Journal of Information Hiding and Multimedia Signal Processing ©2010 ISSN 2073-4212
Ubiquitous International Volume **1**, Number **4**, October **2010**

Speaker Clustering of Stereo Audio Documents Based on Sequential Gathering Process

Halim Sayoud and Siham Ouamour

Institute of Electronics and Computer Engineering
USTHB University
BP 32 Al-Alia, Bab-Ezzouar, Alger, Algeria
halim.sayoud@gmail.com; siham.ouamour@gmail.com

Received April 2010; revised July 2010

Abstract. *This paper focuses on the use of sequential speaker clustering of stereo audio documents to obtain a classification of the different speech segments contained in those documents, according to the speakers who are participating in the audio recording. In general, speaker clustering is used as a second step in a global system of speaker diarization, where the first step deals with the task of speaker segmentation. However, in some applications, the term speaker diarization is confused with speaker clustering. In such applications, the homogeneous segments are automatically separated, like in telephone answering machines or vocal boxes, where the vocal messages are already separated by a sound beep. Even though in our global project we use two main techniques based on speaker localization and speaker discrimination, in this paper we will describe only the second technique, which uses a sequential clustering approach, in order to gather the similar homogeneous segments into classes of speakers. Each class contains the global intervention of only one speaker in the entire audio document. The sequential clustering approach uses a mono-gaussian measure (μ_G) that allows us to assess the degree of similarity between the different homogeneous segments. The application concerns the clustering of stereo debates between several speakers who are located at different positions in the meeting-room. For the evaluation, experiments are conducted on a stereophonic database called DB15, which is composed of 15 scenarios of about 3.5mn each and containing two or three speakers speaking sequentially in every scenario. The new algorithm shows good performances, when the length of the speech segments is over 4s.* **Keywords:** speaker clustering, speaker diarization, sequential clustering, stereo audio document, second order statistical measures, mono-gaussian measures.

1. Introduction.

Voice remains one of the most important means used by human beings to transmit information to external world. Many organisms digitize and archive these information, allowing people to consult them in the future. However, read, listen or watch these multimedia documents in order to extract particular information related to a particular speaker, seems to be a difficult task. It is also interesting to know the sequence of speakers in the conversation. Moreover, in meeting recordings, usually, more than one microphone is available to collect the speaker's intervention. Speaker diarization systems response to these needs by dividing the audio document into homogeneous areas: it is known as the segmentation task [1] [2] [3] and gathering the areas coming from the same speaker into a same cluster: this is the clustering task [4]. However, some applications do not need the segmentation task since the homogeneous areas are already separated by an automatic separation system such as in vocal boxes or telephone answering machines. According to Reynolds and Torres-Carrasquillo [5], there are 3 main applications in speaker diarization, namely: Broadcast news indexing [6], phone conversations indexing and recorded meeting indexing. We propose in this paper a sequential speaker clustering algorithm using a mono-gaussian measure and which is applied to multi-conference (multi-speaker debates) indexing. Speaker clustering consists in gathering the similar homogeneous segments into classes of speakers. At the end of the clustering process, we obtain a number of clusters equal to the number of speakers present in the audio stream. Each cluster contains the global intervention of each speaker in the document. In fact, most of the proposed systems involve hierarchical clustering of the data into clusters where the optimal number of speakers is unknown a priori. A very commonly used method is called bottom-up clustering, where multiple initial clusters are iteratively merged until the optimal number of clusters is reached, according to some stopping criterion [7]. This stopping criterion is often estimated empirically and adapted to the database. However, by using the sequential clustering [8], which consists in gathering the segments sequentially over the time, the problem of the stopping criterion is resolved since the clustering algorithm stops when all the segments are processed. Moreover, on one hand, this clustering takes in consideration the neighborhood relationships between the segments, and on the other hand, it can be used in real time applications because the segments are processed sequentially. Hence, we have opted to use this sequential clustering in order to evaluate its performances. Concerning the mono-gaussian μ_G measure, this one was chosen because it offers the possibility to make discrimination between segments of different durations. Furthermore, the mono-gaussian measures are easy to implement, fast in calculation, do not need learning step and give good performances when the speech segment duration is over 2 s. The sequential clustering algorithm is evaluated on a stereo database that contains 15 meeting recordings (scenarios) and the recording process is ensured by using two distant cardioid microphones that are placed in opposition at fixed positions.

2. **Related works.** Speaker clustering consists in gathering all the homogeneous speech segments belonging to a same speaker. Most of the proposed systems involve some sort of hierarchical clustering of the data into clusters, where the optimum number of speakers or their identities are unknown a priory. A very commonly used method is called bottom-up clustering, where multiple initial clusters are iteratively merged until the optimum number of clusters is reached, according to some stopping criterions [9]. According to the processing time and mode, we can classify the clustering techniques into two types: on-line techniques and off-line techniques. As on-line clustering techniques, we can quote the works of Mori and Nakagawa in 2001 [10], where a clustering algorithm based on the Vector Quantization (VQ) distortion measure [11] is proposed. It starts processing with one speaker in the code-book and incrementally adds new speakers whose VQ distortion exceeds a threshold in the current code-book. In [12] Rougui proposed a GMM based system, using a modified Kullback-Leibler (KL) distance between models. Change points are detected as the speech becomes available and data is assigned to either speaker present in the database or a new speaker is created, according to a dynamic threshold [9]. In 2007, the authors of [13] proposed a new online speaker clustering algorithm using decision tree and decision queue to cluster the segments.

Concerning the offline clustering, most of the reviewed algorithms use hierarchical schemes. We can quote the following works: In 2004, Meignier and Gauvain proposed a Generalized Likelihood Ratio (GLR) based metric with two penalty terms, penalizing for large number of segments and clusters in the model, with tuning parameters [14]. Iterative Viterbi decoding and merging iterations find the optimum clustering, which is stopped using the same metric. Other research is done using GLR as distance metric: Siu et al. for pilot-controller clustering in 1992 [15], and Jin, Laskowski, Schultz and Waibel in 2004 for meetings diarization using the Bayesian information criterion (BIC) as stopping criterion [16]. However, the most commonly used distance and stopping criterion is the BIC criterion. In the same field, in 2006, Xavier proposed a speaker diarization method for meeting rooms. It looks into the algorithms and the implementation of an offline speaker segmentation and clustering system for a meeting recording, where usually more than one microphone is available. He implements a train-free speech/non-speech detection on such signal and processes the resulting speech segments with an improved version of the mono-channel speaker diarization system [9]. Furthermore, most popular speaker-clustering methods employ hierarchical agglomerative clustering (HAC), as it is the case in the following works: Gish et al., 1991 [17]; Jin et al., 1997 [18]; Solomonoff et al., 1998 [19]; Chen and Gopalakrishnan, 1998 [20]; Reynolds et al., 1998 [21]; Johnson and Woodland, 1998 [22]; Ajmera et al., 2002 [23]; and Moh et al., 2003 [24]. Hierarchical agglomerative clustering (HAC) generates a cluster tree by sequentially merging the utterances deemed similar to each other. Then, the tree is cut using the bayesian information criterion [25]; [20]; [26] to retain the appropriate number of clusters. In a different vision, the authors of [27], propose a system of speaker turn detection and clustering, which is ensured by the use of the Direction of Arrival (DOA) information. Purification of the resultant speaker clusters is then done by performing a Gaussian Mixture Model (GMM) modeling on acoustic features. The system achieved a competitive overall DER of 15.32% for the NIST Rich Transcription 2007 evaluation task. In 2009, the authors of [28] presented a novel Fuzzy-based Hierarchical Clustering algorithm (FHC) for speaker clustering and investigated its performances with different similarity thresholds. Comparing with the other conventional clustering algorithms, their method shows quite competitive performances. In the work presented in [29], in 2009, a fusion based speaker clustering system is developed, where the speaker segments are modeled by acoustic and prosodic representations. The idea here is to model the speaker prosodic characteristics and add them to the basic acoustic information estimated from the speaker segments, which leads to a clustering improvement in some cases.

