ELSEVIER

Speech Communication 32 (2000) 111–126

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

DISTBIC: A speaker-based segmentation for audio data indexing ☆

P. Delacourt *, C.J. Wellekens

Institut Eurécom, 2229 route des Crêtes, 06904 Sophia Antipolis Cedex, France

Abstract

In this paper, we address the problem of speaker-based segmentation, which is the first necessary step for several indexing tasks. It aims to extract homogeneous segments containing the longest possible utterances produced by a single speaker. In our context, no assumption is made about prior knowledge of the speaker or speech signal characteristics (neither speaker model, nor speech model). However, we assume that people do not speak simultaneously and that we have no real-time constraints. We review existing techniques and propose a new segmentation method, which combines two different segmentation techniques. This method, called DISTBIC, is organized into two passes: first the most likely speaker turns are detected, and then they are validated or discarded. The advantage of our algorithm is its efficiency in detecting speaker turns even close to one another (i.e., separated by a few seconds).

Zusammenfassung

Dieser Artikel beschreibt Sprecher basierte Segmentierung, den ersten Schritt beim Indexieren von Sprechern. Das Ziel besteht darin, möglichst lange homogene Segmente zu extrahieren, die Laute eines einzelnen Sprechers enthalten. Wir legen zugrunde, daß keinerlei Sprachcharakteristik des Sprechers bekannt ist (weder Sprechermodel noch Sprachmodel). Außerdem wird die Annahme gemacht, daß immer nur ein Sprecher zur Zeit spricht und daß keine Echtzeitanforderungen vorhanden sind. Wir stellen existierende Segmentierungtechniken vor und schlagen eine neue Methode vor, welche zwei gebräuchliche Methoden kombiniert. Unsere Methode (DISTBIC) ist in zwei Phasen aufgeteilt: erst werden die wahrscheinlichsten Sprecherwechsel gefunden, die dann entweder validiert oder verworfen werden. Der Vorteil unseres Algorithmuses liegt in seiner Effizienz Sprecherwechsel aufzufinden, besonders wenn sie sehr nahe beieinander liegen (d.h. Abstände von wenigen Sekunden).

Résumé

Dans cet article, nous nous intéressons au problème de la segmentation en locuteurs, étape préliminaire nécessaire à plusieurs tâches d'indexation. Le but de la segmentation en locuteurs est d'extraire des segments homogènes ne contenant les paroles que d'un seul locuteur et aussi longs que possible. Dans notre contexte, nous faisons l'hypothèse qu'aucune connaissance a priori des locuteurs ou des caractéristiques du signal n'est à notre disposition (pas de modèle de locuteur, pas de modèle de parole). Nous supposons néanmoins que les personnes ne parlent pas simultanément et que nous n'avons pas de contrainte de temps réel. Nous présentons les techniques de segmentation existantes et nous

☆ The financial support of this project from the Centre National d'Etudes des Télécommunications (CNET) under the Grant No. 98 1B is gratefully acknowledged.

* Corresponding author.

0167-6393/00/$ - see front matter
PII: S0167-6393(00)00027-3

proposons une nouvelle méthode qui combine les avantages de deux techniques de segmentation. Cette nouvelle méthode de segmentation, appelée DISTBIC, s'opère en deux passes: les changements de locuteurs les plus probables sont tout d'abord détectés et ils sont ensuite validés ou annulés au cours de la deuxième passe. L'avantage de notre algorithme est son efficacité à détecter des changements de locuteurs proches les uns des autres (i.e. espacés de quelques secondes).

Keywords: Speaker turn detection; Generalized likelihood ratio; Bayesian information criterion

1. Introduction

With the ever increasing number of TV channels and broadcasting radio stations and thanks to the currently available huge storage means, many hours of TV and radio broadcasts are collected every year by information national heritage institutions, like the Institut National de l'Audiovisuel (INA) in France or the BBC archives in the UK. For example, INA possesses 45 years of TV archives consisting of 300,000 hours of national TV programs and 60 years of radio archives consisting of 400,000 hours of radio programs. Moreover, due to systematic digitalization of information, multimedia databases are on a rocketing increase.

Besides the storage and architecture problems underlying the design of such databases, another crucial problem is information retrieval: how to formulate a query in a convenient way and how to efficiently and quickly find the searched information whatever it could be: text, drawings, image, video, audio, music or speech. Pre-indexing is necessary to facilitate and speed-up any kind of query.

Clearly, access to audio documents is much more difficult than access to text; although text retrieval must cope with some variability of spelling by proposing different approximated solutions to the user, it is still easier to detect a name or a string of words in a text than to recognize a speaker or to spot a word within an audio recording, or to recognize a spoken sentence in a large lexicon. Also, listening to an audio recording takes much more time than reading a text. Consequently, it is essential to be able to directly access the significant segments rather than listening to the whole audio recording to retrieve pertinent information.

Audio document indexing associates with each audio document a file describing its structure in terms of retrieval keys. Phoneme strings can be keys for retrieval of a word or a sentence in a speech file (word- and sentence-spotting). Topic spotting plays an essential role in document filtering and understanding. Another key could be speaker identity. The presence of a given speaker in a conversation could be detected if this speaker's voice characteristics have been a priori enrolled. Automatic analysis of conversations recordings requires segmentation into segments containing only one speaker and segment clustering into one-speaker sets.

In this paper, we mainly address the specific problem of audio database segmentation with respect to speakers which is an essential initial step towards full indexing. To stay close to the application, no assumption is made about prior knowledge of the speaker or speech signal characteristics. However, we assume that people do not speak simultaneously. Additionally, since the construction of an index file is an off-line process, we have no real-time constraints. The problem of speaker-based segmentation and indexing is stated in Section 2. Possible application fields are described in Section 2.2. The hypotheses made for this work are discussed: they place our work in the perspective of approaches followed by other authors. Section 3 presents a brief description of the pioneering indexing tool of BBN for application in air traffic control. Section 4 deals with the segmentation operation which is central in this paper and makes a short review of different proposed techniques including the inspiring technique used by Chen and Gopalakrishnan (1998) at IBM. At this point, a new original technique DISTBIC is proposed using a two-pass approach. Since no prior knowledge about speakers is used, our solution turns out to be close to a general change detection algorithm. However, the application to sequences of feature

vectors extracted from the speech waveform and containing speaker information puts specific tunings forward and keeps the general principle in the background. Different criteria are presented as well as the complete algorithm for speaker turn detection. The role of computational effort required is not crucial while the completeness of the segmentation, whatever the speaker intervention lengths are, is essential. Improving this completeness is the aim of the proposed algorithm. The results of DISTBIC are reported in Section 6. We conclude and describe perspectives towards the complete realization of an indexing tool in Section 7.

