Speaker Clustering of Stereophonic Speech Signal Using Spatial and Sequential Gathering

S. Ouamour and H. Sayoud

***Abstract*—In this research work we are interested in gathering all the different speech segments, found after a speaker diarization process of meeting recordings, into homogeneous clusters, where each cluster contains only one speaker. Our application concerns debates or multi-conferences of several speakers who are located at fixed positions in a meeting-room. For that purpose, the stereo speech signals of the speakers are collected by two cardioid microphones, which are placed inside the meeting-room.**
In this investigation, two techniques of clustering have been implemented: the Energy Differential based Spatial Clustering (EDSC) and the Mono-Gaussian based Sequential Clustering (MGSC).
Experiments of speaker clustering are done on a stereophonic database called DB15, which is composed of 15 scenarios of about 3.5 mn each. Every scenario contains the speech of two or three speakers who are speaking sequentially in the meeting room. Experimental results show the large superiority of the energy differential technique in term of precision and speed over the statistical sequential clustering, especially for short speech segments.

***Index Terms*—Artificial intelligence, Speech processing, Speaker clustering, Spatial clustering algorithms, Automatic speaker localization.**

I. Introduction

THE task of speaker clustering requires that we correctly identify how many real speakers participate in the audio recording, by gathering the similar homogeneous segments into classes of speakers [1] in order to obtain, at the end of the process, a number of clusters equal to the real number of speakers who are present in the audio stream. Each cluster contains the global intervention of a speaker participating in the meeting.

The main application, which is focused in this paper, is the speaker clustering of meeting recordings. In such applications, usually more than one microphone is available in the meeting-room [2]. To achieve this task, we propose two techniques: the first one is the Energy Differential based Spatial Clustering (EDSC) and the second method (called MGSC) uses a Sequential Clustering algorithm based on Mono-Gaussian measure [3].

Manuscript received March 09, 2011; revised April 07, 2011.

S. Ouamour is an Associate Professor at the USTHB University of Algiers, Algeria. www.usthb.dz. Email: siham.ouamour@gmail.com.

H. Sayoud is a Full Professor at the USTHB University of Algiers, Algeria. www.usthb.dz. Email: halim.sayoud@gmail.com.

Address: USTHB University, Electronics and Computer Enineering Institute. BP 32 Al-Alia, Bab-Ezzouar, Alger, Algeria.

These two algorithms are evaluated on a stereo database: DB15, which contains 15 meeting recordings (scenarios). Results show that the implemented techniques seem to be promising for the task of speaker clustering.

II. Speaker Clustering Algorithms

II.1. Energy Differential based Spatial Clustering (EDSC)

By assuming that there are, for example, three speakers in a meeting-room, and that the speakers have fixed positions (ie. they are sitting), we can demonstrate that every speaker has a specific energy differential between the two signals collected by two distant microphones placed inside the meeting-room.

In this approach, if we consider that the position of the speakers does not change over the time, we can state that in each homogeneous speech segment, we should retrieve the same energy differential value (same spatial position). Consequently, if two speech segments correspond to the same spatial position, then they should belong to the same speaker (ie. to the same cluster) and can be gathered together to form a unique cluster: this is the principle of the EDSC clustering method.

The algorithm of the EDSC is given as follows:

The first order energy is computed in every speech segment of 1 s for the 2 microphones (signal *x* of the right microphone and signal *y* of the left microphone), with the following manner:

$$E_x = \sum_{1=0}^{N} |x_i| \quad (1)$$

$$E_y = \sum_{1=0}^{N} |y_i| \quad (2)$$

Then, the energy differential is computed as follows:

$$DExy = \log(Ex/Ey) = \log(Ex) - \log(Ey) \quad (3)$$

So it is easy, now, to estimate the relative position of the speaker and then the cluster of the homogeneous segment with regards to the microphones positions. For instance, it is easy to deduce if the speaker is in the right side, left side or in the middle by the following scheme:

Computation of the differential energy;

If $DExy < Threshold_{min}$

then Speaker is in the left

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

If DExy > Threshold$_{max}$

then Speaker is in the right

If Threshold$_{min}$ < DExy < Threshold$_{max}$

then speaker is in the middle

Threshold$_{min}$ and Threshold$_{max}$ are tuned experimentally.

