Hybrid Approach for Unsupervised Audio Speaker Segmentation

Hachem Kadri, Zied Lachiri, Noureddine Ellouze

14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

HYBRID APPROACH FOR UNSUPERVISED AUDIO SPEAKER SEGMENTATION

Hachem KADRI[1] , *Zied LACHIRI*[1,2] , *and Noureddine ELLOUZE*[1]

[1] Unité de recherche signal, image et reconnaissance de formes, ENIT
BP 37, Campus Universitaire, 1002 le Belvédère, Tunis Tunisia
Emails: kadri_hachem@enit.rnu.tn, N.ellouze@enit.rnu.tn
[2] Département Physique et Instrumentation, INSAT, BP 676, Centre Urbain Cedex, 1080, Tunis Tunisia
Email: zied.lachiri@enit.rnu.tn

ABSTRACT

This paper deals with a new technique, DIS_T^2_BIC, for audio speaker segmentation when no prior knowledge of speakers is assumed. This technique is based on a hybrid concept which is organized in two steps: the detection of the most probable speaker turns and the validation of turns already detected. For the detection our new technique uses a new distance measure algorithm based on the Hotelling's T^2-Statistic criterion. The validation is obtained by applying the Bayesian Information Criterion (BIC) segmentation algorithm to the detected speaker turns. For measuring the performance we compare the segmentation results of the proposed method versus recent hybrid techniques. Results show that DIS_T^2_BIC method has the advantage of high accuracy speaker change detection with a low computation cost.

1. INTRODUCTION

Audio Speaker Segmentation (ASS) is considered to be a process that attempts to find speaker segment boundaries in a given audio stream. This partition of audio data into homogeneous regions is of interest to broad class of speech processing applications, including speech recognition, audio transcription and audio indexing. There are several ways in which Audio speaker segmentation could prove to be useful for these applications. It can be an essential pre-processing task for speech and speaker recognition systems, and a necessary step for the automatic indexing of multimedia data. In fact, the Automatic speech recognition performance, which can be more reliable with the application of model adaptation and noise reduction, can be improved using segments relatively homogeneous and short. In audio indexing, speaker segmentation gives a structural summary of the speaker turns and effective cues for boundaries between audio sources, or scenes in multimedia applications.

Most previous unsupervised Audio Speaker Segmentation algorithms can be divided into two classes Metric-based [9] and Model-Selection-based [4]. Metric-based methods are not stable and need thresholds generally selected from experiments results while Model-selection-based methods suffer from the complexity of the model and the high computation cost. A widely used technique for speaker segmentation is based on the Bayesian Information Criterion (BIC) [5]. Indeed, BIC segmentation presents the advantages of robustness and threshold independence. However, this method, extremely computationally expensive, can introduce an estimation error due to insufficient data when the speaker turns are close to each other (about 2 seconds).

In order to minimize these effects, Delacourt [7] tested different metric criteria to associate them to the BIC criterion such as the Kullbach-Leibler distance, the similarity measure and the Generalized Likelihood Ratio measure (GLRM). Still, this method encountered many problems in case of short segments and requires a high computation cost. On another issue, Zhou [11] recommends the use of the T^2-Statistic for metric-based segmentation in the aim to reduce this computation cost. However its technique, T^2-BIC, depends on many empiric parameters which affect the quality of the detection of speaker turns. Therefore, we developed a new technique DIS_T^2_BIC which combines Metric-based and Model-Selection-based segmentation techniques with a complementary manner allowing the improvement of the short segments segmentation. This technique gives better results than T^2-BIC and lower computation cost than DISTBIC [7].

Prior to implementing our algorithm, we considerate the following hypothesis: the number of speakers is unknown, no prior knowledge on the speaker is available, and speakers don't speak simultaneously.

This paper is organized as follows: section 2 illustrates the principle of the hybrid approach for audio speaker segmentation. In section 3, the best well known hybrid audio speaker segmentation techniques and our new technique are introduced. Section 4 discusses our experimental results and section 5 concludes the paper with a summary and discussion.

