Z cepstral acoustic vectors and $L(\mathcal{Z}, M)$ the likelihood of $\mathcal{Z}$ for the model M. The BIC for the model M is given by: $\mathrm{BIC}(M) = \log L(\mathcal{Z}, M) - \lambda \frac{m}{2} \log N_Z$ where m is the number of parameters of the model M and λ is the penalization factor. We assumed that $\mathcal{Z}$ is generated by a multi-Gaussian process. We consider the following hypothesis test for a speaker change point at time i:

• H_0: $z_1 ... z_{N_Z} \sim \mathcal{N}(\mu_Z, \Sigma_Z)$

• H_1: $z_1 ... z_i \sim \mathcal{N}(\mu_{Z_1}, \Sigma_{Z_1})$ and $z_{i+1} ... z_{N_Z} \sim \mathcal{N}(\mu_{Z_2}, \Sigma_{Z_2})$

The maximum likelihood ratio between H_0 and H_1 is then:

$$R(i) = \frac{N_Z}{2} \log |\Sigma_Z| - \frac{N_{Z_1}}{2} \log |\Sigma_{Z_1}| - \frac{N_{Z_2}}{2} \log |\Sigma_{Z_2}| \quad (1)$$

where Σ_Z, Σ_{Z_1} and Σ_{Z_2} are respectively the covariance matrices of the complete data, of $\{z_1 ... z_i\}$ and of $\{z_{i+1} ... z_{N_Z}\}$ and N_Z, N_{Z_1} and N_{Z_2} are respectively the number of acoustic vectors of the complete data, of $\{z_1 ... z_i\}$ and of $\{z_{i+1} ... z_{N_Z}\}$. This hypothesis test can be seen as the comparison of two models : one models the data with two multi-Gaussian and the second with one multi-Gaussian only. The variations of the BIC value between the two modelizations is given by $\Delta\mathrm{BIC}(i) = -R(i) + \lambda P$, where the penalty is given by $P = \frac{1}{2}(p + \frac{1}{2}p(p+1)) \log N_Z$ and λ is the penalization factor. A negative value of $\Delta\mathrm{BIC}(i)$ indicates that the two multi-Gaussian modelization fits best the data $\mathcal{Z}$. In this case, a speaker changing point is detected at time i. In our application, the value of i represents the location of a changing point detected by the method presented in section 2.1. By this mean, we can differentiate between false and correct segmentation points.

3 Experimentations

A correct segmentation should provide the correct speaker changes and therefore segments containing one speaker only. We distinguish two types of errors related to speaker change detection. A *deletion error* occurs when the process does not detect an existing speaker change. The effect is that the resulting segment will contain two or more speakers and therefore will alter the validity of the hierarchical clustering. The *deletion error* rate is defined as the ratio of the number of missed detections versus the expected number of detections. An *insertion error* occurs when a speaker change is detected although it does not exist. The result is an over-segmentation : the utterances of a same speaker are split into several parts. Hierarchical clustering may not be affected if the length of the segments thus obtained remains above an acceptable value. The *insertion error* rate is defined as the ratio of the number of false detections versus the expected number of detections. For evaluating our technique, we will therefore give the rates of these

errors during the segmentation of our test data (see 3.2).

3.1 Data and parameterization

We use four conversations to test our approach. Two of them, referred to as `conv1` and `conv2`, were artificially created by concatenating sentences of about 2s from the TIMIT database. The other conversations are extracted from a database provided by Institut National de l'Audiovisuel. This database contains French TV news broadcasts and we chose two of them, referred to as `extrait1` and `extrait3`.

We use three sets of acoustics vectors : set I contains vectors with 12 lpcc coefficients (cepstral coefficients derivated from the linear predictive coding coefficients), set II contains vectors with mel-cepstral coefficients and set III contains acoustic vectors of dimension 2, the first coefficient is the energy of the signal and the second is the pitch. See [11] for more details.