II.2. Mono-Gaussian based Sequential Clustering (MGSC)

A. 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 [4].

Let $\{x_t\}_{1\le t\le M}$ 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 $\bar{x}$ and the covariance matrix *X*:

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \tag{4}$$

and

$$X = \frac{1}{M}\sum_{t=1}^{M}(x_t - \bar{x})(x_t - \bar{x})^T \tag{5}$$

Similarly, for a speech signal uttered by speaker $\boldsymbol{y}$, a sequence of N vectors $\{y_t\}_{1\le t\le N}$ 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 y_t uttered by speaker $\boldsymbol{y}$ is:

$$G(y_t/\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det X)^{1/2}} e^{-(1/2)(y_t-\bar{x})^T X^{-1}(y_t-\bar{x})} \tag{6}$$

"det" represents the determinant.

If we assume that all vectors y_t are independent observations, the average log-likelihood of $\{y_t\}_{1\le t\le N}$ can be written as:

$$\bar{G}_{\boldsymbol{x}}(y_1^N) = \frac{1}{N}\log G(y_1 \ldots y_N|\boldsymbol{x}) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t|\boldsymbol{x}) \tag{7}$$

by replacing $y_t - \bar{x}$ by $y_t - \bar{y} + \bar{y} - \bar{x}$ and using the property

$$\frac{1}{N}\sum_{t=1}^{N}\left((y_t-\bar{y})^T X^{-1}(y_t-\bar{y})\right) = tr\left(YX^{-1}\right) \tag{8}$$

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

$$\frac{2}{P}\bar{G}_{\boldsymbol{x}}(y_1^N) + \log 2\pi + \frac{1}{P}\log(\det Y) + 1 = \frac{1}{P}\left[\log\left(\frac{\det(Y)}{\det(X)}\right) - tr(YX^{-1}) - (\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right] + 1 \tag{9}$$

The gaussian likelihood measure μ_G is defined by:

$$\mu_G(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{P}\left[-\log\left(\frac{\det(Y)}{\det(X)}\right) + tr(YX^{-1}) + (\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right] - 1 \tag{10}$$

We have:

$$Arg\max_{\boldsymbol{x}}\bar{G}_{\boldsymbol{x}}(y_1^N) = Arg\min_{\boldsymbol{x}}\mu_G(\boldsymbol{x},\boldsymbol{y}) \tag{11}$$

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 $\mu_{G\beta}$ statistical measure is given as follows [5]:

$$\mu_{G\beta}(\boldsymbol{x},\boldsymbol{y}) = (M.\mu_G(\boldsymbol{x},\boldsymbol{y}) + N.\mu_G(\boldsymbol{x},\boldsymbol{y}))/(M+N) \tag{12}$$

where:

$$\mu_G(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{p}\left[tr(YX^{-1}) - \log\left(\frac{\det Y}{\det X}\right) + (\bar{y}-\bar{x})^T X^{-1}(\bar{y}-\bar{x})\right] - 1 \tag{13}$$

B. 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). 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 [6] (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 [7].

C. Sequential Clustering Algorithm

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 [8], [9], 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 $\mu_{G\beta}$ measure, which allows assessing the degree of similarity between 2 homogeneous segments of different lengths [4].

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 [9]. This process continues until all the homogeneous segments are processed chronologically, one after the other (figure 1).

