Speaker Change Detection Method Evaluated on Arabic Speech Corpus

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

[1] Département TIC, ENIT, BP 37, Campus Universitaire, 1002 le Belvédère, Tunis Tunisia
Emails: kadri_hachem@yahoo.fr, 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*—Audio speaker segmentation is a necessary pre-processing step for many applications especially for multimedia data automatic indexing. This paper deals with the problem of audio segmentation when no prior knowledge of speakers is assumed. In this work, we introduce a new audio segmentation method that can detect speaker turns close to each others (about 2 seconds) using an automatic decision threshold independent to the audio stream acoustic characteristics. The proposed method is organized in three steps. First, an energy-based segmentation is considered as a pre-processing task applied to eliminate long silences that can perturb the detection of speaker turns in the audio stream. In the second step and to detect the most probable speaker turns, a new algorithm based on the Hotelling's T^2-Statistic criterion is developed. Then, we use the Bayesian Information Criterion (BIC) to validate the already detected change points. Experimental results on Arabic speech corpus show that our method has the advantage of high accuracy speaker change detection with a low computation cost.**

I. INTRODUCTION

SPEAKER segmentation is the problem of finding speaker segment boundaries when a speaker begins and stop speaking in an audio speaker stream. This segmentation of audio data is of interest to a broad class of applications like surveillance meetings summarization or indexing of broadcast news.

Unsupervised speaker segmentation approaches suppose that there is no information about the speakers and their number is known a priori. It can be classed into three categories [5][8][9]: energy-based segmentation, metric-based selection and model-selection-based segmentation.

- Energy-based segmentation: silence in the input audio stream is detected either by a decoder or directly by measuring and thresholding the audio energy. The segments are then generated by cutting the input at silence locations.
- Metric-based segmentation: the audio stream is segmented at maxima of the distances between neighboring windows placed in evenly spaced time intervals.
- Model-selection-based segmentation [4]: assuming that data are generated by a Gaussian process, speaker changes are detected by using a statistical decision criterion within a sliding window through the audio stream.

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 [6] tested different metric criteria to associate them to the BIC criterion such as the Kullbach-Leibler distance, the similarity measure and the measure derived from the Generalized Likelihood Ratio (GLR). Still, this method encountered many problems in case of short segments and requires a high computation cost. On another issue, Zhou [10] recommends the use of T^2-Statistic as 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 method to improve the segmentation of short segments, giving better results than T^2-BIC and lower computation cost than DISTBIC.

This paper is organized as follows: section 2 describes speaker segmentation techniques based on BIC and Hotelling's T^2-statistic criteria. In section 3, we introduce the proposed method. Section 4 discusses our experimental results, and finally in section 5, we present our conclusions.

II. BIC BASED AND HOTELLING'S T^2 BASED SEGMENTATION

A. BIC segmentation

The BIC is a Bayesian model selection criterion 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].

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

where μ_X, μ_{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.

B. DISTBIC

DISTBIC [6] 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.

C. T^2-BIC

T^2-BIC [10] 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 when we wish 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 (3) 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**.**

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 a speaker changes close to each others. In fact, it requires the use of a time delay τ [10] between two consecutive speaker turns which can lead to ignore some break points.

III. The Proposed Method

The proposed method is a three-step analysis whose purpose is to compensate the speaker turns which can not be detected by T^2-BIC (see figure 2). In fact the time delay t introduced by T^2-BIC has the drawback of ignoring some speaker turns, especially when they are close to each other. A first step eliminates long silences using an energy-based segmentation. The second step uses the T^2-statistic to locate the boundaries in the audio stream in a different way, which will be more detailed hereafter, than that of T^2-BIC method. Then the BIC rule is employed to approve or reject these boundaries in the third step.

A. Energy-Based Segmentation

Energy based-approaches have been widely used and are been particularly easy to implement. Basically, silence periods in the input signal are detected, and segment boundaries are hypothesized in such silence periods if some additional constraints are satisfied, like minimum length of the silence period [8].

In our energy-based segmentation method, a signal energy histogram is generated by computing the energy over a 20ms window. The window is then shifted by 10 ms and the next energy is computed. This histogram presents two modes: the low mode for silence and the high one for speech. The point of intersection between these two modes is the boundary between speech and silence.
Hence, the decision threshold is calculated by using the EM algorithm which can estimate efficiently the parameters of the two Gaussians distributions representing the histogram modes.

B. Detection of speaker change candidate points

Our method detects 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 (3), 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, as described in II.C, 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 1). So, break points can be

detected easily by searching the local maxima of the T^2 curve that verify the criterion 3.

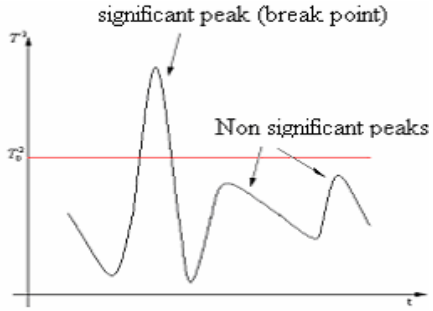

Fig.1. T^2 curve: detection of significant peaks

C. Validation by the BIC criterion

The third step of our technique validates the results of the first step. Denote $\{s_1, \ldots, s_N\}$ as the set of speaker change points founded in step 1, a ΔBIC value is computed for each pair of windows $[s_{i-1}, s_i]$ $[s_i, s_{i+1}]$. If the value is positive, 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). In this step the use of the BIC is more appropriate since the length of the segments is larger than those in the first step and this permits good model estimation [6].