2. The indexing problem

2.1. Description of the problem

Audio speaker indexing consists of the analysis of a speaker sequence. In other words, the question is to know who is speaking and when. Associating speech segments to the same speaker is as important as speaker identification: this information allows the understanding of the structure of a conversation between several persons. Most of the time, no a priori knowledge is available on the content of the recording: neither the number of different speakers nor their identities. As a consequence, no speaker models can be built in advance (except in specific applications like detection of a presenter who regularly appears in broadcasts, but this case is not addressed in this paper). TV/radio recordings also contain music: jingles but also frequently music superposed on a spoken part. Therefore, musical segments must be segregated from speech segments. Background music may seriously degrade the segmentation. Among all the problems, the major difficulty comes from overlapping speech: the problem has not yet been really studied and all publications on indexing (including this one) hypothesize no occurrence of such events.

A simple and somewhat naive idea is to use the silences between speaker utterances to segment a recording. This solution gives acceptable results only for recordings of cooperative speakers accepting a discipline in their elocution mode provided the level of noise is low.

2.2. Applications

Indexing could be used for example to create a database where all speech is indexed with respect to its author or as a preliminary step in transcription tasks (Gauvain et al., 1998; Woodland et al., 1997), in automatic grouping of speech messages (Reynolds et al., 1998) or in speaker tracking (Rosenberg et al., 1998).

An interesting application of speaker segmentation/indexing is in speaker adaptation. When an audio recording must be transcribed i.e. that the sequence of uttered words is recognized off-line and printed, a major problem is the error rate due to the use of generic speech unit models obtained from training on a huge multi-speaker database which is assumed to represent the world of speakers. Enhanced recognition rates are obtained if the recording is segmented in speaker homogeneous chunks or at least in speaker classes (the most trivial solution is to use male and female speaker models) and then speaker-adapted models can be used for enhanced recognition.

Politically correct behavior imposes on candidates that campaign to the Chamber of Representatives or for President, to use equal time for their public TV or radio addresses. The respective duration of the use of broadcasting media is checked manually (in France by the Conseil Supérieur de l'Audiovisuel): automatic segmenting of the debates could ease this task.

3. A pioneering application

The aim of the pioneering work at BBN (Gish et al., 1991) is to automatically retrieve instructions given to pilots among recorded dialogs between pilots and air traffic controllers to improve air traffic at Dallas-Fort Worth airport. Air traffic controllers may all use the same radio-channel so that several of them are engaged in the dialog. Segmentation and indexing constitute the first step of this study. Next steps are: reconstitution of a dialog between one pilot and one controller, flight

identification from the dialog, understanding the dialog. The underlying hypotheses were:

- the number of speakers is unknown,
- no a priori knowledge on the speaker is available,
- no real-time processing is required,
- use is made of differences between the channel qualities respectively used by pilots and controllers,
- pilots are considered as a single class and must not be segregated.

The segmentation applies to the Mel-cepstrum representation of the signal. A brute force approach compares 200 ms segments so that speaker turns are not actually detected. In a second step, segments are labelled into three classes by using the energy variation (MAD: median absolute deviation) which gives acceptable results even if automatic gain control is applied on the signal: noise, speech with high confidence or unreliable class. A mixture of two Gaussians (GMM) is then trained on high confidence segments and used in a new step to refine speech–silence segmentation. Then contiguous segments having the same label are merged, regardless whether they may correspond to different speakers. In a next step, segments are clustered into two classes (controllers and pilots) by using the maximum likelihood ratio as a criterion to decide whether two contiguous segments belong to the same class or not: even at the end of this process, several different controllers or speakers can be mixed in a same segment if not separated by noise segments. Then in arbitrary intervals (expected to contain only one controller and several pilots), segments are clustered into classes using the maximum likelihood ratio, and Gaussian mixtures of two terms are trained (pilots-controllers). Pilots are seen as a single class and controller segregation is obtained by using the Kullback–Leibler criterion to choose the different controller models built on the intervals. The whole process can be iterated to progressively refine the segmentation/indexing.

The BBN system addresses all the major problems encountered in a segmenting/indexing application. However, its goal is very specific since one may take advantage of the different qualities of the channels for segmenting. An access tool for multimedia databases cannot rely on the hypothesis that different speakers are using different telecommunication media. Further research is therefore necessary, in particular for the segmentation which cannot simply be based on checks at every 200 ms window.

4. Segmentation

Segmentation may use different features of the discourse:

- silence detection,
- speaker turn detection,
- frame identification requiring classes of models: speakers, contents (music, speech, noise, ...). This approach requires training material for building the models and cannot be used for general segmentation without a priori knowledge; it is useful in a second step to refine segmentation with models trained on clustered segments.

4.1. Silence detection

The principle of segmentation with respect to speakers based on silence detection relies on the assumption, not always verified, that utterances of different people are separated by significant silences.

To detect inter-speaker silences, Nishida and Ariki (1998, 1999) use the average power of the speech signal. If the power value is below a given threshold, then the signal is identified as silence. The authors do not give any details about how they choose the threshold. It may be tuned for each recording.

Montacié and Caraty (1998) use an energy histogram over 15 s. If it is Gaussian (tests over μ and σ), the interval is assumed homogeneous and can be labelled silence or non-silence. Otherwise, it is assumed bimodal, and means and standard deviations are derived by clustering. From this clustering, a threshold is computed from means and standard deviations and used in a 4-states automaton for decision.

In the previously described air traffic control problem (Gish et al., 1991), BBN proposed a

solution based on energy variability (MAD: median absolute deviation): energy varies more quickly in speech than in noise. However, this criterion left a large number of classifications undecidable.

4.2. Speaker turn detection

Segmentation based on speaker turn detection uses a different policy. The aim is to segment audio recordings into homogeneous segments containing one speaker only. It is mainly used in automatic transcription of news, together with speaker adaptation and speech recognition. The principle behind speaker turn detection is to measure a dissimilarity value between two consecutive parts of the parameterized signal (called windows), assuming that each of these parts is related to one speaker only or to noise (silence).

The objective of the CMU team (Siegler et al., 1997) is to segment an audio recording into classes such as read speech, spontaneous speech, telephonic speech, speech with musical background, speech with background noise, speech of non-native speakers. Mean and covariance of two adjacent windows under Gaussian hypotheses are estimated and the Kullback–Leibler distance between the two distributions is computed. A turn between the windows is assumed if the distance reaches a maximum when the two windows are slid along the time axis. The problem is how to detect maxima of the distance because a transition between speech and silence is not similar to a transition between two speakers.