Concerning our motivation in this research work, there are three main points that have motivated us:

- Firstly, we have noticed a lack in research works involving sequential speaker clustering: that is the reason to choose this type of clustering;
- Secondly, our global application was ”speaker clustering in meeting-rooms”: that is why we have proposed the use of two microphones (stereophonic speech) in order to make a spatial localization (second part of this project);
- And finally, trying to solve the problem of stopping criterion, we have proposed a sequential clustering algorithm that is based on a mono-gaussian measure. This last one has been introduced to assess the degree of similarity between homogeneous segments of different durations and to gather the speech segments belonging to a same speaker, without the need of stopping criterion.

3. **Methods of speaker clustering.** Most systems of speaker clustering use the hierarchical clustering like agglomerative techniques [30] or multiple channels techniques [31]. Other recent researches use the combination between the agglomerative and the sequential clustering for the task of speaker diarization [32]. In this research work, we have chosen the sequential clustering because, on one hand, this technique takes into consideration the neighborhood relationship between the segments, which favors the gathering of the segments that are close in time; on the other hand, and contrarily to hierarchical clustering, sequential techniques can be used in real time applications because the segments are processed sequentially in time when these last ones are collected. For the similarity

measure, we chose the μ_G measure, which allows assessing the degree of similarity between two homogeneous segments with different lengths [33].

3.1. **Sequential clustering.** The principle of this clustering is to consider the first segment as a first cluster, after that, the other homogeneous segments are compared sequentially to it using a similarity distance. If the distance is less than an appropriate threshold, the new segment is added to the old cluster; otherwise, a new cluster is created containing this new homogeneous segment. This process continues until all the homogeneous segments are processed chronologically, one after the other (figure 1).

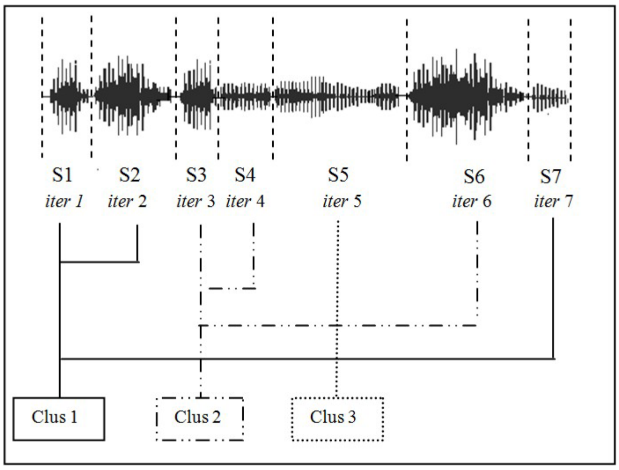

Figure 1. Principle of the sequential clustering. S represents a homogeneous segment, iter represents an iteration and Clus represents a cluster. We can see the resulting clusters in the bottom of this figure (inside the squares).

3.2. **Mono-gaussian measures (or Second Order Statistical Measures).** The proposed method uses mono-gaussian models based on the second order statistics, and provides some similarity measures able to make a comparison between two speakers (speech segments) according to a specific threshold. We recall bellow the most important properties of this approach [33].

Let $\{x_t\}_{1\leq t\leq M}$ be a sequence of M vectors resulting from the P-dimensional acoustic analysis of a speech signal uttered by speaker $\boldsymbol{X}$. These vectors are summarized by the mean vector $\overline{x}$ and the covariance matrix X:

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \tag{1}$$

and

$$X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \tag{2}$$

Similarly, for a speech signal uttered by speaker $\boldsymbol{Y}$, a sequence of N vectors $\{y_t\}_{1\leq t\leq N}$ can be extracted. By assuming that all acoustic vectors extracted from the speech signal uttered by speaker $\boldsymbol{X}$ are distributed like a gaussian function, the likelihood of a single vector yt uttered by speaker $\boldsymbol{Y}$ is:

$$G(y_t/\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}(detX)^{1/2}} e^{-(1/2)(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \tag{3}$$

"det" represents the determinant.
If we assume that all vectors y_t are independent observations, the average log-likelihood of can be written as

$$\overline{G}_x(y_1^N) = \frac{1}{N} logG(y_1...y_N/\boldsymbol{X}) = \frac{1}{N}\sum_{t=1}^{M} logG(y_t/\boldsymbol{X}) \tag{4}$$

by replacing

$$y_t - \overline{x}$$

by

$$y_t - \overline{y} + \overline{y} - \overline{x}$$

and using the property

$$\frac{1}{N}\sum_{t=1}^{N}((y_t - \bar{y})^T X^{-1}(y_t - \bar{y})) = tr(YX^{-1}) \tag{5}$$

where "tr" represents the trace of the matrix,
we get

$$\frac{2}{P}\overline{G}_x(y_1^N) + log(2\pi) + \frac{1}{P}log(det(Y)) = \frac{1}{P}\left[log(\frac{det(Y)}{det(X)}) - tr(YX^{-1}) - (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] \tag{6}$$

the Gaussian likelihood measure μ_G is defined by:

$$\mu_G(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{P}\left[-log(\frac{det(Y)}{det(X)}) + tr(YX^{-1}) + (\bar{y} - \bar{x})^T X^{-1}(\bar{y} - \bar{x})\right] - 1 \tag{7}$$

we have:

$$Argmax_x \overline{G}_x(y_1^N) = Argmin_x \mu_G(\boldsymbol{X}, \boldsymbol{Y}) \tag{8}$$

One possibility for symmetrising this measure is to weight this measure and its dual term by the coefficients M and N. Thus, the formula of the μ_G statistical measure is given as follows [34]:

$$\mu_{G\beta}(\boldsymbol{X}, \boldsymbol{Y}) = (M.\mu_G(\boldsymbol{X}, \boldsymbol{Y}) + N.\mu_G(\boldsymbol{Y}, \boldsymbol{X}))/(M + N) \tag{9}$$

3.3. **Analysis of the homogeneous segments.** Each homogeneous stereo segment is analyzed as follows: At the beginning, we transform the stereo segment into a mono speech segment by choosing the channel for which the speech segment has a higher energy. After that, the speech signal is decomposed in frames of 512 samples (32 ms) at a frame rate of 256 samples (16 ms). The signal is not pre-emphasized. For each frame, a Fast Fourier Transform is computed by providing 256 values representing the short term power spectrum in the 0-8 kHz band. This Fourier power spectrum is then used to compute 37 filter bank coefficients called MFSC or Mel Frequency Spectral Coefficients [35] (figure 2). At the end, each segment is decomposed into several stationary frames (with 37 MFSC coefficients by frame). The next step is to compute the mean vector and covariance matrix in every frame. Thus, the mean vector is represented by 37 components and the covariance matrix is represented by 37x37 components [36].

