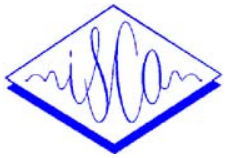

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DETECTION OF SPEAKER CHANGES IN AN AUDIO DOCUMENT

Perrine Delacourt David Kryze Christian J. Wellekens

Institut EURECOM, 2229 route des Crêtes,
BP 193, 06904 Sophia Antipolis Cedex, France
{perrine.delacourt,christian.wellekens}@eurecom.fr
http://www.eurecom.fr/~delacour/speech/

ABSTRACT

This paper addresses the problem of speaker-based segmentation. The aim is to segment the audio data with respect to the speakers. In our study, we assume that no prior information on speakers is available and that people do not speak simultaneously. Our segmentation technique is operated in two passes: first, the most likely speaker changes are detected and then, they are validated or discarded during the second pass. The practical significance of this study is illustrated by applying our technique to synthesized and real data to show its efficiency and to compare its performances with another segmentation technique.

Keywords: segmentation, likelihood ratio

1. INTRODUCTION

The speaker-based segmentation consists in obtaining speaker homogeneous segments: resulting segments should be as long as possible and related to a single speaker. This problem received attention recently in the literature [1, 2, 3, 4, 5] since it can be used as a preliminary step in several indexing applications: news transcription tasks [6, 7], automatic grouping speech messages[8] or speaker tracking [9, 10].

The segmentation algorithm we describe in this paper is designed to be embedded in a speaker-based indexing system. The goal of this system is to know who speaks and when. This indexing system is divided in two parts: first, the speaker-based segmentation step and then, the grouping step which aims at merging speech segments related to a given speaker as described for example in [8] or [11]. In this paper, we specialize in the first step. In our study, we assume that no prior information on speakers is available (no speaker or speech model, no training phase) and that people do not speak simultaneously.

Our segmentation algorithm is operated in two times. First, a distance-based segmentation combined with a thresholding process as robust as possible, is operated to detect the most likely speaker changes. Then, the Bayesian Information Criterion is used during a second pass to validate or discard the previously detected changing points. This criterion has been used for segmentation by S.Chen in [3], but proves to require long speech segments (>3s).

Section 2 details our segmentation technique. Performances of this segmentation algorithm are assessed in section 3 with criteria described in section 3.2. Results are commented in section 3.3 and comparison with the algorithm proposed by S.Chen ([3]) is also made. Finally, section 4 concludes and gives possible tracks for future work.

This work was supported by the Centre National d'Etudes des Télécommunications (CNET) under the grant n° 98 1B

2. SPEAKER-BASED SEGMENTATION

The aim is to segment the speech data every time a speaker change occurs. The speaker-based segmentation algorithm we propose is based on a two step analysis: a first pass uses a distance computation to determine the changing point candidates and a second pass uses the Bayesian Information Criterion (BIC) to validate or discard these candidates.

2.1. Distance-based segmentation

2.1.1. Detection of one speaker change

Given two adjacent portions of parameterized signal (sequences of acoustic vectors) $\mathcal{X}_1 = \{x_1, ..., x_i\}$ and $\mathcal{X}_2 = \{x_{I+1}, ..., x_{N_\mathcal{X}}\}$, we consider the following hypothesis test for a speaker change at time i:

- H_0: both portions are generated by the same speaker. Then the reunion of both portions is modeled by a multi-Gaussian process
$$\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \sim \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})$$
- H_1: each portion is pronounced by a different speaker. Then each portion is modeled by a multi-Gaussian process
$$\mathcal{X}_1 \sim \mathcal{N}(\mu_{\mathcal{X}_1}, \Sigma_{\mathcal{X}_1}) \text{ and } \mathcal{X}_2 \sim \mathcal{N}(\mu_{\mathcal{X}_2}, \Sigma_{\mathcal{X}_2})$$

The Generalized Likelihood Ratio GLR between the hypothesis H_0 and H_1 is defined by:

$$R = \frac{L(\mathcal{X}, \mathcal{N}(\mu_\mathcal{X}, \Sigma_\mathcal{X})}{L(\mathcal{X}_1, \mathcal{N}(\mu_{\mathcal{X}_1}, \Sigma_{\mathcal{X}_1}).L(\mathcal{X}_2, N(\mu_{\mathcal{X}_2}, \Sigma_{\mathcal{X}_2})}$$

The GLR has been used in [1, 12] for speaker identification and has proved its efficiency. The GLR distance is computed from the logarithm of the previous expression: $d_{\mathrm{GLR}} = -\log R$.