The same principle is proposed by researchers at IBM (Beigi and Maes, 1998) for detecting speaker, channel, and environment changes. The data contained in two adjacent sliding windows is clustered into three classes by K-means algorithm: silence, speech and speaker-dependent speech. Using the relative distance between the clusters, a new "distance" is computed and acoustic changes correspond to maxima of this distance. Two drawbacks of this method are:

- a lack of robustness since the three clusters result from an unsupervised algorithm (K-means) which does not guarantee significance of the cluster centers,
- the implicit assumption that feature vectors corresponding to speech can be split in two classes: speaker-dependent feature vectors and common speech feature vectors.

Another approach by people at IBM based on Bayesian Information Criterion (BIC) inspired our contribution. The remaining part of this section is entirely devoted to this method.

4.2.1. The BIC procedure

Dissimilarity measurement between two adjacent windows is based on the comparison of their parametric statistical models. This comparison is performed using the BIC (Chen and Gopalakrishnan, 1998) (also known as Akaike or Rissanen criterion (Rissanen, 1989) or Minimum Description Length (MDL)).

4.2.1.1. Model selection criterion. The BIC criterion is a maximum likelihood criterion penalized by the model complexity, i.e., the number of model parameters. Let us denote the sequence of data to model $X = \{x_1, \ldots, x_{N_X}\}$ and M a parametric model. The likelihood $L(X, M)$ is maximized for this model. If m represents the number of parameters, the BIC criterion for M is defined as

$$\mathrm{BIC}(M) = \log L(X, M) - \lambda \frac{m}{2} \log N_X. \qquad (1)$$

The first term accounts for the quality of the match between the model and the data while the second one is a penalty for model complexity with λ allowing the tuning of the balance between the two terms (the theoretical value of λ is 1). In coding theory, the BIC expression (1) with λ equal to 1 represents the shortest code length with which long sequences of data can be encoded relative to a model M (see (Rissanen, 1989)).

The BIC permits the selection of a model out of a set of models for the same data: this model will match the data while keeping low complexity.

Besides, the BIC criterion can also be viewed as a general change detection algorithm since it does not take into account any prior knowledge on speakers.

4.2.1.2. Use of the BIC for one speaker turn detection. We assume that X is generated by a multi-dimensional Gaussian process and we consider the following hypothesis test for speaker turn at time i:

- H_0: $(x_1,\ldots,x_{N_X}) \sim N(\mu_X,\Sigma_X)$: the sequence has been uttered by a single speaker and thus is assumed to be represented by a single multi-dimensional Gaussian process,
- H_1: $(x_1,\ldots,x_i) \sim N(\mu_{X_1},\Sigma_{X_1})$ and $(x_{i+1},\ldots,x_{N_X}) \sim N(\mu_{X_2},\Sigma_{X_2})$: the sequence has been uttered by two different speakers. Two multi-dimensional Gaussian processes justify the production,

where μ and Σ, respectively the mean and full covariance matrix, are the parameters of the models.

The maximum likelihood ratio between hypothesis H_0 (no speaker turn) and H_1 (speaker turn at time i) is then defined by

$$R(i) = \frac{N_X}{2}\log|\Sigma_X| - \frac{N_{X_1}}{2}\log|\Sigma_{X_1}| - \frac{N_{X_2}}{2}\log|\Sigma_{X_2}|, \tag{2}$$

where Σ_X, Σ_{X_1} and Σ_{X_2} are respectively the covariance matrices of the complete sequence, of the subset $\{x_1,\ldots,x_i\}$, and of the subset $\{x_{i+1},\ldots,x_{N_X}\}$, and N_X, N_{X_1} and N_{X_2}, are respectively the number of acoustic vectors in the complete sequence, in the subset $\{x_1,\ldots,x_i\}$, and in the subset $\{x_{i+1},\ldots,x_{N_X}\}$. Then speaker turn point is estimated via maximum likelihood ratio criterion as

$$\widehat{t} = \arg\max_i R(i). \tag{3}$$

The variations of the BIC value between the two models (one Gaussian versus two different Gaussians) is then given by

$$\Delta\mathrm{BIC}(i) = -R(i) + \lambda P, \tag{4}$$

where the penalty is given by $P = \frac{1}{2}(p + \frac{1}{2}p(p+1)) \times \log N_X$, p being the dimension of the acoustic space. The symmetry of the covariance matrix is taken into account.

A negative value of $\Delta\mathrm{BIC}(i)$ indicates that the two multi-dimensional Gaussian models best fit the data X, which means that a speaker turn occurs at time i such that

$$\{\max_i \Delta\mathrm{BIC}(i)\} < 0. \tag{5}$$

According to Chen and Gopalakrishnan (1998), an advantage of the BIC procedure is to avoid the use of any threshold as in most of the previously described methods. Threshold estimation is a critical point in most of segmentation processes and is usually left to the user's inspiration. Actually, the role of λ is equivalent to the definition of a threshold.

4.2.1.3. Detection of multiple speaker turns with BIC. The computation of BIC values is efficiently implemented in three steps, as described in (Tritschler, 1998):

1. A first pass is performed to determine the approximate location of the turns. The ΔBIC value is computed between two adjacent windows $[a,b]$ and $[b,c]$, where the boundaries a and c are fixed, and where b takes its values in $[a,c]$ and is increased at each iteration by a certain resolution step. The distance $d(a,c)$ is increased when no negative value is found for ΔBIC. When a negative value is found, the turn becomes the new value for a.
2. The second pass uses the same method for refining the results of the first pass: the exploration intervals $[a,c]$ are chosen much smaller, and centered around the points previously selected as candidates.
3. The third pass validates the results of the second pass. If $\{s_1,\ldots,s_N\}$ is the set of speaker turn candidates found in step 2, a ΔBIC value is computed for each pair of windows $[s_{i-1},s_i]$ $[s_i,s_{i+1}]$. If the value is negative, a speaker turn is identified at time i. Otherwise, the point s_i is discarded from the candidate set, so that the ΔBIC value is now computed for the new pair of windows $[s_{i-1},s_{i+1}]$ $[s_{i+1},s_{i+2}]$ (with the old indexes), as shown in Fig. 1.

This method, which consists in merging segments as long as positive values for BIC are found, is necessary for a correct estimation of the Gaussian parameters, since the model accuracy highly depends on the amount of available information. Thus, the reliability of the results is a function of the length of the sequence of acoustic vectors used for computation.