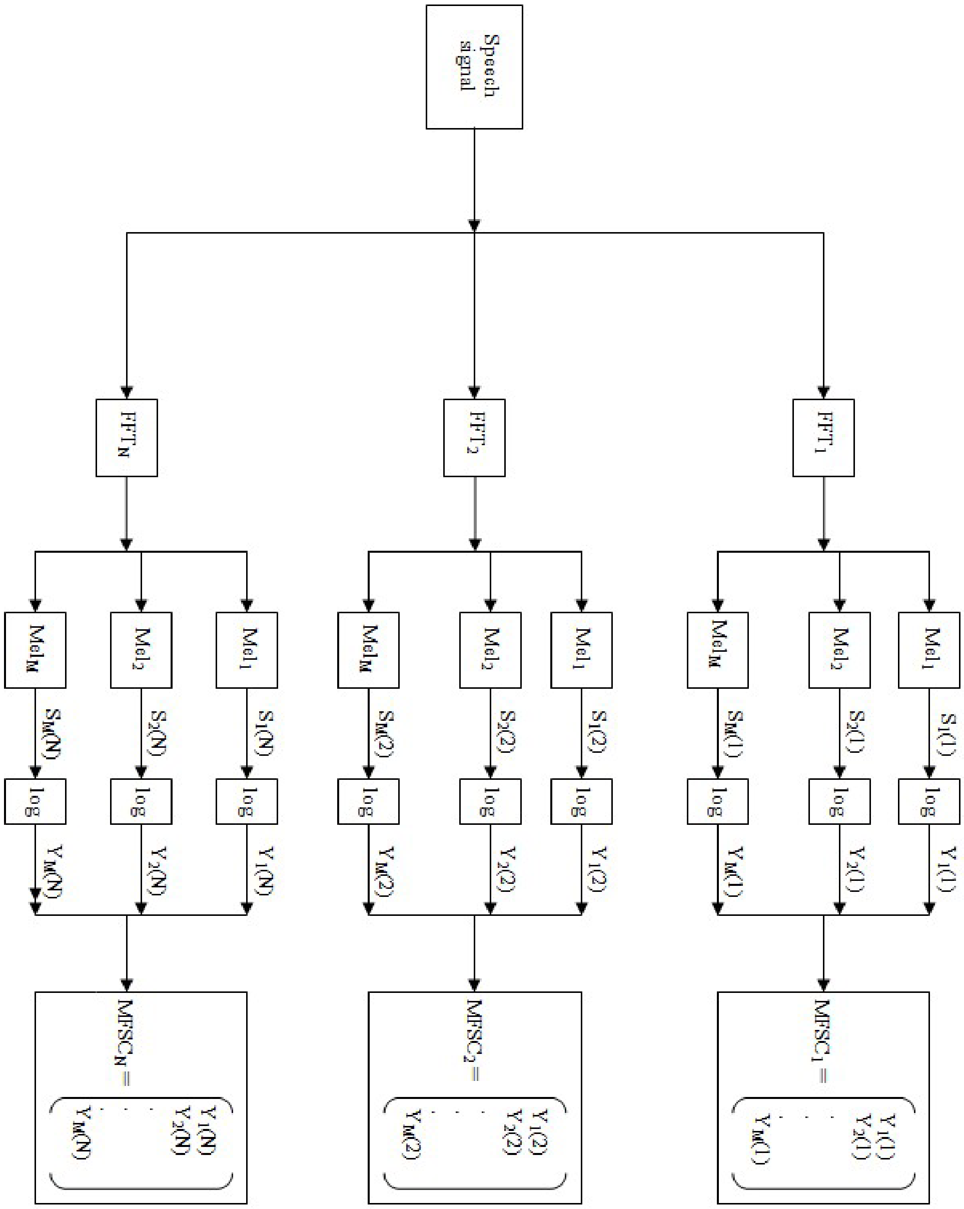

Figure 2. Principle of the MFSC extraction. The speech signal is spitted into N frames and in each frame the FFT transform is estimated in order to compute the Mel-logarithmic energies (MFSC).

3.4. **Clustering algorithm.** To achieve the clustering task, we have developed an algorithm based on a sequential technique, which is characterized by the following points:

- The different homogeneous segments are represented by their instants of beginning and end, and their numbers in the audio document (see the labels in the bottom of figure 3);

- The new technique consists in the application of the μ_G similarity measure between every pair of segments, in order to gather the similar homogeneous segments with regard to the speakers present in the audio document (see figure 4). This is ensured by using a sequential process of all the segments (processed over the time), as described in the following algorithm:

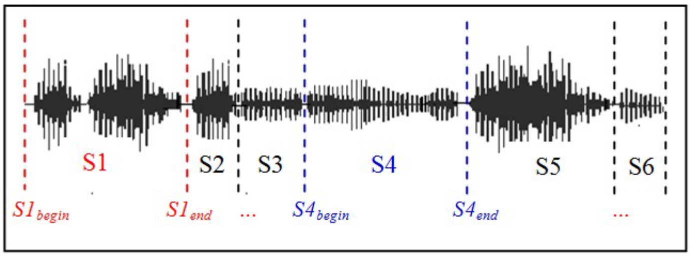

Figure 3. Every homogeneous segment is indexed by two labels: instant of beginning and instant of end (see the labels in the bottom).

If distance[segment(i), segment(j)] $\leq$ *threshold*
$\rightarrow$*Then segment(i) and segment(j) come from the same speaker;*

If distance[segment(i), segment(j)] $>$ *threshold*
$\rightarrow$*Then segment(i) and segment(j) belong to different speakers;*

Redo the process by incrementing the indices.

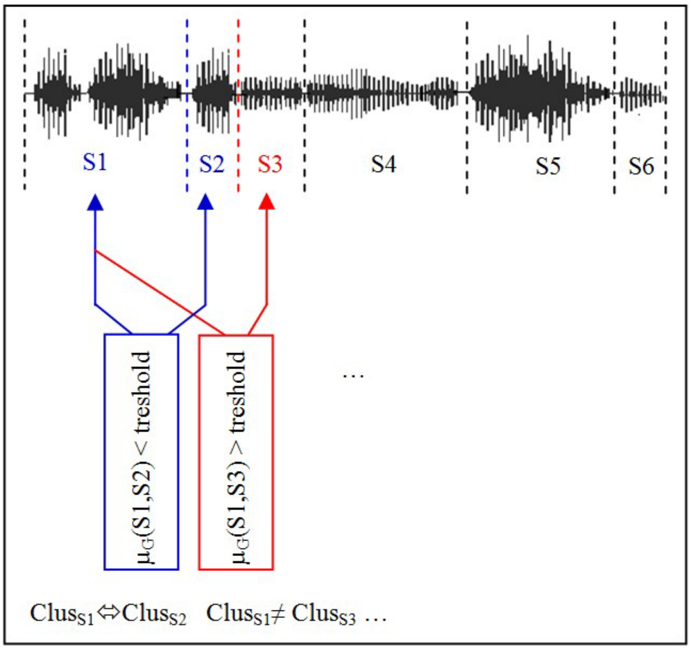
S1
S2
S3
S4
S5
S6
$\mu_G(S1,S2) <$ treshold
$\mu_G(S1,S3) >$ treshold
...
$Clus_{S1} \Leftrightarrow Clus_{S2}$ $Clus_{S1} \neq Clus_{S3}$ ...