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

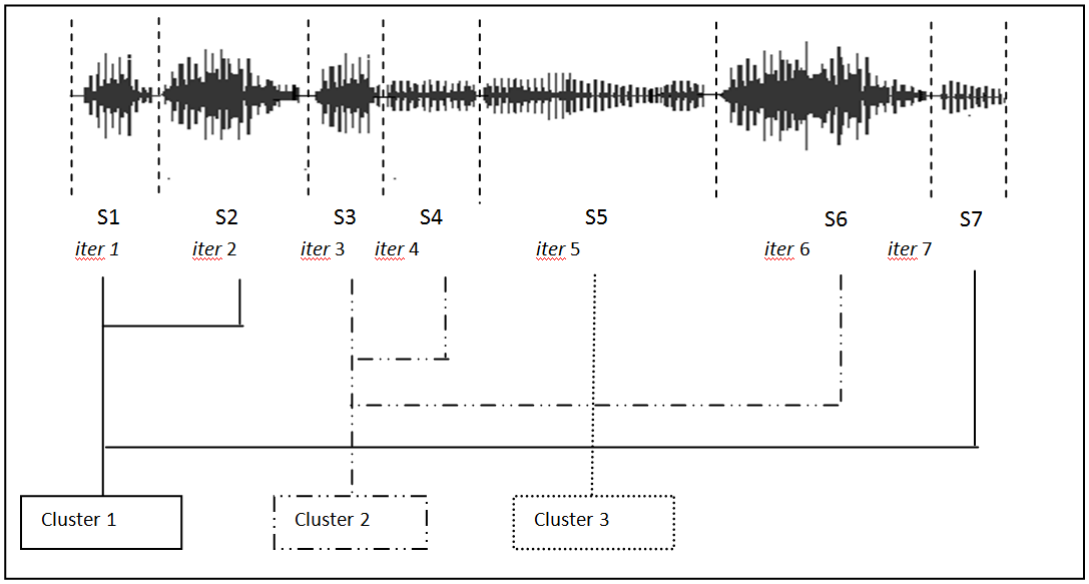

Fig. 1: Principle of the sequential clustering.
S represents a homogeneous segment and *iter* represents an iteration.

III. The Stereophonic Speech Database

The sequential clustering algorithm is evaluated on a stereo database called DB15. The audio database includes 15 meeting recordings divided into 10 conversations between 2 speakers and 5 conversations between three different speakers speaking alternatively in a natural manner. The speech recording is acquired at 16kHz and in a stereo form by two cardioid microphones placed in opposition and separated by a fixed distance. The duration of each scenario is between 3 mn and 4 mn, and the total speech duration is about 40 mn. The speakers are seated at one of the 3 fixed positions of the meeting room: Left, Middle or Right (figures 2). The distance between the 2 microphones is 1m and the global number of speakers used to construct these scenarios is six (4 females and 2 males).

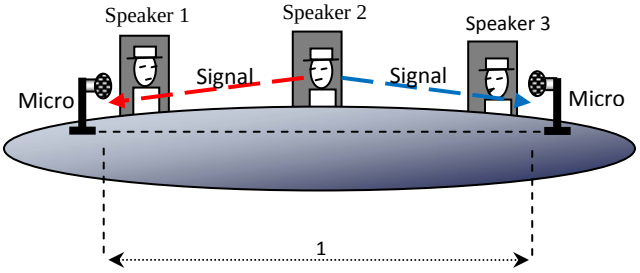

Fig. 2. Disposition of the speakers in the meeting-room: the stereo speech signal is recorded by 2 cardioid microphones.

IV. Experimental Results

In order to evaluate the results given by the two techniques, two types of scores have been proposed:

- Score of Good Clustering (GC) defined by the ratio between the number of homogeneous segments which are well gathered and the total number of homogeneous segments, given by the formula below:

$$GC = \frac{\text{number of segments which are well gathered}}{\text{total number of segments}} * 100 \quad (14)$$

- Score of Cluster Homogeneity (CH) represents the mean of all the cluster homogeneities of the scenario (eg. in case of a scenario with 3 clusters, we will have three cluster homogeneities: CH1, CH2 and CH3).
  The cluster homogeneity of each cluster i (CHi) is defined by the ratio between the number of clusters that belongs really to this cluster and the number of all the segments gathered in that cluster (real segments plus false alarms). Thus the corresponding formulas are given as follows:

$$CHi = \frac{\text{number of segments belonging to cluster } i}{\text{number of all the segments of cluster } i} * 100 \quad (15)$$

$$CH = \frac{1}{N}\sum_{i=0}^{N} CHi \quad (16)$$

with N representing the number of clusters in the scenario.

The different scores of clustering and homogeneity, obtained in each experiment, are given in figures 3 and 4, we can deduce the following results:

- We can notice that, for the sequential algorithm (presented in green), the GC score reaches 100% for 8 scenarios, it is between 85% and 91% for 4 scenarios and between 66% and 78% for three scenarios (figure 3). Concerning the CH score, this one is over 91% and reaches 100% for most of the scenarios and it is between 79% and 89% for three scenarios. However, for the 7th scenario, the system falls down (figure 4). In the case of the EDSC (represented in red), the GC and CH scores reach the rate of 100% for 13 scenarios and are over 91% for two scenarios (figures 3 and 4).