A direct consequence is that the use of the BIC algorithm alone for the speaker segmentation is

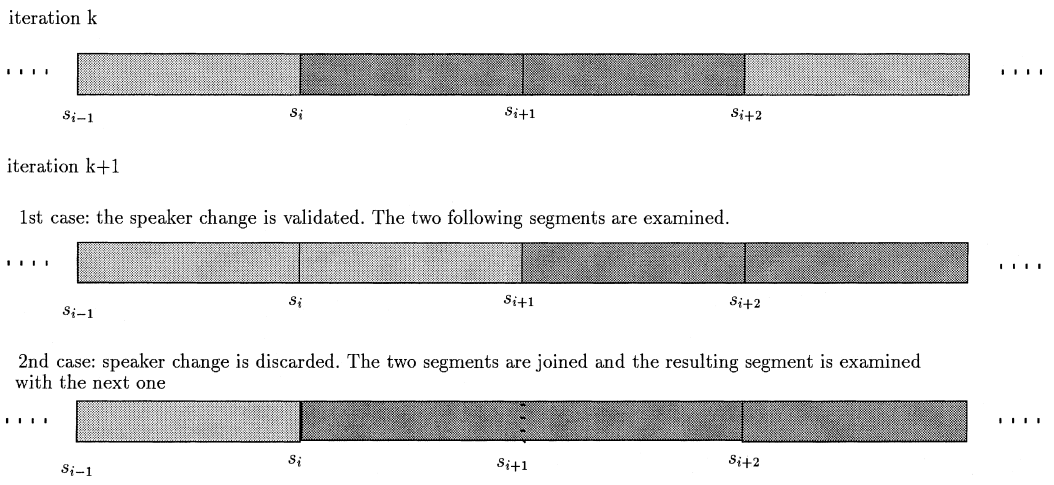

Fig. 1. Principle of the BIC final pass.

not adapted for small size segments. Indeed, the algorithm cannot detect two speaker turns closer to one another than the second pass window duration, which is of about 2 s.

Another problem comes from the tuning of the penalty factor λ, which showed to be dependent on the type of analyzed data. In Chen's work (Chen and Gopalakrishnan, 1998), λ is set to 1 but in our experiments the empirical factor λ took its values between 1.0 and 2.0 (see Section 6.3.2.1).

We will therefore use a more robust technique based on distance computation for the first pass that generates longer segments which can be used with the BIC algorithm for refinement in a second pass.

5. DISTBIC: a new two-pass segmentation technique

The method proposed in this paper is based on a two-step analysis: a first pass uses a distance computation to determine the turn candidates and a second pass uses the BIC (in fact, the third pass of BIC) to validate or discard these candidates. Our segmentation technique shows less dependence on the average segment size.

5.1. First pass: detection of speaker turn candidate points

The first pass of our segmentation technique relies on a distance-based segmentation defined from the likelihoods of adjacent windows. The concatenation of two windows is considered as a third one. In each window, the data is assumed to result from a single multi-dimensional Gaussian process, as shown in Fig. 2.

We are faced with the problem of deciding whether the data in the large window fits better with a single multi-dimensional Gaussian or whether a two-window representation justifies the data better. The length of the windows is the result of a trade-off between the number of frames inside the windows required for significant statistical estimation and the speaker homogeneity. A typical length of each window is 2 s. The windows are slid by steps of 100 ms for which the criterion is

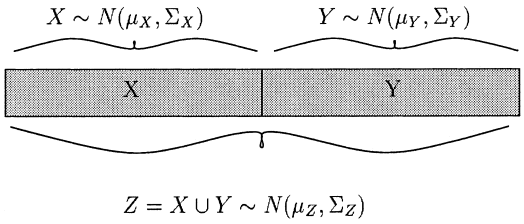

Fig. 2. Acoustic segment models.

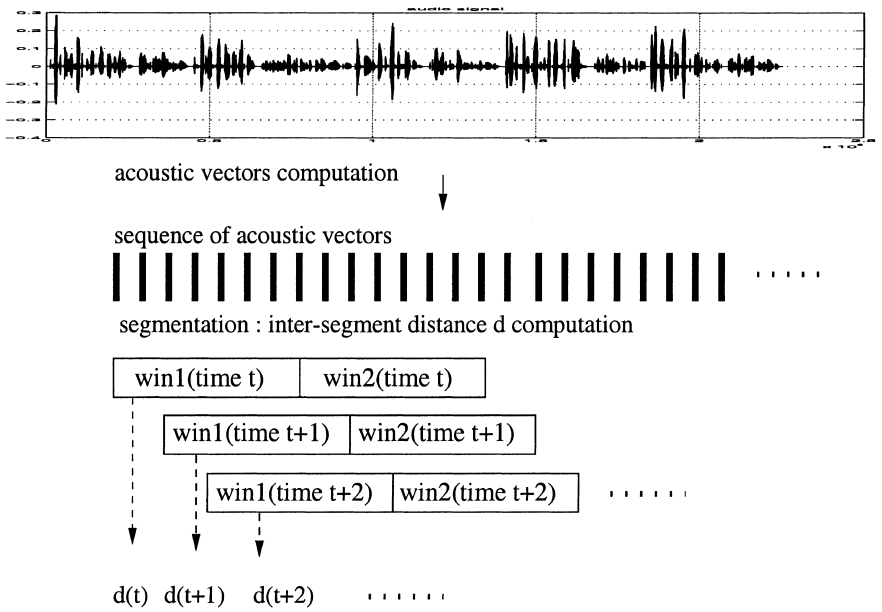

Fig. 3. Sliding windows.

computed (see Fig. 3). The resolution of speaker turns is thus 100 ms (see Section 6.3.2.1).

Six criteria will be tested but the segmentation procedure is similar in all cases: Generalized Likelihood ratio, 4 similarity measures based on covariances and the Kullback–Leibler criterion. We describe these criteria in the next section.

5.1.1. Criteria

5.1.1.1. The Generalized Likelihood Ratio. The Generalized Likelihood Ratio (GLR) is used by Gish and Schmidt (1994) and Gish et al. (1991) for speaker identification. The main features are outlined here. Let us consider the hypothesis test:

- H_0: segments have been uttered by the same speaker. Then, the union of the two segments is produced by a unique multi-dimensional Gaussian process.
- H_1: segments are uttered by several speakers. Then, they are assumed to be generated by different multi-dimensional Gaussian processes.

The likelihood ratio associated with this test is

$$R = \frac{L(z,\mu;\Sigma)}{L(x;\mu_1;\Sigma_1)L(y;\mu_2;\Sigma_2)}, \tag{6}$$

where $L(X, N(\mu_X, \Sigma_X))$ represents the likelihood of the sequence of acoustic vectors X given the multi-dimensional Gaussian process $N(\mu_X, \Sigma_X)$. To obtain a distance between two segments, the log-value of this ratio is considered:

$$d_R = -\log R. \tag{7}$$

5.1.1.2. The Kullback–Leibler distance. The Kullback–Leibler distance (KL) (or divergence) measures the distance between two distributions,

$$\mathrm{KL}(X,Y) = E_X\langle(\log P(X) - \log P(Y))\rangle, \tag{8}$$

where $E_X\langle\cdot\rangle$ denotes the expectation computed with the density P of X (see (Siegler et al., 1997)).