Figure 4. The μ_G measure is computed between every pair of segments: a distance less than the threshold means that the segments belong to the same speaker (same cluster).

- Then, a new reorganization of the different clusters is applied by gathering the segments belonging to each speaker and assigning them new numbers, with the corresponding time of beginning and time of end. The estimated number of speakers will be equal to the new total number of clusters that are found in the audio document (see figure 5).

- Finally, a graphical representation over the time of the different homogeneous segments is done, which will indicate the speaker (cluster number) who has spoken at each segment (see figure 6).

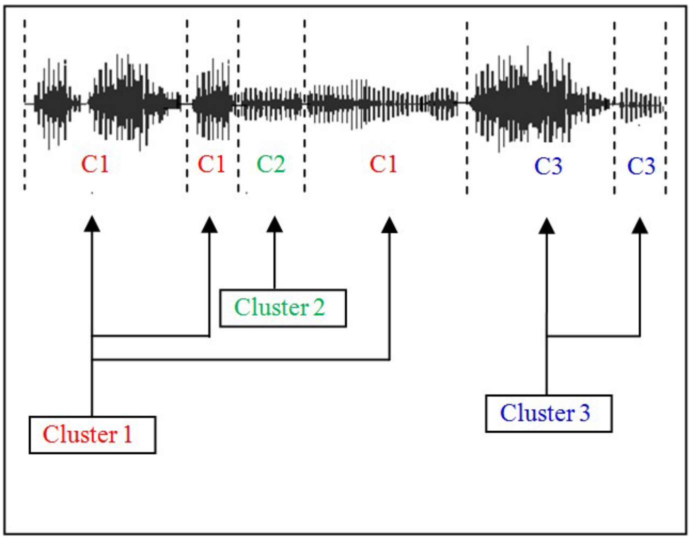

Figure 5. All the equivalent segments are given the same cluster number. The final number of clusters will indicate the number of speakers sharing the discussion.

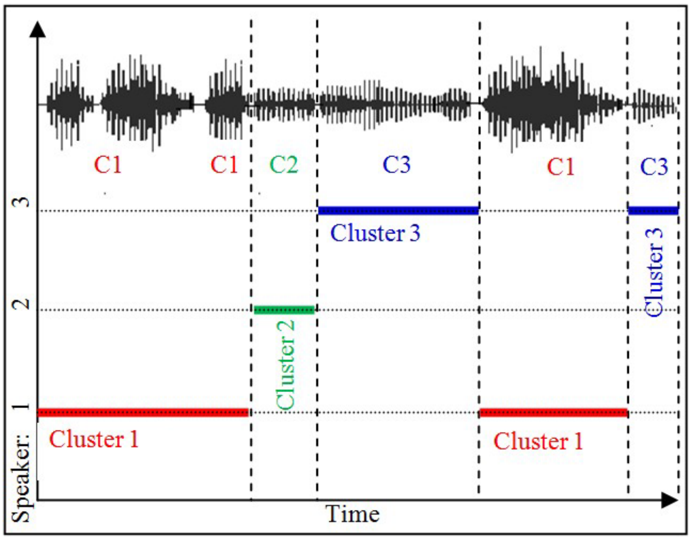

Figure 6. Graphic representation of the clusters over the time.

At this stage, it is easy to collect the speech of any speaker present during the discussion, since all his speech clusters are delimited in time and already memorized.

4. **Speech database.** The sequential clustering algorithm is evaluated on a stereo database which we called DB15. The audio database includes 15 recordings of multi-speaker meetings that are divided into 10 conversations between 2 speakers and 5 conversations between three different speakers who are speaking alternatively in a natural manner. Each speech recording is performed in stereo form by two cardioid microphones placed in opposition and separated by a fixed distance. The duration of each scenario is between 3mn and 4mn. Thus, the total duration is about 40mn of speech. The speakers are seated at one of the 3 fixed positions of the meeting room: Left, Middle or Right (figures 7 and 8). The distance between the 2 microphones is 1m and the global number of speakers used to construct these scenarios is 6 different speakers: 4 females and 2 males.

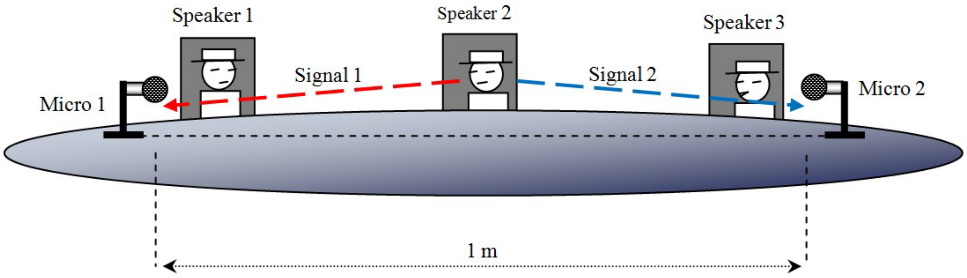

FIGURE 7. Example of a disposition with 3 speakers who are present in the meeting-room: the speech signal is recorded by 2 cardioid microphones.

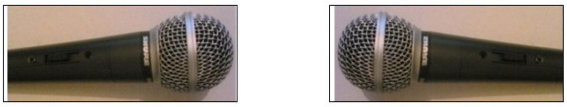

FIGURE 8. Pictures of the cardioid microphone: left disposition and right disposition respectively. The left one is oriented toward the right and the right one is oriented toward the left.

5. **Results and discussion.** We have implemented a sequential algorithm using a mono-gaussian measure (μ_G) in order to gather the different homogeneous segments into a same cluster. The overall results are exposed and discussed in detail. For concreteness and in order to see the evolution of the algorithm, we present, here below, some results obtained with scenario 1, which is taken as an example:
- Table 1 presents the different homogeneous speech segments of the scenario 1, numbered from 1 to 9 (duration of about 3min 30s). This table displays the moments of beginning and end of each segment.

TABLE 1. Distribution of the homogeneous speech segments of the scenario 1. This table shows the number, the instant of beginning and the instant of end of each segment.

Segment begining (s)	1	19	52	76	93	113	152	174	201
Segment end (s)	19	52	76	93	113	152	174	201	210
Segment number	1	2	3	4	5	6	7	8	9

- For the clustering step, we use the μ_G similarity measure which allows gathering the similar homogeneous segments (belonging to the same speaker). The clustering algorithm, in this example, finds 3 clusters against 2 clusters really existing in the considered audio stream (see table 2). Once we gather the different segments into clusters, the algorithm displays the different segments with their corresponding cluster number obtained after the clustering process (as described in table 2).