A high value of R (i.e. a low value of d_{GLR}) signifies that the one multi-Gaussian modeling (hypothesis H_0) fits best the data. By contrast, a low value of R (i.e. a high value of d_{GLR}) indicates that the hypothesis H_1 should be preferred so that a speaker change is detected at time i.

2.1.2. Detection of all speaker changes

The GLR distance is computed for a pair of adjacent portions (windows) of the same size (about 2s), and the windows are then shifted by a fixed step (about 0.1s) along the complete parameterized speech signal. The distance values computed for each pair of windows are stored to form at the end of the process a distance graph. Then, the significant peaks of this graph are detected since they correspond to the speaker changes. A local maximum is regarded as "significant" when the differences between its value and those of the minima surrounding it are above a certain threshold (calculated as a fraction of the graph variance), and when there is no greater local maximum in its vicinity. Thus, the selection of the local maxima is not done considering the absolute value of the peaks, but rather by considering the "form factor" of the peaks, as detailed in [13].

Since a missed detection (an actual speaker change has not been detected) is more severe for the grouping process than a false alarm (a speaker change has been detected although it does not exist), parameters involved in the speaker change detection have been tuned to avoid missed detection to the detriment of false alarms. Thus, the parameterized signal is likely over-segmented (utterances of a given speaker are split in several segments). A second pass using the Bayesian Information Criterion (BIC) is required to merge the segments corresponding to the same speaker, and thus to decrease the number of false alarms. The BIC applied to segmentation has been proposed by S.Chen in [3].

2.2. BIC refinement

The BIC (also called Minimum Description Length principle) is a likelihood criterion penalized by the model complexity. With the same notations as before, the BIC value is determined by: $\mathrm{BIC}(M) = \log L(\mathcal{X}, M) - \lambda \frac{m}{2} \log N_{\mathcal{X}}$, where $L(\mathcal{X}, M)$ is the likelihood of $\mathcal{X}$ for the model M, m is the number of parameters of the model M and λ the penalty factor. The first term reflects how good the model fits the data and the second term corresponds to the complexity of the model. Thus, the modeling which maximized this criterion is favored. The variations of the BIC value between the two models (one Gaussian function versus two different Gaussian functions) is then given by: $\Delta\mathrm{BIC}(i) = -R(i) + \lambda P$ where $R(i)$ denotes the maximum likelihood ratio between hypothesis H_0 (no speaker change) and H_1(speaker change at time i) and the penalty is given by $P = \frac{1}{2}(d + \frac{1}{2}d(d+1)) \log N_{\mathcal{X}}$, d being the dimension of the acoustic space, and λ is the penalty factor. A negative value of Δ-BIC(i) indicates that the two multi-Gaussian models best fit the data $\mathcal{X}$, which means that a change of speaker occurred at time i.

For each pair of segments delimited by the speaker changes previously found (during the first pass), a Δ-BIC value is computed. If the value is negative, the speaker changing point separating both segments is validated and then a Δ-BIC value is computed for the next pair of segments. If not, the speaker changing point separating both segments is discarded and then, both segments are merged to form one segment for the next pair of segments.

3. EXPERIMENTATIONS

3.1. Data

Different types of speech data have been used to compare our segmentation technique with the algorithm proposed by S.Chen, referred to as the BIC procedure (see [3, 13]):

- 2 conversations which are artificially created by concatenating sentences of 2 s on average from the TIMIT database (clean speech, short segments).
- 2 conversations created by concatenating sentences of 1 to 3 s from a French language database provided by CNET (Centre National d'Etudes des Télécommunnications) (clean speech, short segments).
- 3 TV news broadcasts extracted from the database optovided by INA (Institut National de l'Audiovisuel) in French language (segments of any length).
- 3 phone conversations extracted from SWITCHBOARD ([14]) database (segments of any length, spontaneous speech).

