Speaker Based Clustering Using the Differential Energy

S. Ouamour
USTHB University, Electronics and Computer Enineering Institute.
Alger, Algeria
siham.ouamour@gmail.com

H. Sayoud
USTHB University, Electronics and Computer Enineering Institute.
Alger, Algeria
halim.sayoud@uni.de

***Abstract*— A new approach of speaker clustering is presented and discussed in this paper. The main technique consists in grouping all the homogeneous speech segments obtained at the end of the segmentation process, by using the spatial information provided by the stereophonic speech. The proposed system is suitable for debates or multi-conferences for which the speakers are located at fixed positions. The new method uses the differential energy of the two stereophonic signals collected by two cardioid microphones, in order to group all the speech segments that are uttered by the same speaker. The total number of clusters obtained at the end should be equal to the real number of speakers present in the meeting room and each cluster should contain the global intervention of only one speaker.**

The new proposed approach (which we called Energy Differential based Spatial Clustering or EDSC) has been experimented comparatively with a classic statistical approach called "Mono-Gaussian Sequential Clustering".

Experiments of speaker clustering are done on a stereophonic speech corpus called DB15, composed of 15 stereophonic scenarios of about 3.5 minutes each. Every scenario corresponds to a free discussion between several speakers seated at fixed positions in the meeting room.

Results show the strong performances of the new approach in terms of precision and speed, especially for short speech segments.

Keywords: Speaker based clustering; Spatial clustering; Spatial speaker localization; Speaker recognition; Stereophonic speech.

I. Introduction

Voice remains one of the most important means used by human being 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.

In most cases, we deal with diarization applications without any prior knowledge about the speakers' identities or their number in the audio document. This situation makes the task of segmentation and clustering more difficult.

The first goal of a speaker diarization system is to divide an audio stream into a sequence of speaker-homogeneous regions. Thus, the output of such a system provides the answer to the following question: Who spokes and when? [1, 2, 3].

Thereafter, another important refining task is required: namely, the speaker clustering task, which consists in gathering the similar homogeneous segments into classes of speakers [4]. This last process will provide a number of clusters equal to the real number of speakers who are present in the audio stream.

Several researchers have proposed many techniques for solving the speaker clustering problem. Most of these methods involve hierarchical clustering of the data into clusters where the optimal number of speakers is unknown in advance. 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 criteria [5]. These stopping criteria are often estimated empirically and adapted to the speech database.

According to Reynolds and Torres-Carrasquillo [6], there are three main applications in speaker diarization, namely: Broadcast news indexing [7], phone conversation indexing and meeting indexing.

The application focused in this paper concerns the speaker clustering of meeting recordings. In such applications, usually, more than one microphone is available in the meeting-room.

978-1-4799-7100-8/14/$31.00 ©2014 IEEE

To achieve this task, we propose two techniques: the first one is a new spatial clustering algorithm that we called Energy Differential based Spatial Clustering or EDSC. This method is based on the calculation of the logarithmic energy difference between the two speech signals collected by two microphones [8]. The second method (called MGSC) uses a simple Sequential Clustering algorithm based on a Mono-Gaussian distance measure [9]. This distance has been chosen because it offers the possibility to measure the similarity between speech segments with different durations [10, 11]. Moreover, this distance is simple to implement, does not require any training and has the advantage to present good performances in speaker identification.

These two techniques are evaluated in the same conditions and applied on a speech stereophonic database called DB15, which is composed of 15 multi-speakers scenarios.

II. Methods of Speaker Clustering

In this section, we describe the theoretical principle of the two clustering algorithms, namely: the Energy Differential based Spatial Clustering (EDSC) and the Sequential Clustering based on Mono-Gaussian Measures (MGSC). These two techniques were evaluated on the same corpus (DB15).

A. The new proposed clustering technique: Energy Differential based Spatial Clustering (EDSC)

We have proposed a new clustering algorithm based on the spatial localization of the speakers in the meeting-room, which we called Energy Differential based Spatial Clustering (EDSC). This algorithm investigates the energy differential between the speech signals recorded by two microphones (left and right). Its principle is detailed in the following paragraphs.

A.1. Algorithm of the Energy Differential based Method

The first order energy is computed in every speech segment of 0.25 s for the 2 microphones (signal *x* of the right microphone and signal *y* of the left microphone), with the following manner:

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

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

where Ex and Ey denote the corresponding energies and N represents the length of these vectors.