Figure 9 represents the final result of the sequential clustering using the μ_G measure, where, we notice 3 different clusters (3 speakers) presented versus their chronological time of participation in the audio document (3mn 30s of length). For a comparison purpose, we have also represented on figure 10 the 2 real speakers present in the same scenario. The comparison between the two figures 9 and 10 shows that the algorithm has made 2 errors in this scenario: the first error concerns the segment 6 that is gathered in the first cluster

Table 2. Display of the segments with their cluster numbers in scenario 1. This table shows the number, the instant of beginning and the instant of end of each segment, after the clustering process.

Segment begining (s)	1	19	52	76	93	113	152	174	201
Segment end (s)	19	52	76	93	113	152	174	201	210
Segment number	1	2	1	2	1	1	1	3	1

(false alarm error) while this last one belongs to the second cluster really, and the second error concerns the segment 8 which is considered coming from another speaker (cluster 3) while it really belongs to the second speaker (this error is called missed detection error). Thus, we have defined two scores:

- Score of Good Clustering (GC) defined by the ratio between the number of homogeneous segments that are well gathered and the total number of homogeneous segments:

$$\mathrm{GC} = \frac{\text{number of segments that are well gathered}}{\text{total number of segments}} * 100 \tag{10}$$

- Score of Homogeneity of the Clusters (CH) represents the mean of all the cluster homogeneities of the scenario (eg. in the case of the scenario 1, there are 2 clusters, so two cluster homogeneities). The cluster homogeneity of each cluster i (CHi) is defined by the ratio between the number of clusters that belong really to this cluster and the number of all segments gathered in cluster i (number of real segments plus false alarms):

$$\mathrm{CHi} = \frac{\text{number of segments belonging to cluster i}}{\text{number of all segments of cluster i}} * 100 \tag{11}$$

$$\mathrm{CH} = \frac{1}{N}\sum_{i=0}^{N} \mathrm{CHi} \tag{12}$$

with N representing the number of clusters in the scenario.

The different scores of clustering and homogeneity, estimated in each experiment, are given by figures 11 to 14.

- Figures 11 and 12 display respectively the clustering score and the homogeneity score obtained for each scenario in the DB15 stereo database. We can notice that the GC reaches the 100% rate for 8 scenarios; it is between 85% and 91% for 4 scenarios and between 66% and 78% for 2 scenarios (figure 10). Concerning the CH, this last one is over 91%, it can reach 100% for most of the scenarios, and it is between 79% and 89% for three recordings. However, for the 7th recording, the system presents a total failure (figure 11).

- In figures 13 and 14, we present the mean values of the GC and CH scores, calculated from the 10 scenarios containing 2 speakers (in light gray), from the 4 scenarios containing three speakers without considering the 7th scenario (in dark gray), and those of all the scenarios without considering the 7th scenario (in black). We remark that the GC and CH scores for the clustering of 2 speakers (GC equal to 95% and CH is about 96%) are better than those obtained with 3 speakers (GC equal to 86% and CH is about 90%), as shown in figures 12 and 13 respectively, which means that these scores decrease when the

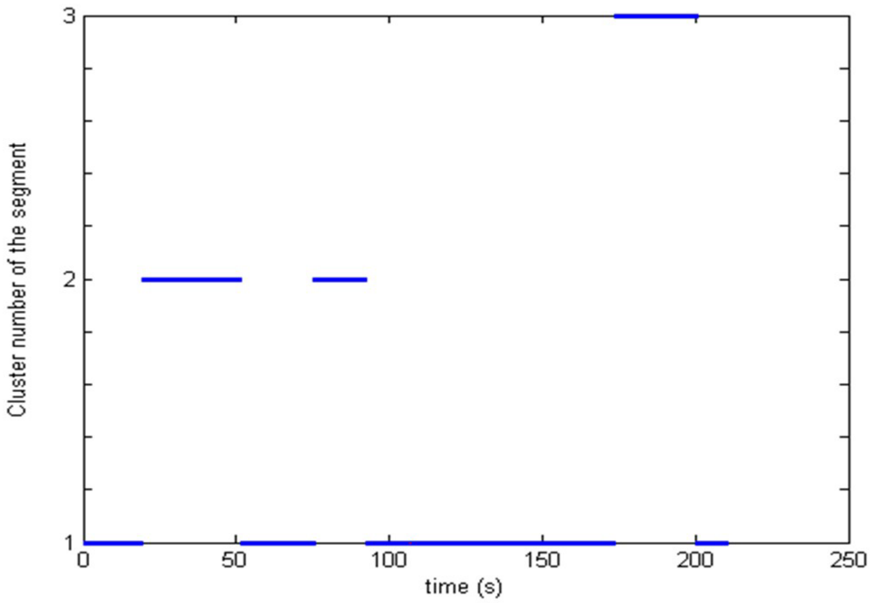

Figure 9. The new clusters found after the clustering technique: 3 different speakers have been found in the audio stream. The first speaker (cluster) is denoted by 1, the second by 2 and the third by 3.

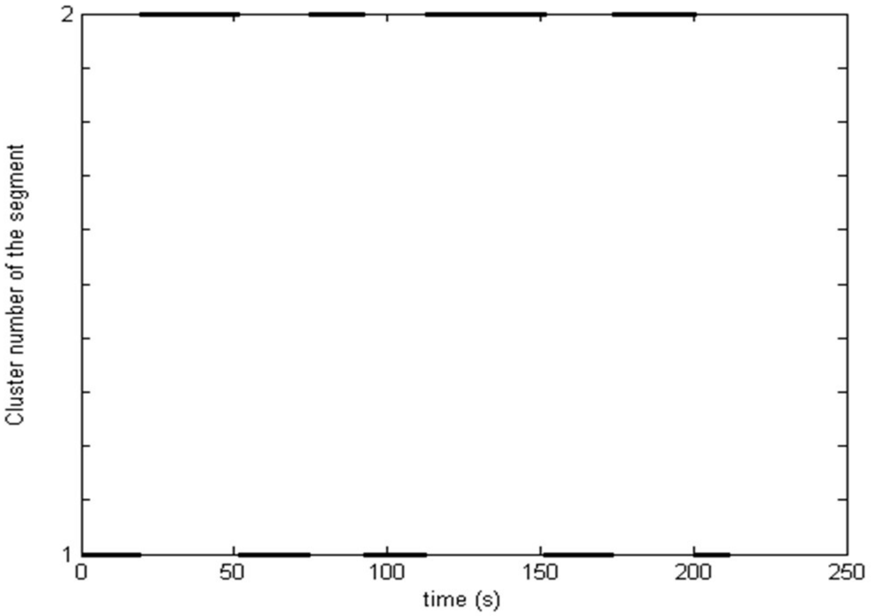

Figure 10. The 2 clusters of reference: each cluster represents a real speaker. In the indexing file of reference, two speakers (clusters) are present, denoted here by speaker 1 and speaker 2.

number of speakers increases.
However in the overall the average GC is about 93% and the average CH is about 95% for all the scenarios (except the 7th scenario), which represents an encouraging result.