A symmetrical measure is obtained as

$$\mathrm{KL2}(X,Y) = \mathrm{KL}(X,Y) + \mathrm{KL}(Y,X). \tag{9}$$

For Gaussian variables X and Y, KL2 can be written

$$\mathrm{KL2}(X,Y) = \frac{\sigma_X^2}{\sigma_Y^2} + \frac{\sigma_Y^2}{\sigma_X^2} + (\mu_X - \mu_Y)^2\left(\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2}\right) - 1, \tag{10}$$

and becomes for Gaussian vectors

$$\mathrm{KL2}(X,Y) = \frac{1}{2}(\mu_Y - \mu_X)^{\mathrm{T}}(\Sigma_X^{-1} + \Sigma_Y^{-1})(\mu_Y - \mu_X) + \frac{1}{2}\mathrm{tr}\left((\Sigma_X^{1/2}\Sigma_Y^{-1/2})(\Sigma_X^{1/2}\Sigma_Y^{-1/2})^{\mathrm{T}}\right) + \frac{1}{2}\mathrm{tr}\left((\Sigma_X^{-1/2}\Sigma_Y^{1/2})(\Sigma_X^{-1/2}\Sigma_Y^{1/2})^{\mathrm{T}}\right) - p, \tag{11}$$

where tr denotes the trace of a matrix and with the same notation as before, p is the dimension of the feature vectors.

5.1.1.3. Similarity measures. All the *similarity measures* presented in this section are described in more details in (Bimbot et al., 1995). They rely on the hypothesis that two segments X and 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 two speaker segments X and 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

$$\mu_G(X,Y) = a - \log g + \frac{1}{p}(\mu_X - \mu_Y)\Sigma_X^{-1}(\mu_X - \mu_Y)^{\mathrm{T}} - 1, \tag{12}$$

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., $X = Y$), 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,

$$\mu_{GC}(X,Y) = a - \log g - 1. \tag{13}$$

The third similarity measure is a sphericity test for the matrix Γ

$$\mu_{SC}(X,Y) = \log\frac{a}{g}. \tag{14}$$

The last similarity measure μ_{DC} is based on the absolute deviation of the eigenvalues of Γ when compared to 1,

$$\mu_{DC}(X,Y) = \frac{1}{p}\sum_{i=1}^{p}|\lambda_i - 1|. \tag{15}$$

All these similarity measures do not satisfy the symmetry property of a distance. Therefore, they are made symmetrical as follows:

$$\mu_S(X,Y) = \mu(X,Y) + \mu(Y,X), \tag{16}$$

where μ represents μ_G, μ_{GC}, μ_{SC} or respectively μ_{DC}.

For all distance measures described in this section, a high value indicates a turn of speaker, whereas low values signify that the two portions of signal correspond to the same speaker.

5.1.2. Speaker turn detection

The criterion (“distance”) is computed for a pair of adjacent windows of the same size (about 2 s), and the windows are then shifted by a fixed step (about 0.1 s) along the whole parameterized speech signal. This process (see Fig. 3) gives the graph of distance as output with respect to time which is smoothed by a low-pass filtering operation. Then, all the “significant” local maxima are searched. 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 higher 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. To be more formal, if σ and μ respectively denote the standard deviation and the mean of the distances along the plot, a peak is significant if

$$|d(\max) - d(\min_{\mathrm{r}})| > \alpha\sigma$$

and

$$|d(\max) - d(\min_{\mathrm{l}})| > \alpha\sigma, \tag{17}$$

where α is real, and $\min_{\mathrm{r}}$ and $\min_{\mathrm{l}}$ are respectively the right and left minima around the peak max. This is illustrated in Fig. 4.

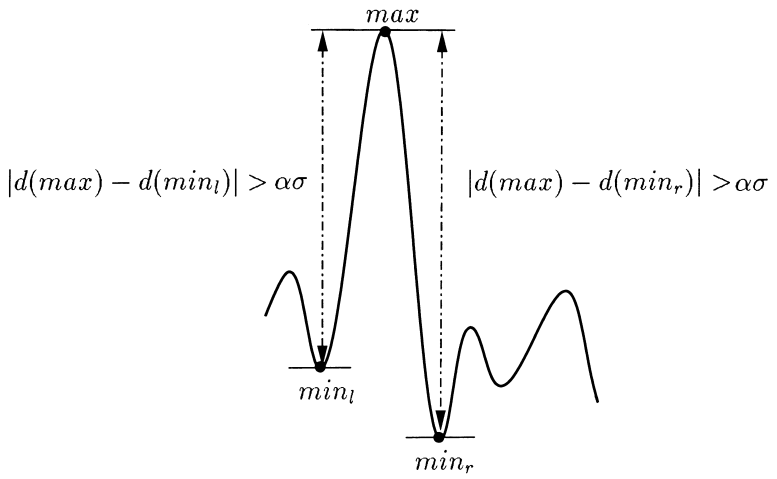

Fig. 4. Distance plot: characterization of a speaker turn.

In addition, we impose some minimal duration between two maxima: if two maxima are too close, the lowest one is discarded. Clearly, this constraint gives the lowest bound for short duration segments (typically, 1 s). This type of detection meets the following requirements:

- It does not depend on the type of speech data (TV news, phone conversations, studio).
- In this step, the emphasis is placed on minimizing missed detections (not detecting an actual turn) resulting in a high number of false alarms (detecting speaker turns although they do not exist). This number will be reduced by merging contiguous segments during the second pass by means of the BIC criterion.

5.2. Second pass: BIC refinement

The second pass is the exact copy of the third pass of the BIC analysis presented in Section 4.2.1 (see Fig. 1). A ΔBIC value is computed for each turn candidate to validate the result of the first pass. The value of the empirical factor λ has to be tuned in order to reduce the number of false alarms without increasing the number of new missed detections. The use of the BIC is now much more appropriate as the length of the considered segments is large enough for a good parameter estimation.

6. Experiments and results

In order to fully evaluate the DISTBIC segmentation technique, we first perform several tests on the possible configurations that form this technique. For example, the most accurate distance measure is first determined by these pre-tests. Once this optimal DISTBIC procedure is constituted, we compare it with the BIC procedure in Section 6.3.2. Finally, a more thorough analysis of DISTBIC results is conducted on TV news in Section 6.3.3.

6.1. Data and parameterization

Different types of speech data have been used to compare both segmentation techniques BIC and DISTBIC:

- 2 conversations artificially created by concatenating sentences of 2 s on an average from the TIMIT database (clean speech, short segments, 60 speaker turns).
- 2 conversations created by concatenating sentences of 1–3 s from a French language database provided by Centre National d'Etudes des Télécommunnications (CNET) (clean speech, short segments, 45 speaker turns).
- 3 TV news broadcasts extracted from the database provided by Institut National de l'Audiovisuel (INA) in French language (segments of

any length, prepared and spontaneous speech, 85 speaker turns, 30 minutes).
- 3 phone conversations extracted from SWITCHBOARD database (Godfrey et al., 1992) (segments of any length, spontaneous speech, 120 speaker turns, 30 minutes).