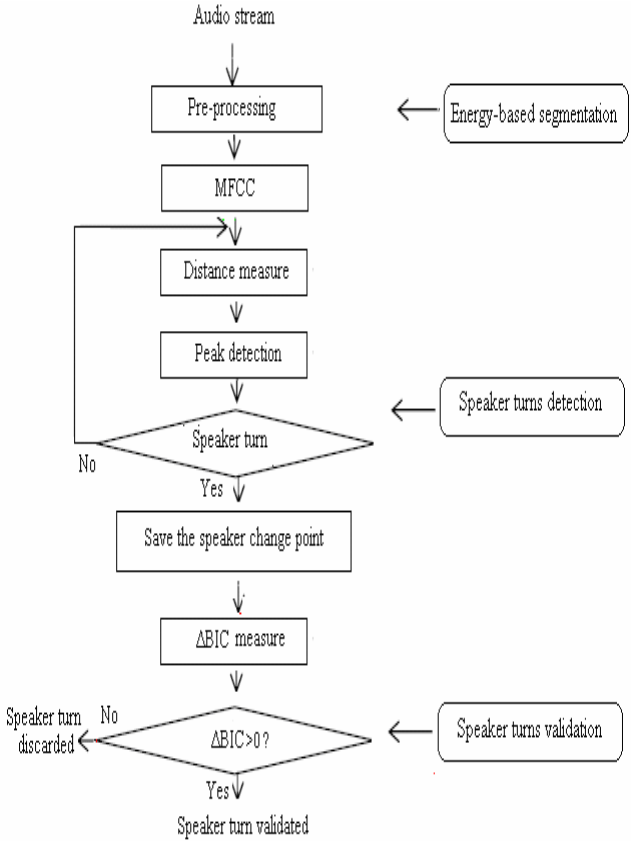

Fig.2. Diagram of the proposed method

IV. Experiments and Results

A. Data description and parameterization

In order to evaluate the new segmentation method, experiments by computer is carried out. The speech data uttered by male and female speakers are obtained from four resources sampled at 16 KHz:

- A conversation created by concatenating sentences of 2s on an average 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).

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 our method 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 [6]. 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.

$$MDR = 100 \times \frac{number\ of\ MD}{number\ of\ actual\ speaker\ turns}\% \qquad (4)$$

$$FAR = 100 \times \frac{numberFA}{number\ of\ speaker\ turns + numberFA}\% \qquad (5)$$

B. Experimental results

Table I report the FAR and the MDR results of segmenting the different type of audio streams described above using DISTBIC, T^2-BIC and the proposed techniques. The segmentation of the LISTE conversation that contains short segments presents a low MDR (4.82%), contrarily to the T^2-BIC segmentation (40.83%). This is due to the elimination of the time delay τ between two consecutive speaker changes necessary for the T^2-BIC algorithm. The non presence of silence between the speeches of the same speaker explains the low values of FAR for this conversation (8.19 for our method vs 7.69% for DISTBIC and 6.61% for T^2-BIC).

ALJAZEERA conversation presents sequences that contain simultaneously the speeches of different speakers. For this reason the MDR of this conversation with our method presents a high enough value (31.2%) but remains lower than with DISTBIC and T^2-BIC segmentation (respectively 44% and 54%). On the contrary, the FAR of both segmentation techniques are comparable (about 35%). Our method, presents many advantages. The detection of speaker change points with the curve of T^2 allows the detection of break points close each other. Come to the point

TABLEI
DISTBIC, T^2-BIC AND OUR METHOD RESULTS

	DISTBIC		**T^2-BIC**		**Our method**	
	FAR (%)	MDR (%)	FAR (%)	MDR (%)	FAR (%)	MDR (%)
LISTE	7.69	8.75	6.61	40.83	8.19	**4.82**
ALJAZEERA	36.70	44	33.33	54	35.18	**31.20**

we assume that the covariance matrices are equals, we can use more data to estimate the covariance and reduce the impact of insufficient data in the estimation.

The experiments show that the proposed method is more accurate than T^2-BIC segmentation in the presence of shorts segments. On the other hand, our method and DSTBIC techniques are comparables with audio stream containing long and short segments but our method 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.

V. CONCLUSION

In this paper, we proposed a new segmentation method 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. The proposed method 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, our algorithm 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 the proposed method 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", XIX ème JEP, Bruxelles, Mai, 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] P.Delacourt and C.J.Wellekens, "DISTBIC: a speaker based segmentation for audio data indexing", Speech Communication, vol. 32, pp. 111-126, Sept. 2000.

[7] 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.

[8] T.kemp, M.schmidt, M.Westphal and A. Waibel "Strategies for automatic segmentation of audio data", in Proc. IEEE ICASSP'00, Istambul, Turkey, june 2000,vol. 3 pp.1423-1426.

[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] 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, eijing, China, Oct. 2000, pp.714-717.