- The last figure (figure 15), gives the number of clusters (speakers) in each scenario obtained after the clustering process. We can notice that for the 10 scenarios containing 2 speakers (from scenario 1 to 6 and from scenario 10 to 14), the clustering algorithm

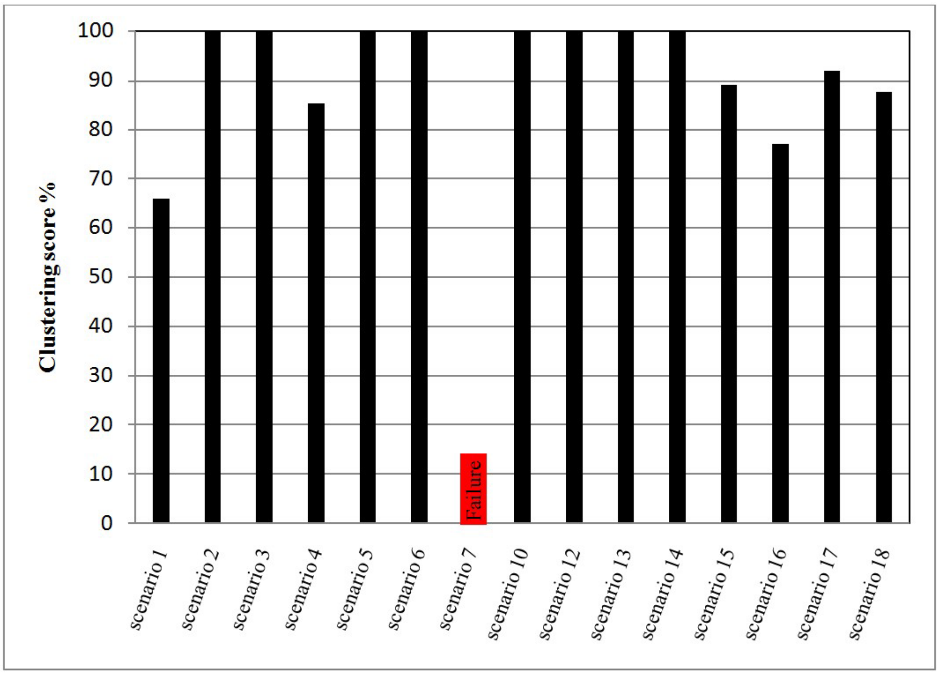

Figure 11. Score of good clustering for each scenario. We notice that the clustering is performed successfully except for the 7th scenario, which presents a failure.

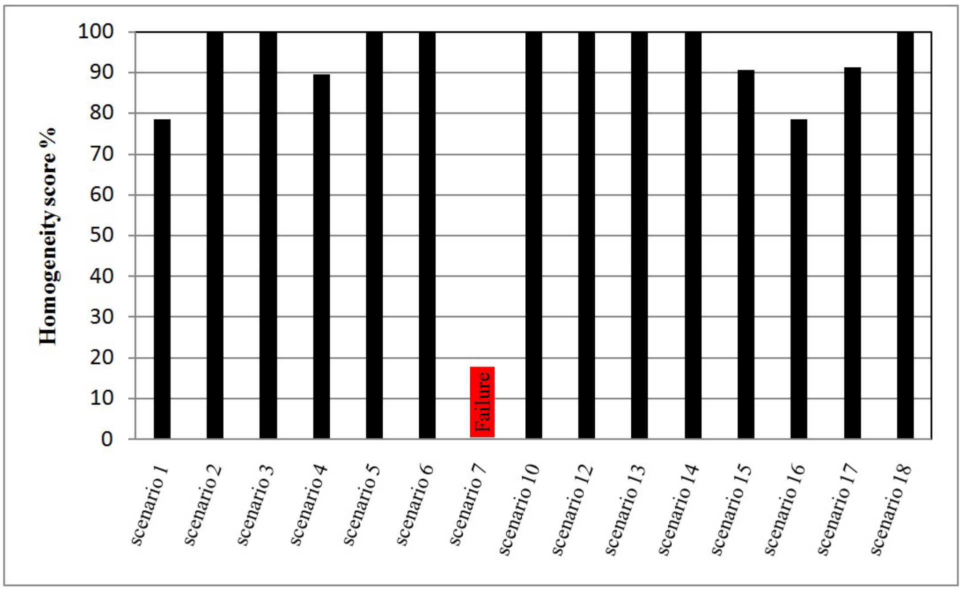

Figure 12. Score of cluster homogeneity for each scenario. We notice that the clustering is performed successfully except for the 7th scenario, which presents a failure.

manages to detect successfully the real number of speakers in 9 scenarios over 10. In the case of the first scenario, one more cluster is detected (3 clusters instead of 2). Concerning the scenarios with three speakers, the algorithm detects 3 clusters in 3 scenarios and 6 clusters in one scenario. However in the case of the 7th scenario, the algorithm detects 12 clusters while the real number of clusters is only 3, which represents a failure of the system. The cause of this failure is explained in the next paragraph.

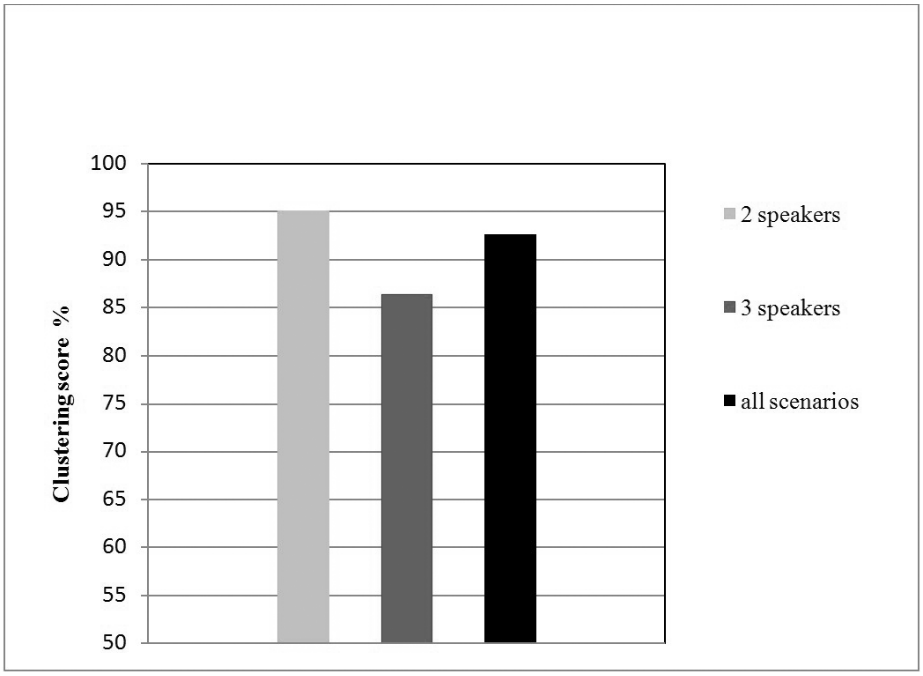

Figure 13. Score of good clustering for the different scenarios. We can notice that the score got with 2 speakers is better than that obtained with 3 speakers.