We also used 4 French TV news broadcasts (referred to as *jt*) collected in our lab to test more accurately our approach.

The speech signal is parameterized with 12 mel-cepstral coefficients. The addition of the Δ-coefficients (first derivatives) does not improve the results and increases the time of computation. For this reason, the Δ-coefficients are not used (see [15]).

3.2. Assessment methods

A good segmentation should provide the correct speaker changes and therefore segments containing a single speaker. We distinguish two types of errors related to speaker change detection. A *false alarm* (FA) occurs

when a speaker change is detected although it does not exist. A *missed detection* (MD) occurs when the process does not detect an existing speaker change. In our context, a missed detection is more severe than a false alarm. Indeed, a missed detection may damage the grouping step: a "corrupted" segment (containing two or more speakers) will contaminate the cluster it is attached to. By contrast, false alarms may be resolved during the grouping step: if the utterances of a given speaker have been split in several segments, then they will be grouped in the same cluster during the grouping step. We can then define the false alarm rate (FAR) where 'sc' denotes 'speaker changes':

$$\text{FAR} = \frac{\text{number of FA}}{\text{number of actual sc + number of FA}}$$

and the missed detection rate (MDR):

$$\text{MDR} = \frac{\text{number of MD}}{\text{number of actual sc}}$$

A good segmentation is then characterized by low values of FAR and MDR.

3.3. Results

In order to evaluate our segmentation technique, we compare it with the BIC procedure, described in [13]. For both techniques, we mention the false alarm rate (FAR) and the missed detection rate (MDR). Concerning our segmentation technique, we distinguish the distance-based segmentation (first pass) and the BIC refinement (second pass). Table 1 reports performances of the BIC procedure applied on different types of data described in 3.1 and table 2 reports performances of the two passes of our segmentation technique applied on the same data.

For both segmentation techniques, the parameters they involved are set up for each database. Their value depend on the length of the segments contained in the audio document. For instance, the longer the speaker segments are, the higher parameter λ (involved in the BIC) should be.

The MDR and FAR respectively with the BIC procedure (see table 1) and with the second pass of our segmentation algorithm (see table 2) applied on the TV broadcast news (INA) are quite equal. That means that both segmentation techniques are equivalent with conversations containing long speech segments. One can notice the significant reduction of the FAR between the first and the second pass of our algorithm. The distance-based segmentation seems to be more sensitive to environment changes or speaker intonation.

The phone conversations (referred to as SWITCHBOARD in tables 1 and 2) also contain long segment but of spontaneous speech. That means that the conversation is scaterred with small words like 'Yeah' or 'Hum-hum'. When these small words are uttered while the other person is speaking, our hypothesis that people do not speak simultaneously is not respected. The segmentation process is degraded by these small words since they are too small to be correctly detected. They are not considered as relevant for speaker-based indexing: a short intervention to say 'Yeah' has no significance in this context. Therefore, they are not taken into account for the assessment of both segmentation techniques. However, the distance-based segmentation, as seen above, is sensitive to environment changes. It detects one of the both boundaries of the small words. That explains the high value of the FAR with the first pass of our segmentation algorithm (see table 2). This value remains higher with the second pass than with the BIC procedure (table 1). On the other hand, the MDR of both segmentation techniques are comparable.

	BIC	
	FAR	MDR
TIMIT	31.5	30.5
CNET	14.3	50.0
INA	18.3	15.7
SWITCHBOARD	20.3	30.6

Table 1: FAR and MDR with the BIC procedure

Concerning the conversations containing short segments (referred to as TIMIT and CNET in tables), our segmentation technique shows better results than the BIC procedure: for these two types of conversations the MDR with our technique is twice lower than with the BIC procedure with comparable values of FAR. The CNET conversations are made of shorter segments than the TIMIT conversations: that explains the higher value of MDR and that shows also the limit of segment length for our segmentation technique. One can also notice that parameters are not influenced by the language: parameters of both segmentation techniques used with American and French synthetic conversations (TIMIT and CNET) are quite the same. The small differences are probably due to the recording conditions.