2. THE HYBRID APPROACH

The hybrid approach consists on the association of Metric-based and model-selection-based segmentation techniques with a complementary manner. First, we use a distance measure algorithm and we adjust its parameters in order to detect the most possible speaker changes. At the end of this step the audio stream is over-segmented; so that we get a high false alarm rate but a low missed detection one. The wrong detection points will be eliminated on the second phase using a model selection criterion to validate speaker turns detected by the metric-based first phase (figure 1).

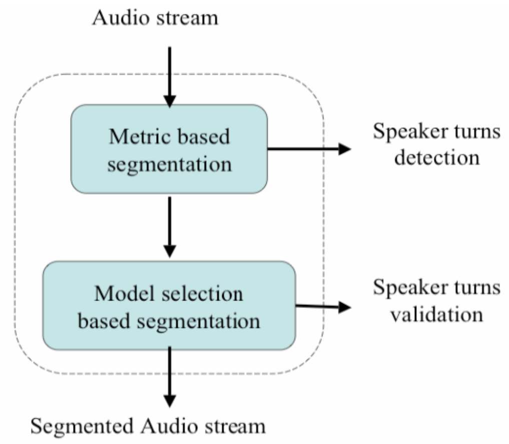

Figure 1- The hybrid based audio speaker segmentation concept

2.1 Speaker turns detection

Speaker turns are detected using a metric-based audio segmentation.In this method, various acoustic distance measures have been defined to evaluate the similarity between two adjacent windows shifted along the audio stream to form a distance curve. This distance curve was often low-pass filtered and the locations of peaks were chosen to be acoustic changing points by a generally selected from experiments results is needed determined threshold. Most of the distance measure criterions come from the statistical modelling framework. The feature vectors in each of the two adjacent windows are assumed to follow some probability density (usually Gaussian) and the distance is represented by the dissimilarity of these two densities. The Kullbak-Leibler distance and the Generalized Likelihood ratio are the most used methods for speaker segmentation [7]; they present the advantage of precise detection and low computation cost. However, they have many drawbacks, it is difficult to decide an appropriate threshold and each acoustic changing point is detected only by its neighbouring acoustic information.

2.2 Speaker turns validation

A model selection based segmentation techniques permit the validation of speaker turns. This technique uses a statistical decision criterion to detect speaker turns [4] by the means of a sliding window through the audio stream. Assuming that data are generated by a Gaussian process, speaker changes are detected by comparing two hypotheses:

- the window contains data generated by the same distribution;
- the left and right semi-windows, with respect to the speaker change, contain data drawn by two different distributions.

The Bayesian Information Criterion (BIC) is a widely used model selection criteria for speaker segmentation which permits to choose one model among a set of models for the same data. Lets $X = \{x_1,...,x_n\} \subset R^d$ be a sequence of framed-based cepstral vectors extracted from an audio stream in which there is at most one speaker turn. If we suppose that X is generated by a multivariate Gaussian process, a speaker change is detected at frame $i \in \{1,...,n\}$ by calculating the ΔBIC value at this instant [5].

$$\Delta BIC(i) = \frac{n}{2}\log\left|\Sigma_X\right| - \frac{i}{2}\log\left|\Sigma_{X_1}\right| - \frac{(n-i)}{2}\log\left|\Sigma_{X_2}\right| - \lambda P. \qquad (1)$$

where $\mu_{X,}\mu_{X_1}$ and μ_{X_2} are the sample mean vectors, Σ_X, Σ_{X_1} and Σ_{X_2} are the sample covariance matrices, knowing that $X_1 = \{x_1,...,x_i\}$, $X_2 = \{x_i,...,x_n\}$ and $P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log(n)$. The value of i that maximizes $\Delta BIC(i)$ is the most probable speaker change point and if $\Delta BIC(i_{\max}) > 0$ then $i_{\max}$ is confirmed to be a change.