Concerning the artificially created conversations, the inter-speaker silences have been reduced so that they sound like those of a real conversation. More precisely, resulting inter-speaker silences are intentionally too short to be detected by our algorithm. Each segment of a speaker is followed by a segment of another speaker.

For additional tests on DISTBIC (Section 6.3.3), we used French TV news broadcasts to assess our algorithm and to analyze the nature of the observed errors:
- 4 French TV news broadcasts collected in our lab, referred to as *jt* (segments of any length, spontaneous and prepared speech, 830 speaker turns, 135 minutes).

We used 12 Mel-cepstral coefficients since they have proved in most papers to be very efficient for speaker recognition. They are computed with a 32 ms analysis window, shifted by 10 ms (the sample frequency of our audio signals is of 8 kHz). We also experimented using the same set of feature 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 turns more difficult. In addition, the BIC is sensitive to the dimension of the feature vectors.

6.2. Assessment methods

A good segmentation should provide the correct speaker turns and therefore segments should contain a single speaker. We distinguish two types of errors related to speaker turn detection. A *false alarm* (FA) occurs when a speaker turn is detected although it does not exist. A *missed detection* (MD) occurs when the process does not detect an existing speaker turn. 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 are likely to be grouped in the same cluster during the grouping step. We can then define the false alarm rate (FAR),

$$\mathrm{FAR} = 100 \times \frac{\text{number of FA}}{\text{number of actual speaker turns} + \text{number of FA}}\%. \tag{18}$$

A high value of FAR signifies that the speech signal has been over-segmented. The missed detection rate (MDR) is defined by

$$\mathrm{MDR} = 100 \times \frac{\text{number of MD}}{\text{number of actual speaker turns}}\%. \tag{19}$$

A high value of MDR means under-segmentation.

Since a missed detection is more severe than a false alarm, as seen above, our system is tuned to get low values of FAR and MDR but with MDR < FAR, which is not the traditional EER (equal error rate) objective.

A reference segmentation is required for using this kind of error definitions. However, since the human ear detects speaker turns with a limited accuracy, this reference segmentation (when it exists) should manage some tolerance. This is due for instance to breaths or sighs before utterances. It results that if speaker-based hand-segmentation is performed by several people on real conversations, it may result in different references. But for synthetic signals, speaker turns are obviously known by construction. One can account for this tolerance by defining accuracy windows around reference and detected speaker turns. A detected speaker turn is a false alarm if no reference speaker turn is found in the surrounding window. On the contrary, the absence of a speaker turn candidate in a window around a reference speaker turn corresponds to a missed detection (see also (Liu and Kubala, 1999)). However, real performance evaluation should use speaking rate-dependent accuracy windows and might not be independent of the semantic context of the

conversation. So that the ultimate test is performed by listening to the segments in isolation and deciding upon the quality of their ending point detection.

6.3. Results

6.3.1. Choice of the distance measure

Fig. 5 shows the distance graphs obtained with the different distance measures detailed in Section 5.1.1. We use a speech file from TIMIT to produce these graphs. The vertical lines indicate the localization of the real speaker turns and the stars ($*$) represent the speaker turns resulting from the distance-based segmentation. We first point out the ability of our segmentation technique to detect speaker turns, even when they are close to one another. Although the KL distance and the GLR are the most computationally costly, they produce the best results: one can distinguish easily the peaks corresponding to the speaker turns. The μ_G measure may be a good compromise since its computational cost is lower than with the KL and the GLR distances and it provides similar results.

With audio files containing spontaneous speech, we recommend the use of the measure derived from the GLR, as it is done in the following, since it proves to be the most efficient one, showing high and narrow peaks at speaker turn, and low amplitude variation within single speaker segments.

6.3.2. Comparison of the BIC and DISTBIC segmentation techniques

In order to evaluate our segmentation technique, we compare it with the BIC procedure, described in Section 4.2.1. For both techniques, we mention the FAR and the MDR. For the DISTBIC technique, we distinguish the distance-based segmentation (first pass) and the BIC refinement (second pass).

6.3.2.1. Parameters. Since both segmentation techniques are based on local signal properties only, it is not surprising that the tuning of parameters will play an essential role in the segmentation by allowing over-segmentation to justify slight intra-speaker local variations. The ultimate solution will take trained speaker models into account and will be implemented in a later stage. By now, parameters are adjusted to meet our constraints of low FAR and MDR with MDR $<$ FAR.

Parameters have been tuned on the test sets. Table 1 gives the parameter values for the BIC algorithm:

- λ is the penalty weight for the BIC criterion (see Eq. (1)),
- *win*1 is the duration (in seconds) of the window $d(a,c)$ and *res*1 is the resolution step described in step 1, Section 4.2.1.3,
- likewise, *win*2 is the duration (in seconds) of the window $d(a,c)$ and *res*2 is the resolution step described in step 2, Section 4.2.1.3.

Table 2 reports the parameter values for the DISTBIC algorithm:

- λ is the penalty weight for the BIC criterion (see Eq. (1)),
- α is the coefficient defined in Eq. (17) (see also Fig. 4),
- *win* is the duration (in seconds) of a window and *shift* is the window shift between two iterations (see Section 5.1 and Fig. 3).

Concerning window duration (*win*, *win*1 and *win*2), as seen in Section 5.1, it results from a trade-off between:

- a short duration to assume that windows contain utterances of a single speaker,
- a long duration to have a good estimation of speaker models,

*res*1, *res*2 and *shift* give the accuracy of the speaker turn location.

For high values of λ, more Δ-BICs are positive since $P > 0$ (see Eq. (4)), so that fewer speaker turns will be detected.

It is quite easy to justify (Eq. (17)) that large values of α also reduce the number of detected speaker turns.

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 similar. This is also true for real conversations (see Tables 1 and 2). The small differences are probably due to the recording conditions.