3.2 Results and discussions

Evaluation of the sets of acoustic vectors The lpcc-coefficients and the mel-cepstral coefficients produce the best results for the distance-based segmentation (see table 1). However, the mel-cepstral coefficients give better results than lpcc-coefficients for synthetic data for the BIC pass (see table 2). The mel-cepstral should be preferred because the BIC values of a real and of a false speaker changing points computed with mel-cepstral coefficients show more contrast than those computed with lpcc coefficients. Thus, it is easier to fine-tune automatically the parameter λ to distinguish real from false changing points. The set III (signal energy and pitch) is not suitable to detect the speaker changing points. This probably comes from our cepstral coefficient pitch detection method, which is not rather reliable. We have also experimented using the same sets of acoustic vectors completed with the Δ-coefficients (first derivatives). The use of Δ-coefficients deteriorates the performances of both passes: the peaks of the distance graph are smoothed away thus making the detection of the speaker changing points more difficult and the BIC seems to be very sensitive to the dimension of the acoustic vectors.

	deletion error (I)	insertion error (I)	deletion error (II)	insertion error (II)
conv1	0	0	0	0
conv2	0	0	12.5%	0
extrait1	0	52.6%	0	52.6%
extrait3	5.9%	158%	5.9%	158%

Table 1. Errors with the GLR distance measure for set I and set II

Evaluation of the distance measures The distance graphs obtained with the different distance measures are shown in figure 3. The vertical lines indicate the localization of the real speaker changing points and the stars ($*$) represent the speaker changing points resulting from the distance-based segmentation. We first point out the ability of our segmentation technique to detect speaker changing points, even when they are close to one another. Although the Kullbach-Leibler (KL) distance and the Generalized Likelihood Ratio (GLR) are the most computationally costly, they produce the best results: one can distinguish easily the peaks corresponding to the speaker changing points. The μ_G measure seems to be a good compromise since its computational cost is lower than with the KL and the GLR distances and it provides almost similar results.

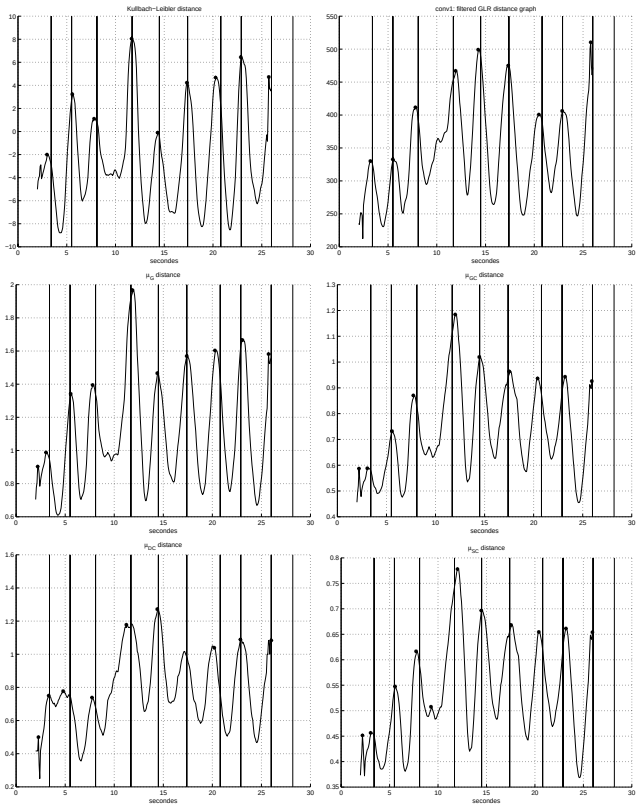

Figure 3. Distance-based segmentation with several distances

Discussion about parameter λ Concerning the parameter λ, its value should be equal to 1 in theory (see [7]). Besides, equation (1) implies that the number of changing points detected should decrease as the value of λ increases. In practice, we have better results by setting it empirically to 1,25. We can deduce from tables 1 and 2 that the second pass is not necessary for the synthetic data (`conv1` and `conv2`). Indeed, the BIC procedure tends to deteriorate the results from the distance-based segmentation. By contrast,

the evolution of deletion and insertion error rates between the first pass (table 1) and the second pass (table 2) clearly demonstrates that the second pass improves results from the first pass for the real data (`extrait1` and `extrait3`). It should be noticed that a substantial part of the insertion errors in `extrait1` and `extrait3` is due to significant pauses within single-speaker utterances.