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

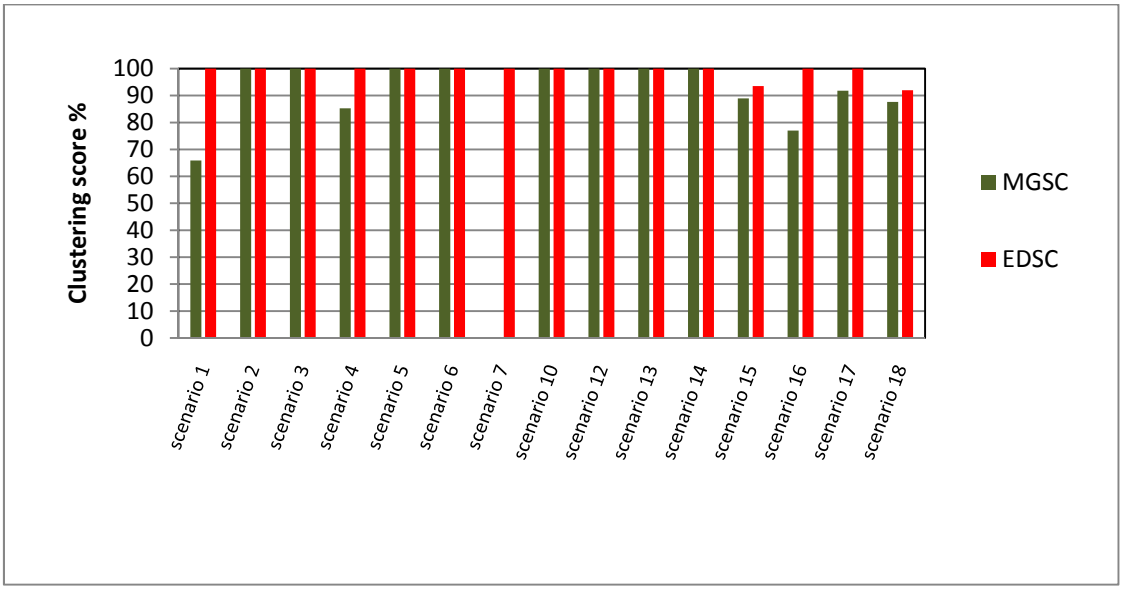

Fig. 3. Scores of Good Clustering (GC) of each scenario.