3. HYBRID SEGMENTATION TECHNIQUES

3.1 DISTBIC

DISTBIC [7] segmentation is composed of a metric-based algorithm to detect speaker turns, followed by the BIC algorithm to validate them. The principle of this technique is the measure of a distance, derived from the Generalized Likelihood Ratio (GLR), between two adjacent frames shifted by a fixed step along the whole parameterized speech signal. This process gives the graph of distance as output. Then, a threshold is fixed to detect local maxima points that represent a speaker change. Speaker change points detected by the curve of distance will be confirmed as speaker turns by the BIC criterion. The use of the curve of distance to detect speaker change points permit improving segmentation for short segments but the threshold dependence and the high computation cost are the majority disadvantages.

3.2 T^2-BIC

T^2-BIC [11] is a hybrid technique which validates each speaker change point detected by Hotelling's T^2-statistic using the BIC criterion. Hotelling's T^2-statistic is a multivariate analogue of the square of the t-distribution [2]. The T^2-statistic is used to test if the mean of one normal population is equal to the mean of the other where the covariance matrices are assumed equal but unknown. In terms of segmentation, the problem can be viewed as testing the hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_1 : \mu_1 \neq \mu_2$ where μ_1, μ_2 are, respectively, the means of two samples of the audio stream, one containing the frame [*1,b*] and the second contains [*b,N*]. The likelihood ratio test is given by the following T^2-statistic:

$$T^2 = \frac{b(N-b)}{N}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2). \qquad (2)$$

where Σ represent the common covariance matrix. The T^2 value defined in (2) can be considered as a distance measure of two samples. Obviously, the smaller the value of T^2, the more similar the two samples distributions [10][11].

The T^2-BIC algorithm operates by fixing an analysis frame with L second length from the beginning of the parameterized audio stream and calculating the T^2 value in different points situated on this frame; the point that represents the highest value of T^2 is more probable to be a real speaker turns; then it can be validated by the BIC criterion. The T^2-BIC segmentation presents certainly some advantages. The selection, from the statistical criteria T^2, of a candidate speaker change permits to reduce computational costs. Thus, T^2-BIC offers a reduced calculation time compared to the BIC and DISTBIC segmentation. Besides, this technique works with an automatic threshold and presents a low false alarm. However, T^2-BIC is not reliable for the segmentation of audio documents that contain speaker changes close to each other. In fact, it requires the use of a time delay τ [11] between two consecutive speaker turns which can lead missing some break points.

3.3 DIS_T^2_BIC

The fast localization of speaker changes via the T^2-statistic criterion and the dependence of T^2-BIC and DISTBIC segmentation of empirical parameters incite us to develop a more robust and more reliable hybrid segmentation technique called DIS_ T^2_BIC. DIS_ T^2_BIC algorithm takes place in two phases. The first phase presents a new principle of detection of speaker changes that use the T^2-statistic as distance measure. The second phase is the validation of speaker changes already detected using the BIC criteria. A flowchart of the DIS_ T^2_BIC technique is represented in figure 2.

3.3.1 Detection of speaker change candidate points

DIS_T^2_BIC detect speaker turns by computing the value of T^2 between a pair of adjacent windows of the same size shifted by a fixed step along the whole parameterized speech signal. In the end of this procedure we obtain the curve of the variation of T^2 in time. The analysis of this curve shows that a speaker change point is characterized by the presence of a "significant" peak. A peak is regarded as "significant" when it presents a high value. In order to differentiate high peaks from low peaks, we define a fixed threshold from the proprieties of T^2-statistic. In fact, the T^2 value, given by (2), is distributed as T^2 with N-2 degrees of freedom [2] and the critical region is:

$$T^2 \geq \frac{(N-2)p}{N-p-1}F_{p,N-p-1}(\alpha) = T_0^2. \qquad (3)$$

where $F_{p,N-p-1}(\alpha)$ is the F-point for p and N-p-1 degrees of freedom with significance level α. A T^2 value lower than T_0^2 shows that the two samples are homogenous and consequently don't present a speaker change. Therefore a "significant" peak has to present a T^2 value higher than T_0^2 (see figure 3). So, break points can be detected easily by searching the local maxima of the T^2 curve that verify the criterion 3.