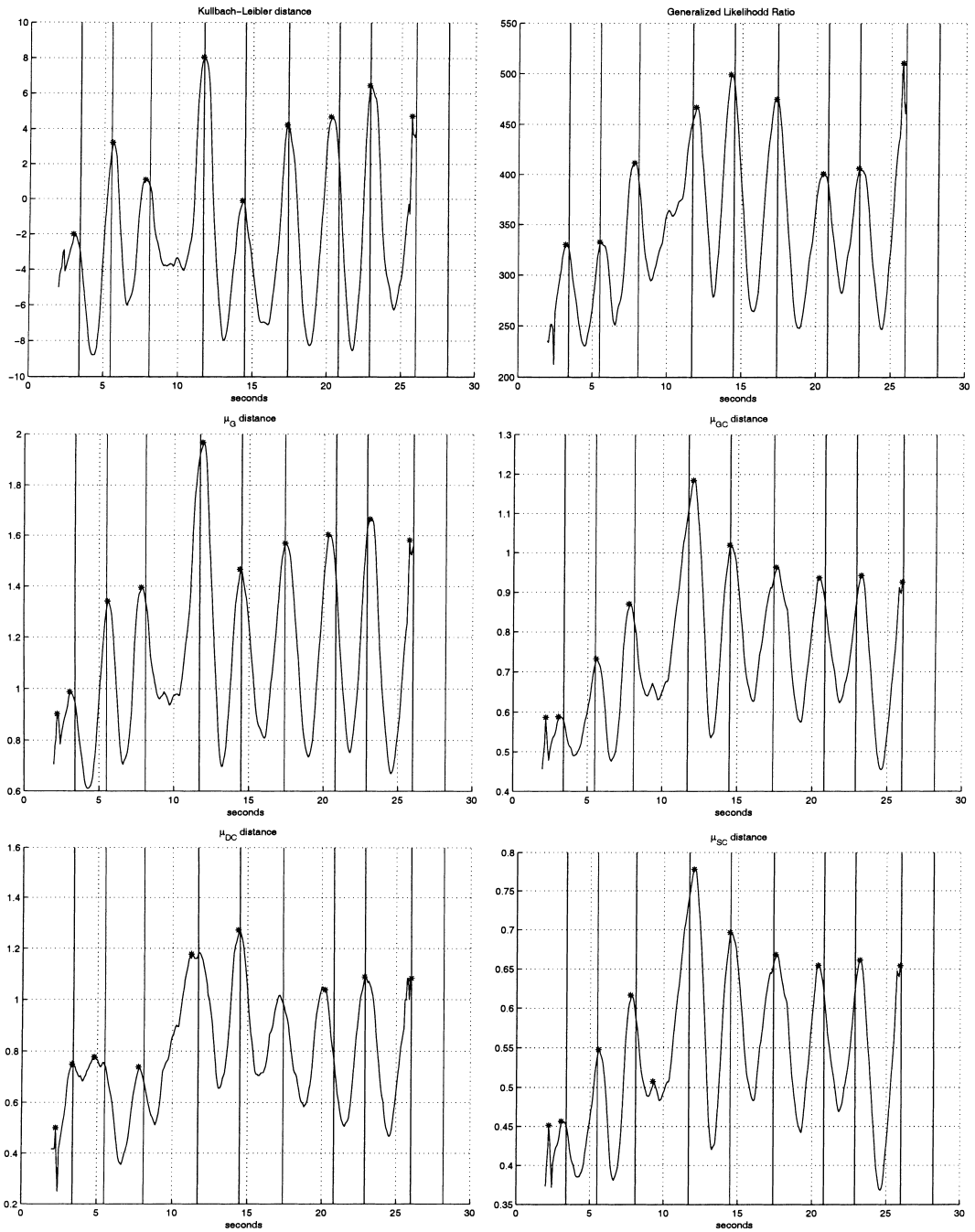

Fig. 5. Distance-based segmentation with several distances. From left to right and from top to bottom: the KL, the GLR, the μ_G, the μ_{GC}, the μ_{DC} and the μ_{SC} criteria.

6.3.2.2. Computational cost. Concerning the computational cost, our algorithm (as well as the BIC algorithm) uses two passes so that a real-time processing is not possible. However, the computing time required to find speaker turns only corresponds to a few percents of the recording

Table 1
Parameter values of the BIC procedure for the different speech data

BIC	λ	*win*1	*res*1	*win*2	*res*2
TIMIT	1.3	3 s	0.6 s	2 s	0.2 s
CNET	1.3	3 s	0.6 s	2 s	0.2 s
INA	2.0	3 s	0.6 s	1.8 s	0.15 s
SWITCHBOARD	2.0	3 s	0.6 s	1.8 s	0.15 s

Table 2
Parameter values of the DISTBIC method for the different speech data

DISTBIC	λ	*win*	*shift*	α
TIMIT	1.2	1.96 s	0.7 s	15 %
CNET	1.0	1.96 s	0.7 s	15 %
INA	1.8	2 s	0.1 s	50 %
SWITCHBOARD	1.5	2 s	0.1 s	50%

playback duration (a few minutes for a 45-minute audio document). BIC is faster than DISTBIC but this difference is irrelevant for the considered applications (speaker-based indexing).

$$\text{BIC} < \text{DISTBIC} \ll \text{recording playback}. \tag{20}$$

6.3.2.3. Performance comparison. Table 3 reports performances of the BIC procedure applied on different types of data described in 6.1 and Table 4 reports performances of the two passes of the DISTBIC segmentation technique applied on the same data. The MDR respectively with the BIC procedure (15.7%) and with the second pass of our segmentation algorithm (13.5%) applied on the TV broadcast news (INA) are almost equal. Similar results are observed on the FAR: respectively 18.3% with the BIC procedure and 18.5% with our algorithm. 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: from 37.4% to 18.5%. The distance-based segmentation seems to be more sensitive to environment changes or speaker intonations related to the semantic content.

Table 3
FAR and MDR with the BIC procedure

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

Table 4
FAR and MDR respectively with the first and the second pass of the DISTBIC method

	First pass		Second 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
SWITCH-BOARD	39.0	29.1	25.9	29.1

Telephone conversations (referred to as SWITCHBOARD in Tables 3 and 4) also contain long segments but of made of spontaneous speech. That means that the conversation is scattered 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 detected correctly. Also depending on the context of the segmentation, they may not be relevant. On the contrary, if the accuracy level required for a transcription task is very high, then it becomes necessary to detect these small words correctly. In our context, we decide not to take them into account.

The distance-based segmentation, as seen above, is sensitive to environment changes. It detects one of both boundaries of the small words. That explains the high value of the FAR with the first pass of our segmentation algorithm: 39.0%. This value remains higher with the second pass (25.9%) than with the BIC procedure (20.3%). On the contrary, the MDR of both segmentation techniques are comparable: 29.1% with DISTBIC and 30.6% with BIC.

Concerning the conversations containing short segments (referred to as TIMIT and CNET in the tables), the DISTBIC method shows better results than the BIC procedure: for these two types of conversations the MDR with our technique is half

(15.6% for TIMIT and 21.4% for CNET) that with the BIC procedure (30.5% for TIMIT and 50.0% for CNET) with comparable values of FAR (28.2% for TIMIT and 16.9% for CNET with DISTBIC versus 31.5% for TIMIT and 14.3% for CNET with BIC). The CNET conversations are made of shorter segments than the TIMIT conversations: that explains the higher value of MDR and also shows the limit of segment length for our segmentation technique.