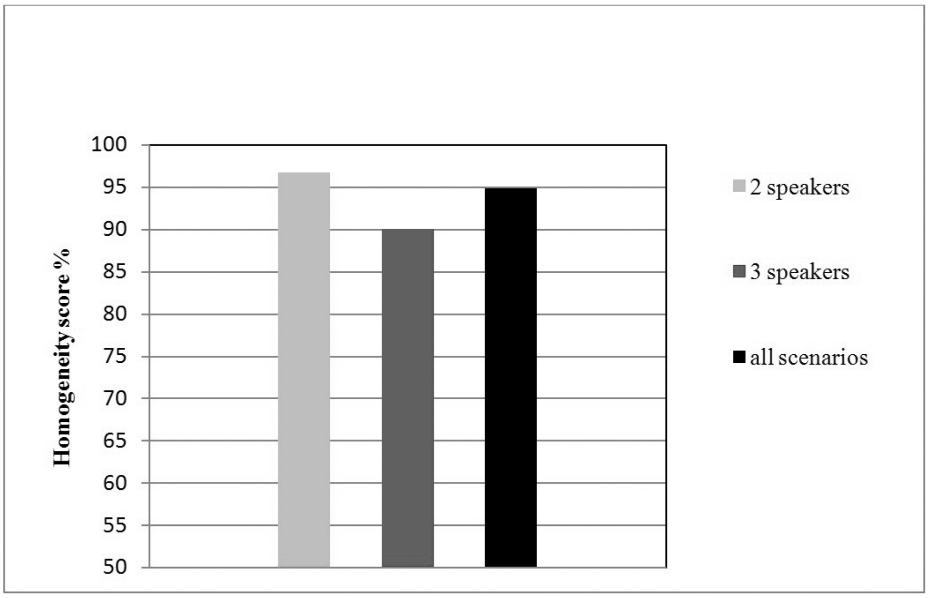

Figure 14. Score of cluster homogeneity for the different scenarios. We can notice that the score got with 2 speakers is better than that obtained with 3 speakers.

Strengths and weaknesses of the method. Trying to response to the question: *Why the clustering algorithm presents a failure for certain scenarios?*, we have represented the duration of the shortest speech segment in each scenario in table 3 below.
According to this table, we can deduce that the clustering algorithm using the mono-gaussian measure (μ_G) cannot give good performances if the duration of the speech segments is less than 3s.

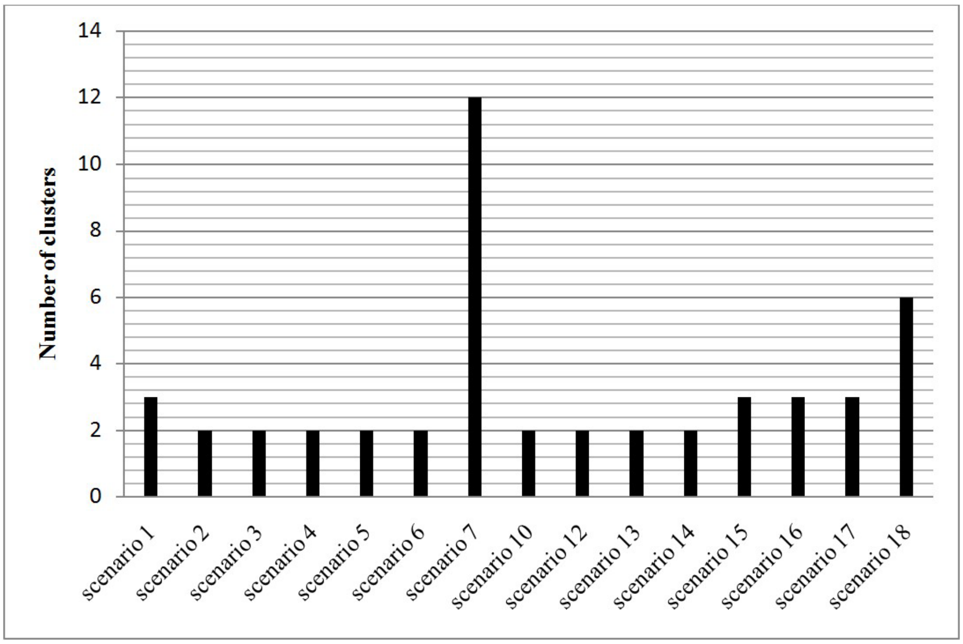

Figure 15. Estimated number of clusters in each scenario. Note that the maximum number of speakers, in reality, is 3 speakers: some of the scenarios have 2 speakers (clusters) and the others have 3 speakers (clusters).

This is really the case in the 7th scenario, where the duration of some segments do not exceed 2s in several cases of this audio stream. This situation causes the failure of the similarity measure to gather such segments and the result is then the creation of additional clusters (12 clusters instead of 3 clusters in scenario 7). In another situation, like in scenario 1, the error of clustering is due to the resemblance of the speech features between the two speakers of the scenario, which causes confusion between their speech segments. Therefore, in the overall, we can state that this method gives good performances when the duration of the homogeneous speech segment exceeds 4s in the audio recording.

Table 3. Duration of the shortest homogeneous segment in each segment. This table is displayed in order to show the effect of the short segments on the failure of the system.

Scenario	Duration of the shortest segment in second
scenarios 1, 2, 3, 4, 6, 10, 12, 13, 14, 15 and 16	between 8s and 22s
scenario 5, 17 and 18	between 3s and 5s
scenario 7	2s *(failure)*

6. **Conclusion.** The application of our investigation is the speaker clustering of audio documents related to meetings that are recorded by the means of two cardioids microphones (stereo audio documents). This task is made in a purpose of gathering the different homogeneous segments with regards to the speakers present in the considered document. Although, most existing clustering techniques are based on hierarchical clustering as agglomerative schemes, however, such systems present two problems: finding the stopping criterion and choosing the threshold of clustering decision. In this research work, we have tried to solve the first problem by proposing a sequential clustering approach based on

second order statistical measures that are used in order to gather the similar homogeneous segments correctly. Our sequential clustering approach uses a mono-gaussian statistical measure, called μ_G , which is able to assess the degree of similarity between the different homogeneous segments (even if they have different lengths). Experiments are done on a stereophonic database containing 15 scenarios and the corresponding results can be summarized by the obtained scores of good clustering (GC) and scores of cluster homogeneity (CH), as follows:

- score of 95.11% of good clustering, in case of scenarios containing 2 speakers;
- score of 86.36% of good clustering, in case of scenarios containing 3 speakers;
- score of 92.61% of good clustering for the whole scenarios;
- score of cluster homogeneity of 96.8%, in case of scenarios containing 2 speakers;
- score of cluster homogeneity of 90.06%, in case of scenarios containing 3 speakers;
- score of cluster homogeneity of 94.87% for the whole scenarios.

In the overall, we can notice that the results are interesting and very promising if the durations of the homogeneous speech segments contained in the audio file exceed 4s. However, when the audio recording contains several speech segments that are shorter than 3s, the system presents a failure. Furthermore, the simplicity of the proposed sequential clustering shows that the implemented algorithm can be interesting for the task of speaker diarization of meeting recordings such as debates, interviews or multiconferences.

In the future, we will try to use competitive clustering methods using other stereophonic techniques like the differential energy based technique which seems to be very appropriate for the case of stereo audio streams.

Acknowledgments. We would like to thank all the persons who contributed, directly or indirectly, to our research achievements. We wish also to thank the reviewers and the editors for their useful comments.

REFERENCES

[1] S. Meignier, Indexation en Locuteurs de Documents Sonores : Segmentation d'un Document et Appariement d'une Collection, Ph. D Thesis, Laboratoire Informatique d'Avignon (LIA), Universit d'Avignon et des Pays de Vaucluse, Avignon ,France, 2002.