Contrary to T^2-BIC, the proposed method doesn't present a time delay τ which has the drawback of ignoring some speaker turns, and then it can detect changes even they are close to each other.

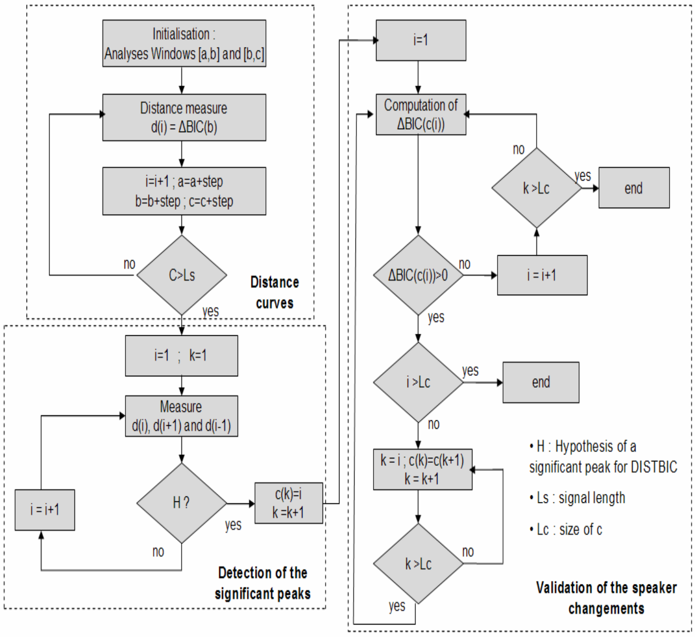
Initialisation : Analyses Windows [a,b] and [b,c]
Distance measure d(i) = ΔBIC(b)
i=i+1 ; a=a+step b=b+step ; c=c+step
no
C>Ls
yes
Distance curves
i=1 ; k=1
Measure d(i), d(i+1) and d(i-1)
i = i+1
H ?
c(k)=i k =k+1
yes
no
Detection of the significant peaks
i=1
Computation of ΔBIC(c(i))
no
k >Lc
yes
end
ΔBIC(c(i))>0
no
i = i+1
yes
i >Lc
yes
end
no
k = i ; c(k)=c(k+1) k = k+1
k >Lc
no
yes
• H : Hypothesis of a significant peak for DISTBIC
• Ls : signal length
• Lc : size of c
Validation of the speaker changements

Figure 2- Flowchart of the DIS_T^2_BIC approach

3.3.2 Improvement with the BIC criterion

Once the speaker turns are detected, we move to the validation of these break points using the BIC criterion in order to exclude the wrong results. Denote $\{T_1, \ldots, T_N\}$ as the set of speaker turns found in the first step, a ΔBIC value is computed for each pair of windows $[T_{i-1},T_i]$ $[T_i,T_{i+1}]$. When this value is positive, a speaker turn is identified at time i. Otherwise, the point s_i is discarded from the candidate set, then the ΔBIC value is now applied for a larger pair of windows $[T_{i-1},T_{i+1}]$ $[T_{i+1},T_{i+2}]$. At this stage, when segments are large enough, BIC criterion gives better validation results since model estimation becomes more accurate. DIS_T^2_BIC presents many advantages.

The determination of speaker changes from the curve of T^2 gives to DIST_T^2_BIC the possibility to detect speaker turns close to each others. Moreover, the use of the T^2-statistic criteria permits to reduce the computation cost and to have an automatic threshold decision independent of the type of the audio stream.