The difference of performance between synthetic (TIMIT and CNET) and real conversations (SWITCHBOARD and INA) for both segmentation algorithms is explained by the difference of the actual length of speaker segments in the conversations. In the ideal case, during the last pass(es) of both segmentation algorithms, speaker models (the multi-dimensional Gaussians) are estimated on the actual length. Thus, the longer the actual speaker segments are, the better (more reliable and more robust) the speaker model estimation is and also, the better the resulting segmentation will be.

The difference of performance between the two types of conversations may also be explained by the recording conditions. Indeed, when speakers use different communication channels, turn detection is made easier by the discrimination enhanced by the channels characteristics, like in the BBN experiments. On the contrary, in synthetic conversations (like TIMIT and CNET), turn detection only relies on differences between speakers. In other words, for real conversations, segmentation algorithms detect changes of speaker together with recording conditions and for our synthetic conversations, algorithms only detect speaker changes.

Our experiments show that the DISTBIC segmentation technique is more accurate than the BIC procedure in the presence of short segments, although both techniques are equivalent when applied to conversations containing long segments, except for SWITCHBOARD which shows a higher FAR.

6.3.3. Qualitative analysis of the error occurrences

We conducted further experiments with the DISTBIC technique applied to TV broadcast news *jt* collected in our lab in order to study the error occurrences. The parameters of the segmentation were not retuned. We used parameters tuned in earlier tests on the INA database because data are of the same nature. Results are reported in Table 5. In order to assess our segmentation technique more accurately, we consider the shift rate (SR) defined as

Table 5
jt: FAR, MDR and SR respectively with the first and the second pass of the DISTBIC method

	First pass			Second pass		
	FAR	MDR	SR	FAR	MDR	SR
jt	59.0	8.9	8.4	23.7	9.4	8.4

$$\mathrm{SR} = 100 \times \frac{\text{number of shifts}}{\text{number of actual speaker turns}}\%. \tag{21}$$

A shift denotes a speaker turn which has been shifted by less than 1 s (it corresponds to a false alarm and a missed detection which are very close). As a consequence of the speaker turn shift, one of the segments contains few data from the contiguous speaker. But it will not affect the grouping step provided the ratio between odd data to the current segment is low.

Most of the missed detections are due to short sentences, especially during interviews. Questions of journalists are in general very short. In fact, parameters have been set for long segments, so that short segments are poorly detected. Two main reasons explain the high value of the FAR. The first reason 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 with in the report are commented by a journalist but the sound track corresponding to the events is not completely removed. A change in the sound track corresponding to the events often causes a false alarm within the journalist comment.

7. Conclusion and further work

We proposed a segmentation technique composed of a distance-based algorithm followed by a BIC-based algorithm. This segmentation

technique proved to be as accurate 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 showed that parameters mainly depend on the length of speech segments contained in the conversation. A problem still remains: parameters can be tuned to detect short segments rather than long segments but not both lengths simultaneously. For that reason, preference is given to over-segmentation. Indeed, this segmentation algorithm is one of the parts of a speaker-based indexing system and the next step will consist in grouping similar segments to form the complete indexing process (i.e., the recognition of the sequence of speakers engaged in a conversation). For other applications like speaker tracking, over-segmentation can be dealt with as explained in (Bonastre et al., 2000).

Acknowledgements

The authors would like to thank S. Marchand-Maillet for his help and are grateful to the anonymous reviewers for helpful comments on an earlier version of this paper.

References

Beigi, H.S.M., Maes, S., 1998. Speaker, channel and environment change detection. In: World Congress of Automation.

Bimbot, F., Magrin-Chagnolleau, I., Mathan, L., 1995. Second-order statistical measures for text-independent speaker identification. Speech Communication 17 (1–2), 177–192.

Bonastre, J.-F., Delacourt, P., Fredouille, C., Merlin, T., Wellekens, C.J., 2000. A speaker tracking system based on speaker turn detection for NIST evaluation. In: IEEE International Conference on Acoustics Speech and Signal Processing. To be published.

Chen, S.S., Gopalakrishnan, P.S., 1998. Speaker environment and channel change detection and clustering via the Bayesian Information Criterion. In: DARPA Speech Recognition Workshop.

Gauvain, J.-L., Lamel, L., Adda, G., 1998. Partitioning and transcription of broadcast news data. International Conference on Spoken Language Processing 4, 1335–1338.

Gish, H., Schmidt, N., 1994. Text-independent speaker identification. In: IEEE Signal Processing Magazine, October, 18–32.

Gish, H., Siu, M.-H., Rohlicek, R., 1991. Segregation of speakers for speech recognition and speaker identification. In: IEEE International Conference on Acoustics Speech and Signal Processing. pp. 873–876.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: IEEE International Conference on Acoustics Speech and Signal Processing. Vol. I, pp. 517–520.

Liu, D., Kubala, F., 1999. Fast speaker change detection for broadcast news transcription and indexing. In: Eurospeech. Vol. 3, pp. 1031–1034.

Montacié, C., Caraty, M.-J., 1998. A silence/noise/music/speech splitting algorithm. In: International Conference on Spoken Language Processing. Vol. 4, pp. 1579–1582.

Nishida, M., Ariki, Y., 1998. Real time speaker indexing based on subspace method: applications to TV news articles and debate. In: International Conference on Spoken Language Processing. Vol. 4, pp. 1347–1350.

Nishida, M., Ariki, Y., 1999. Speaker indexing for news articles debates and drama in broadcasted TV programs. In: IEEE International Conference on Multimedia Computing and Systems. pp. 466–471.

Reynolds, D.A., Singer, E., Carlson, B.A., O'Leary, G.C., McLaughlin, J.J., Zissman, M.A., 1998. Blind clustering of speech utterances based on speaker and language characteristics. In: International Conference on Spoken Language Processing. Vol. 7, pp. 3193–3196.

Rissanen, J., 1989. Stochastic Complexity in Statistical Inquiry. Series in Computer Science, Vol. 15. World Scientific, Singapore, Chapter 3.

Rosenberg, A.E., Magrin-Chagnolleau, I., Parthasarathy, S., Huang, Q., 1998. Speaker detection in broadcast speech databases. In: International Conference on Spoken Language Processing. Vol. 4, pp. 1339–1342.

Siegler, M.A., Jain, U., Raj, B., Stern, R.M., 1997. Automatic segmentation classification and clustering of broadcast news audio. In: DARPA Speech Recognition Workshop. pp. 97–99.

Tritschler, A., 1998. A segmentation-enabled speech recognition application using the BIC criterion. Master's thesis. Institut EURECOM, France.

Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J., 1997. The Development of the 1996 HTK broadcast news transcription system. In: DARPA Speech Recognition Workshop. pp. 97–99.