	deletion error (I)	insertion error (I)	deletion error (II)	insertion error (II)
conv1	0	0	0	0
conv2	12.5%	0	25%	0
extrait1	5.3%	26.3%	5.3%	26.3%
extrait3	5.9%	117%	5.9%	117%

Table 2. Errors with BIC

4 The complete indexing task

Before concluding, we recall the complete procedure of our indexing task and explicit where the segmentation step takes place. As illustrated in figure 4, the audio signal is first parameterized (i.e. the acoustic vectors are computed on small analysis windows called frames). Then, the distance-based segmentation is operated, as detailed in section 2.1. This typically results in an over-segmentation : the utterances of a same speaker are split in several parts. A refinement is performed by applying the BIC: short consecutive segments related to a same speaker are merged. The final indexing results consist in the sequence of speaker utterances which constitutes the conversation. A hierarchical clustering step is therefore required to locate all the segments corresponding to a given speaker (see [7]).

5 Conclusion and future work

In this paper, we proposed a new segmentation algorithm, which combines advantages of metric-based and BIC-based techniques. A first pass locates accurately changing points even close one to another. The second pass improves the results by reducing the insertion error rate. Our experimentations show encouraging results. Our efforts will now concentrate on improving our technique of speaker changing points for distance-based segmentation and on automating the choice of parameter λ. At this point, it will be necessary to fully validate our approach on larger databases, such as HUB4 or SWITCHBOARD. The final step will be to combine our technique with hierarchical clustering to form a complete indexing procedure.

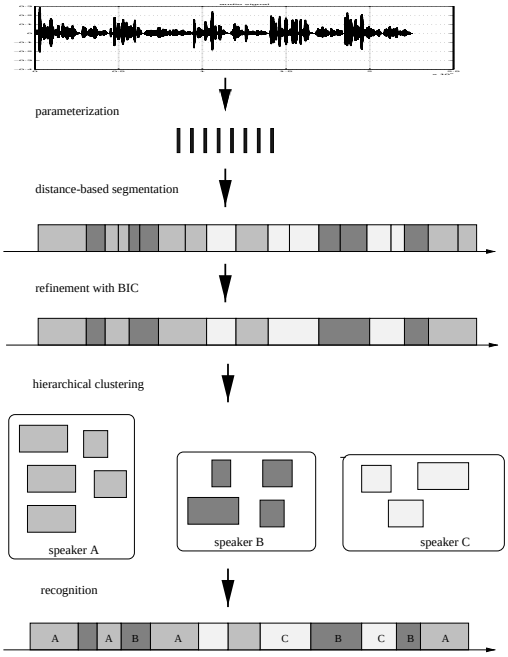

Figure 4. the complete indexing task

References

[1] R. Duda and P. Hart, *Pattern classification and scene analysis*. John Wiley and Sons, Inc., 1973.

[2] A. E. Rosenberg and al., "Speaker detection in broadcast speech databases," in *ICSLP98*, 1998.

[3] D. Reynolds and al., "Blind clustering of speech utterances based on speaker and language characteristics," in *ICSLP98*, 1998.

[4] M. A. Siegler and al., "Automatic segmentation, classification, and clustering of broadcast news audio," in *DARPA speech recognition workshop*, 1997.

[5] C. Montacié and M.-J. Caraty, "Sound channel video indexing," in *Eurospeech*, pp. 2359–2362, 1997.

[6] H. Beigi and S. Maes, "Speaker, channel and environment change detection," in *World congress of automation*, 1998.

[7] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA speech recognition workshop*, 1998.

[8] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *ICASSP*, pp. 873–876, 1991.

[9] H. Gish and N. Schmidt, "Text-independent speaker identification," *IEEE signal processing magazine*, oct. 1994.

[10] F. Bimbot and al., "Second order statistical measures for text-independant speaker identification," *Speech communication*, vol. 17, Aug 1995.

[11] L. Rabiner and R. Schafer, *Digital processinf of speech signals*. Prentince-Hall, 1978.