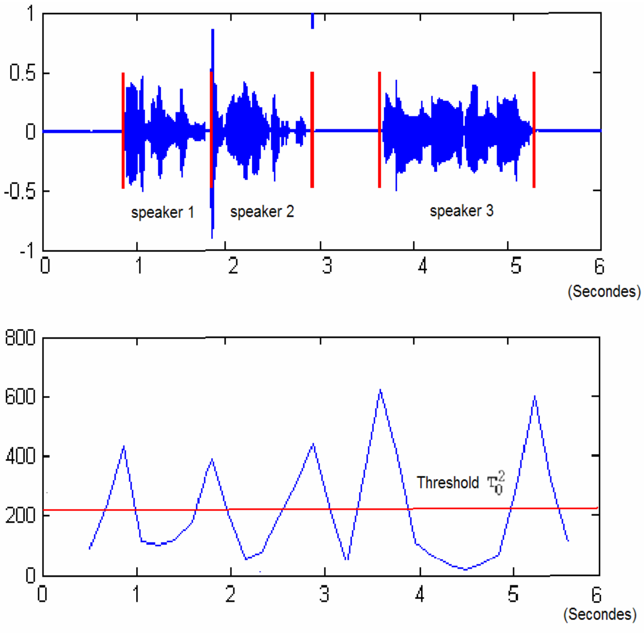

Figure 3- T^2 curve: detection of significant peaks

4. EXPERIMENTS AND RESULTS

4.1 Data description and parameterization

In order to evaluate the DIS_T^2_BIC segmentation method, experiments by computer is carried out. We first perform several tests of the possible configurations and we compare it with the most known hybrid approaches, including DISTBIC and T^2-BIC. The speech data uttered by male and female speakers are obtained from four resources sampled at 16 KHz:

- A conversation created from an Arabic audio base phonetically equilibrated "LISTE" [3] (short segments, 240 speaker turns).
- A TV news broadcast registered for ALJAZEERA channel [1] (50 speaker turns, 20 minutes).
- A conversation extracted from the IDIAP database [6] (85 speaker turns, 30 minutes).
- A conversation created from the TIMIT database (long segments, 48 speaker turns).

Experiments were conducted using 12 Mel-Frequency Cepstral Coefficients (MFCC) with a frame rate of 100 frames per second, each frame lasts 20 ms with an overlap of 50%. The evaluation of DIS_T^2_BIC is realized from the quantification of two error types: missed detection (MD) and false alarm (FA). An existed speaker change represents a missed detection when it is not detected and a detected changing point is counted as false alarm if there is no actual changing point. The missed detection rate (MDR) and false alarm rate (FAR) are defined and described in [7].

Generally, audio segmentation precedes a second task which consists in speaker clustering. Therefore, for audio segmentation, we have to focus more on the MDR since the FAR can be automatically ameliorated in audio clustering.

4.2 Experimental results

Table 1 report the FAR and the MDR results of segmenting the different type of audio streams described above using DISTBIC, T^2-BIC and DIS_T^2_BIC techniques.

The segmentation of the LISTE conversation that contains short segments presents a low MDR (6.66%), contrarily to the T^2-BIC segmentation (40.83%). The non presence of silence between the speeches of the same speaker explains the low values of FAR for this conversation (9.09% for DIS_T^2_BIC vs 7.69% for DISTBIC and 6.61% for T^2-BIC). The MDR respectively with the DISTBIC and T^2-BIC procedure (625% and 15.58 %) and with our segmentation algorithm (8.33%) for the TIMIT conversation are almost equal. Similar results are observed on the FAR (30.43 for BIC vs 34.24 for DIS_T^2_BIC). That means that both segmentation techniques are nearly equivalent with conversations containing long speech segments.

ALJAZEERA and IDIAP conversations are not exactly conform to our hypothesis fixed in the introduction and present sequences that contains simultaneously the speeches of different speakers. For this reason the MDR of these conversations with DIS_T^2_BIC method presents a high value (38% for ALJZEERA and 41.17% for IDIAP) but remains lower than with DISTBIC and T^2-BIC segmentation (respectively 44% and 54% for ALJZEERA and 45.88% and 50.58% for IDIAP). On the contrary, the FAR of both segmentation techniques are comparable (about 40 % for ALJZEERA and 43% for IDIAP).