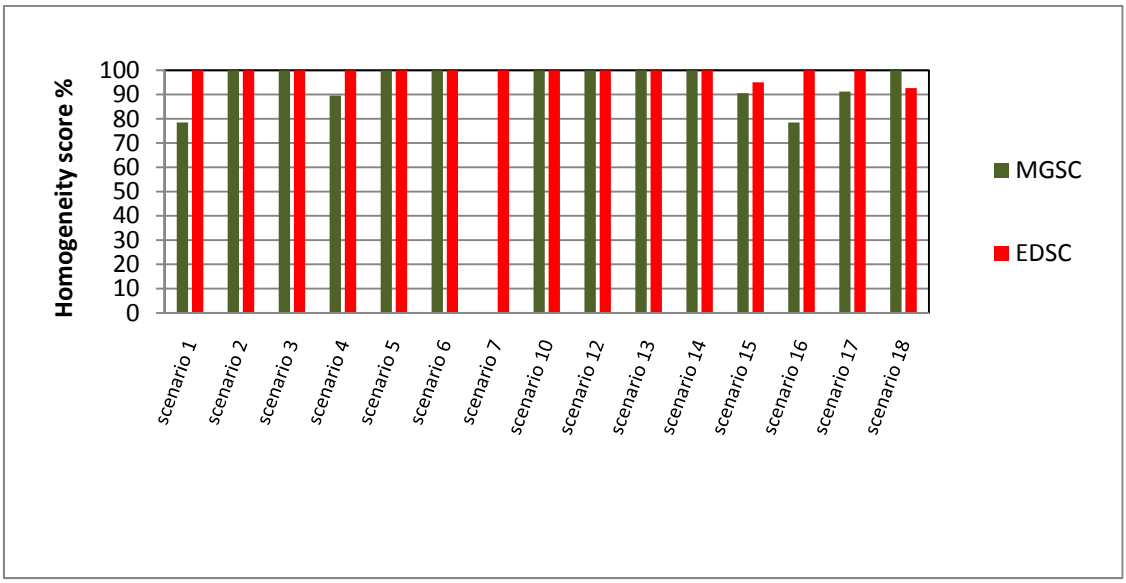

Fig. 4. Scores of Cluster Homogeneity (CH) of each scenario.

- We also notice that the EDSC gives a GC score and CH score of 100% for the 7th scenario, whichcontains several short homogeneous segments (less than 3s), whereas the MGSC method presents a total failure for the same scenario.
- For of all the scenarios (except the 7th scenario), the average GC is about 93% and the average CH is about 95% for the MGSC clustering, whereas their corresponding scores obtained by the EDSC clustering, are about 99% for both the GC and CH scores, which represents an interesting result.

V. Conclusion

Speaker clustering is the task of grouping a set of speech utterances into speaker-specific classes. In this framework, we proposed two techniques of speaker clustering. The first technique uses an algorithm based on the energy differential which we called Energy Differential based Spatial Clustering (EDSC), and the second method uses a Sequential Clustering algorithm associated to the Mono-Gaussian measure (MGSC).

Experiments are done on a stereophonic database; the corresponding results can be summarized by the obtained scores of good clustering GC and scores of cluster homogeneity CH, as follows:

• GC score of 92.61% for the MGSC clustering and 98.97 % for the EDSC technique, for all the scenarios;

• CH score of 94.87 % for the MGSC clustering and 99.12 % for the EDSC technique, for all the scenarios.

In the overall, we can notice, on one hand, that the results are interesting in case of the technique based on the sequential algorithm if the duration of the homogeneous

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

speech segments contained in the audio file exceeds 4s. However, when the audio recording contains several speech segments that are shorter than 3s, the system presents a failure. On the other hand, very good scores are obtained by the proposed technique (EDSC), which gives quite better performances than the MGSC clustering using the mono-gaussian measure in all the experiments. Especially when the scenarios contain short homogeneous segments, the EDSC algorithm seems to be not affected by the speech utterances durations at all.

Finally, we can deduce from this investigation that the EDSC technique seems to be very promising for the task of speaker clustering in case of meeting indexing because in addition to its simplicity, it presents very good results even in the case of short segments, which represents an important result because most techniques present a failure in such situations.

As perspectives, we propose to use the proposed clustering algorithm associated with other discriminative classifiers as the Support Vector Machines or Neural Networks, which usually present high discriminative capacities. The objective would be to make a fusion between those clustering systems in order to further enhance the clustering accuracy.

REFERENCES

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

[2] A. M. Xavier, "Robust Speaker Diarization for meetings", PhD Thesis, Speech Processing Group Department of Signal Theory and Communications Universitat Politecnica de Catalunya Barcelona (Espagne), October 2006.

[3] S. Ouamour, "Indexation Automatique des Documents Audio en vue d'une Classification par Locuteurs -Application à l'Archivage des

[4] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan, "Second-Order Statistical Measures for text-independent Broadcaster Identification", Speech Communication, volume 17, number 1-2, August 1995, pp. 177-192.

[5] H. Sayoud et al., "'ASTRA' An Automatic Speaker Tracking System based on SOSM measures and an Interlaced Indexation", Acta Acustica, volume 89, number 4, 2003, pp. 702-710.

[6] 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.

[7] 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), April-June 2009, pp. 36-47.

[8] S. Meignier, "Indexation en locuteurs de documents sonores : Segmentation d'un document et Appariement d'une collection", Phd Thesis, Laboratoire Informatique d'Avignon (LIA), Université d'Avignon et des Pays de Vaucluse, Avignon (France), 2002.

[9] P. Delacourt, "La segmentation et le regroupement par locuteurs pour l'indexation de documents audio", Thèse de Doctorat, Institut Eurecom, Nice (France), 2000.

ISBN: 978-988-19251-4-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)