Then, the energy differential is computed as follows:

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

Hence, the position of the active speaker is estimated by deducing his spatial position from the two microphones (left and right) as follows:

```
Computation of the differential energy;

If  DExy < Threshold_min
            then Speaker is in the left
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.

Then, the different speech segments corresponding to a same spatial position (i.e. same speaker) are grouped together into the same cluster.

A.2. Spatial Clustering Principle

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

Consequently, the closest speaker to the right microphone will have the highest energy differential according to the formula (3), the speaker who is at equal distance from the two microphones will have a null energy differential, and so on.

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 come from 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.

B. Sequential Clustering based on Mono-Gaussian Measures

The second algorithm of clustering developed and tested on the same database, is based on sequential technique and uses the mono-gaussian measure. We called this algorithm, Sequential Clustering based on Mono-Gaussian Measures (EDSC). The sequential technique has been chosen in order to make a comparison between this last one and the spatial clustering. Concerning our motivation about the choice of the EDSC clustering: -firstly we have noticed a lack in research works involving sequential speaker clustering; secondly, our global application deals with "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. Furthermore, the new proposed technique lets us solving the problem of stopping criterion, which is needed in the case of hierarchical clustering. The mono-gaussian measure has been introduced to assess the degree of similarity between segments with different durations and to gather the speech segments belonging to the same speaker.

The details of the EDSC clustering are described in the following paragraphs.

B.1. Mono-Gaussian Measures (or second order statistical measures)

The sequential clustering 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 [12].

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 ***x***. These vectors are summarized by the mean vector $\bar{x}$ and the covariance matrix *X*:

$$\bar{x} = \frac{1}{M}\sum_{t=1}^{M} x_t \qquad (4)$$

and

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

Similarly, for a speech signal uttered by speaker ***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 ***x*** are distributed like a gaussian function, the likelihood "G" of a single vector y_t uttered by speaker ***y*** is:

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

where $\bar{y}$ denotes the mean vector and *Y* the covariance matrix of speaker ***y*** and "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:

$$\overline{G}_{\boldsymbol{x}}(y_1^N) = \frac{1}{N}\log G(y_1 ... y_N | \boldsymbol{x}) = \frac{1}{N}\sum_{t=1}^{N}\log G(y_t | \boldsymbol{x}) \qquad (7)$$

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

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

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

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

the gaussian likelihood measure μ_G is defined by:

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

We have:

$$\underset{\boldsymbol{x}}{Arg\max}\ \overline{G}_x(y_1^N) = \underset{\boldsymbol{x}}{Arg\min}\ \mu_G(\boldsymbol{x,y}) \qquad (11)$$

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 [13]:

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

where:

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

B.2. Sequential Clustering Algorithm

The principle of this clustering is to consider the first segment as a first cluster, after that, the following homogeneous segment is compared to the first cluster (containing the first segment) 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. the other homogeneous segments are processed sequentially in the same manner, the homogeneous segments is compared to existing clusters, to assess the degree of similarity between this segment and the different clusters created previously. in case of similarity with any cluster, the segment will be assigned to that cluster. Now, if the features of the processed segment are different from the features of the other clusters, a new cluster is added to the existing ones. This cluster will represent a new speaker. Thus, this process continues until all the homogeneous segments are processed chronologically, one after the other.

The algorithm of our sequential clustering is given below:

- the different homogeneous segments are represented by their instants of beginning and end, thus their numbers in the audio document;
- The new step consists in the application of the $\mu_{G\beta}$ 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. This is ensured by using a sequential process of all the segments (processed over the time):