Our experiments show that our segmentation technique is more accurate than the BIC procedure in presence of short segments, although both techniques are equivalent when applied on conversations containing long segments.

We made other experiments with our segmentation technique applied on TV broadcast news collected in our lab in order to study the error occurrences. Results are reported in table 3. In order to assess more accurately our segmentation technique, we consider the shift rate (SR) defined as:

$$\text{SR} = \frac{\text{number of shifts}}{\text{number of actual sc}}$$

A shift denotes a speaker change which has been shifted by some tenths of second (it corresponds to a false

	1^{rst} pass		2^{nd} pass	
	FAR	MDR	FAR	MDR
TIMIT	40.3	14.3	28.2	15.6
CNET	18.2	16.7	16.9	21.4
INA	37.4	9.03	18.5	13.5
SWITCHBOARD	39.0	29.1	25.9	29.1

Table 2: FAR and MDR respectively with the first and the second pass of our segmentation

	1^{rst} pass			2^{nd} pass		
	FAR	MDR	SR	FAR	MDR	SR
jt	59.0	8.9	8.4	23.7	9.4	8. 4

Table 3: JT: FAR, MDR and SR respectively with the first and the second pass

alarm and a missed detection which are very close) and which may not affect the grouping step . Resulting from the speaker change shift, one of the segments contains data from to different speakers. But the proportion of data of one of the speakers (some tenths of second) compared to the proportion of data of the other speaker (several seconds) is insignificant.

Most of the missed detections are due to short sentences, especially during interviews. Question of journalists are in general very short during interviews. In fact, parameters have been set for long segments, so that short segments are not well detected. Two main reasons explain the high value of FAR. The first reasons is speech translations: foreigners are interviewed and their speeches are translated in parallel (once again, our hypothesis is not respected). The second reason for a high value of FAR is environment changes during reports. Most of the reports are built as follows: events dealt in the report are commented by a journalist but the soundtrack corresponding to the events is not completely removed. A change in the soundtrack corresponding to the events often causes a false alarm inside the journalist comment.

Finally, the assessment rates we use do not really reflect the quality of the resulting segmentation: although the MDR is not negligible, the most significant segments according to their length are detected and listening quality is worthwile.

4. CONCLUSION AND FURTHER WORK

In this paper, we proposed a segmentation technique, composed of a distance-based algorithm followed by a BIC-based algorithm. This segmentation technique proves to be as efficient as the BIC procedure in the case of conversations containing long segments and to give better results than the BIC procedure when applied to conversations containing short segments. Our experiments show that parameters depend especially on the length of speech segments contained in the conversation. A problem still remains: parameters can be tuned to detect rather short segments than long segments but not both lengths. Our efforts will now concentrate on adapting parameters to the actual length of speech segments. Finally, as this segmentation is a part of a speaker-based indexing system, our future work will consist in combining segmentation with the grouping stage to form the complete indexing process (i.e. the recognition of the sequence of speakers engaged in a conversation).

5. REFERENCES

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

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

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

[4] C. Montacié and M.-J. Caraty, "A silence noise music speech splitting algorithm," in *ICSLP*, 1998.

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

[6] P. Woodland and al., "The development of the 1996 HTK broadcast news transcription system," in *DARPA speech recognition workshop*, 1997.

[7] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *ICSLP*, 1998.

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

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

[10] I. Magrin-Chagnolleau and al., "Detection of target speakers in audio databases," in *ICASSP*, 1999.

[11] S. Johnson and P. Woodland, "Speaker clustering using direct maximisation of the MLLR-adapted likelihood," in *ICSLP*, 1998.

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

[13] P. Delacourt, D. Kryze, and C. J. Wellekens, "Speaker-based segmentation for audio data indexing," in *ESCA workshop: accessing information in audio data*, 1999.

[14] J. Godfrey and al., "SWITCHBOARD: telephone speech corpus for research and development," in *ICASSP*, 1992.

[15] P. Delacourt and C. J. Wellekens, "Audio data indexing: use of second-order statistics for speaker-based segmentation," in *ICMCS*, 1999.