[2] S. Ouamour, M. Guerti, and H. Sayoud, Speaker based segmentation on broadcast news -On the use of ISI technique, *Proc. of ISCA Tutorial and Research Workshop on Experimental Linguistics*, Athens, Greece, pp. 28-30, August 2006.

[3] S. Ouamour, and H. Sayoud, A new approach for speaker change detection using a fusion of different classifiers and a new relative characteristic, *The Mediterranean Journal of Computers and Networks (MedJCN)*, vol. 5, no 3, pp. 104-113, 2009.

[4] J. Ajmera, and C. Wooters, A robust speaker clustering algorithm, *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, US Virgin Islands, USA, 2003.

[5] D. A. Reynolds, and P. Torres-Carrasquillo, *The MIT Lincoln Laboratories RT-04F diarization systems: Applications to broadcast audio and telephone conversations*, Rich Transcription Workshop (RTW' 04), Palisades, NY, Fall 2004.

[6] S. Tranter, and D. Reynolds, *Speaker diarisation for broadcast news*, *Proc. of ISCA Odyssey 2004 Workshop on speaker and language recognition*, Toledo, June 2004.

[7] A. M. Xavier, *Robust Speaker Diarization for meetings*, Ph. D Thesis, Speech Processing Group Department of Signal Theory and Communications Universitat Politecnica de Catalunya Barcelona , Espagne, October 2006.

[8] S. Ouamour, *Indexation Automatique des Documents Audio en vue d'une Classification par Locuteurs-Application l'Archivage des missions TV et Radio*, Ph. D thesis, USTHB University, October 2009.

[9] M. Xavier, *Robust Speaker Diarization for meetings*, Ph. D Thesis, Speech Processing Group Department of Signal Theory and Communications Universitat Politecnica de Catalunya Barcelona , Espagne, October 2006.

[10] K. Mori, and S. Nakagawa, Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, vol. 1, pp. 413-416, 2001

[11] S. Nakagawa, and H. Suzuki, A new speech recognition method based on VQ-distortion and HMM, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Minneapolis, USA, pp. 676-679, 1993.

[12] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006.

[13] Wang Wei, Lv Ping, Zhao Qingwei, and Yan Yonghong, *A Decision-Tree-Based Online Speaker Clustering*, *Proc. of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I table of contents. Girona*, Spain, pp. 555-562, 2007.

[14] C. Barras, X. Zhu, S. Meignier, and J. L. Gauvain, *Improving speaker diarization*, Fall Rich Transcription Workshop (RT04), Palisades, NY, 2004.

[15] M. H. Siu, G. Yu, and H. Gish, An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, USA, vol. 2, pp. 189-192, 1992.

[16] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, *Speaker segmentation and clustering in meetings*, NIST 2004 Spring Rich Transcrition Evaluation Workshop, Montreal, Canada, 2004.

[17] H. Gish, M. H. Siu, and R. Rohlicek, Segregation of speakers for speech recognition and speaker identification, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, vol. 2, pp. 873-876, 1991.

[18] H. Jin, F. Kubala, and R. Schwartz, *Automatic speaker clustering*, DARPA Speech Recognition workshop, Chantilly, USA, 1997

[19] A. Solomonov, A. Mielke, M. Schmidt, and H. Gish, Clustering speakers by their voices, *Proc. of IEEE International Conference on Acoustics*, Speech and Signal Processing, Seattle, USA, vol. 2, pp. 757-760, 1998

[20] S. S. Chen, and P. Gopalakrishnan, Clustering via the bayesian information criterion with applications in speech recognition, *Proc. of IEEE International Conference on Acoustics*, Speech and Signal Processing, Seattle, USA, vol. 2, pp. 645-648, 1998

[21] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. J. McLaughlin, and M. A. Zixxman, Blind clustering of speech utterances based on speaker and language characteristics, *Proc. of International Conference on Speech and Language Processing*, Sidney, Australia, 1998.

[22] S. Johnson, and P. Woodland, Speaker clustering using direct maximization of the MLLRadapted likelihood, *Proc. of International Conference on Speech and Language Processing*, vol. 5, pp. 1775-1779, 1998.

[23] J. Ajmera, H. Bourlard, and I. Lapidot, *Improved unknown-multiple speaker clustering using HMM*, Technical report, IDIAP, 2002.

[24] Y. Moh, P. Nguyen, and J. C. Junqua, Towards domain independent speaker clustering, *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, 2003.

[25] G. Schwarz, *Estimating the dimension of a model*, The Annals of Statistics 6, pp. 461-464, 1978.

[26] B. Zhou, and J. H. Hansen, Unsupervised audio stream segmentation and clustering via the bayesian information criterion, *Proc. of International Conference on Speech and Language Processing*, Beijing, China, vol. 3, pp. 714-717, 2000.

[27] Eugene Chin Koh, Hanwu Sun, Tin Lay Nwe, Trung Hieu Nguyen, Bin Ma, Eng-Siong Chng, Haizhou Li and Susanto Rahardja, *Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information*: The I2R-NTU Submission for the NIST RT 2007 Evaluation In Multimodal Technologies for Perception of Humans, Springer Berlin / Heidelberg, pp. 484-496, 2008.

[28] Haipeng Wang, Xiang Zhang, Hongbin Suo, Qingwei Zhao, and Yonghong Yan, A Novel Fuzzy-Based Automatic Speaker Clustering Algorithm, *Proc. of the 6th International Symposium on Neural Networks: Advances in Neural Networks- Wuhan*, China Section: Clustering and Classification, pp. 639-646, 2009.

[29] J. Zibert, and F. Mihelic, Fusion of Acoustic and Prosodic Features for Speaker Clustering, *Proc. of the 12th International Conference on Text, Speech and Dialogue*, Pilsen, Czech Republic, Section, pp. 210-217, 2009.

[30] M. Ben, M. Betser, F. Bimbot, and G. Gravier, Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs, *Proc. of International Conference on Spoken Language Processing*, (ICSLP4), Jeju Islands, South Corea, October 2004.

[31] F. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J. F. Bonastre, Pre-processing techniques and speaker diarization on multiple microphone meetings, *Proc. of NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation (RT'05S)*, Lecture Notes in Computer Sciences (LNCS), Springer, July 2005.

[32] D. Vijayasenan, F. Valente, and H. Bourlard, Combination of agglomerative and squential clustering for speaker diarization, *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[33] F. Bimbot, I. Magrin-Chagnolleau , and L. Mathan, Second-Order Statistical Measures for text-independent Broadcaster Identification, Speech Communication, vol. 17, no. 1-2, pp. 177-192, August 1995.

[34] H. Sayoud, et al., *'ASTRA' An Automatic Speaker Tracking System based on SOSM measures and an Interlaced Indexation*, Acta Acustica, vol. 89, no. 4, pp. 702-710, 2003.

[35] K. Schutte, and J. Glass, *Features and Classifiers for Robust Automatic Speech Recognition*, Research Abstracts-2007, Research Project. MIT CSAIL Publications and digital archives, 2007.

[36] S. Ouamour, H. Sayoud, and M. Guerti, Optimal Spectral Resolution in Speaker Authentication, Application in noisy environment and Telephony, *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, pp. 36-47, 2009.