Our method, DIS_T^2_BIC, presents many advantages. The detection of speaker change points with the curve of T^2 allows DIS_T^2_BIC to detect break points close each other. Come to the point we assume that the covariance matrices are equals, we can use more data to estimate the the covariance and reduce the impact of insufficient data in the estimation.

The experiments show that DIS_T^2_BIC method is more accurate than T^2-BIC segmentation in the presence of shorts segments.

On the other hand, DIS_T^2_BIC and DSTBIC techniques are comparables with audio stream containing long and short segments but DIS_T^2_BIC presents a lower computation cost. This is due to the use of T^2-statistic that avoids the computation of two full computation matrices at each point.

	DISTBIC		***T^2-BIC***		***DIS_T^2_BIC***	
	FAR (%)	*MDR (%)*	*FAR (%)*	*MDR (%)*	*FAR (%)*	*MDR (%)*
TIMIT	30.43	**6.25**	26.15	15.58	34.24	**8.33**
LISTE	7.69	8.75	6.61	40.83	9.09	**6.66**
ALJAZEERA	36.70	44	33.33	54	40.47	**38**
IDIAP	45.16	45.88	47.85	50.58	42.17	**41.17**

Table 1 - DISTBIC, T^2-BIC and DIS_T^2_BIC results

5. CONCLUSION

In this paper, we proposed a new hybrid segmentation technique that associates the metric-based segmentation and the model-based segmentation. Our technique applies first a new measure distance algorithm using the Hotelling's T^2 statistic to detect the most probable speaker turns. In the second step it uses the BIC criterion to validate changes already detected and then compensate the false alarm rate. DIS_T^2_BIC has the advantage to detect more break points than T^2-BIC algorithm. In fact, in a contrast with T^2-BIC, which introduces a time delay, ignoring some break points, DIS_T^2_BIC does not neglect anyone of them thanks to the use of the T^2 curve with a fixed threshold. Then, contrary with DISTBIC, the use of T^2-statistic as distance measure allows detecting break points for short segments with a low computation cost. Our experiments show that DIS_T^2_BIC can detect speaker turns even close to each other and give better results than T^2-BIC and DISTBIC technique. In the future and to ameliorate the detection of speaker turns in the first step of our technique, we plan to develop an automatic threshold variable with acoustic characteristics of the audio stream.

REFERENCES

[1] Aljazeera broadcasting channel.
http://www.aljazeera.net

[2] T.Anderson, An introduction to multivariate statistical analysis, John Wiley & Sons, Inc., New York, NY, 1985.

[3] M.Boudraa, B.Boudraa, and B.Guerin, "Mise en place de phrases arabes phonétiquement équilibrées",XIXème JEP, Bruxelles, 1992.

[4] M.Cettolo and M.Federico, "Model selection criteria for acoustic segmentation", in Proc. ISCA Tutorial and Research Workshop ASR 2000, Paris, France, September 2000.

[5] S.Chen and, P.Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion", in Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[6] Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), http://www.idiap.ch.

[7] P.Delacourt and C.J.Wellekens, "DISTBIC: a speaker based segmentation for audio data indexing", Speech Communication, vol. 32, pp. 111-126, Sept. 2000.

[8] R. Huang and J.H.L. Hansen, "Advances in unsupervised audio segmentation for the broadcast news and ngsw corpora" in Proc. ICASSP'04, Montreal, May 2004, pp. 741-744.

[9] M.Siegler, U.Jain, B.Raj, and R.M.Stern "Automatic segmentation, classification and clustering of broadcast news audio", in Proc. DARPA Speech Recognition Workshop, Chantilly, Virginia, USA, 1997, pp. 97-98.

[10] S. Wegmann, P. Zhan and L. Gillick, "Progress in Broadcast News Transcription at Dragon Systems", in ICASSP 99, Phoenix, Arizona, March 1999.

[11] B.W.Zhou and J.H.L.Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion," in Proc. ICSLP'2000, Vol. 1, Beijing, China, Oct. 2000, pp.714-717.