```
if distance[segment(i), segment(j)] ≤ threshold
    then segment(i) and segment(j) come from the
    same speaker. So, belong to the same cluster;
if distance[segment(i), segment(j)] > threshold
    then segment(i) and segment(j) come from
    different speakers. S, belong to different clusters;
Redo the process by incrementing the indices.
```

- Then, a new reorganization of the different clusters is applied by gathering the segments belonging to each speaker and assigning to them new numbers, with the corresponding time of beginning and time of end.
- Finally, a comparison is done between the number of clusters (speakers) obtained at the end of the clustering process and the real number of the speakers.

III. Stereophonic Speech Database: DB15

The spatial and sequential clustering algorithms are evaluated on a stereophonic 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 with 2x16 bits, 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 (figure 1). 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.

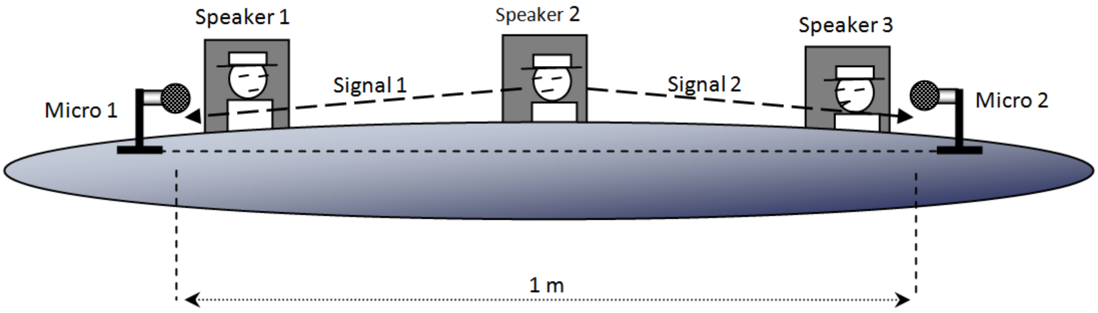

Figure 1: The two cardioid microphones: left disposition and right disposition respectively.

IV. Experiments and Results

The two clustering algorithms, namely: the Energy Differential based Spatial Clustering (EDSC) and Mono-Gaussian based Sequential Clustering (MGSC), are tested in the same conditions on DB15 database. The goal is to gather the different homogeneous segments into classes of speakers, with a maximum of accuracy.

In order to evaluate the different results given by the two techniques, a Score of Good Clustering has been proposed:

The Score of Good Clustering (GC) is defined by the ratio between the number of homogeneous segments which are well gathered and the total number of homogeneous segments of the processed scenario. This score is given by the following formula:

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

Figure 2 displays the clustering scores obtained for the 15 scenarios of the stereo DB15 database. We can notice that, for the sequential algorithm (presented in gray), 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. However, for the 7th scenario, the system falls down (figure 2). However, in the case of the EDSC (represented in black), the GC score reaches the rate of 100% for 13 scenarios and are over 91% for two scenarios.

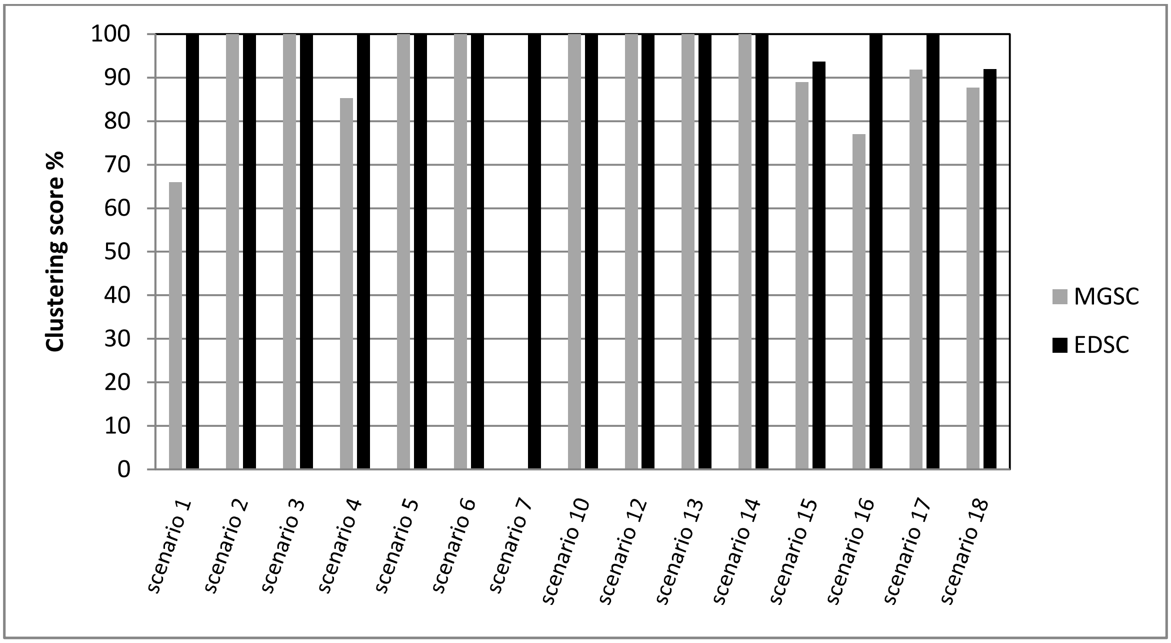

Figure 2: Scores of Good Clustering (GC).

We also notice that the EDSC gives a GC score of 100% for the 7th scenario, which contains several short homogeneous segments (less than 3s), whereas the MGSC method presents a total failure for the same scenario.

Typical results, reported from the 10 scenarios containing 2 speakers, from the 4 scenarios containing three speakers without considering the 7th scenario, and those of all the scenarios without considering the 7th scenario, have shown that the GC score for the clustering of 2 speakers (score of 95% for the MGSC clustering) is better than the one obtained with 3 speakers (GC score of 86%), which means that this score decreases when the number of speakers increases. The same remark is observed for the EDSC clustering (GC score of 100% for scenarios containing 2 speakers and decreasing to 96% for scenarios containing 3 speakers.

For all the scenarios (except the 7th scenario), the average GC is about 93% for the MGSC clustering, whereas its corresponding score obtained by the EDSC clustering, is about 99%, which represents an interesting result.

The last observation concerns the comparison between the two techniques used for the speaker clustering task, we can notice that the EDSC clustering gives quite better performances than the MGSC clustering. Especially, when the duration of the homgeneous segments is less than 3 seconds (case of the 7th scenario), the EDSC technique remains robuste (clusteing score of 100%), whereas the MGSC clustering presents a total failure in such cases.

V. Conclusion

Several clustering experiments, using the new EDSC algorithm, have been conducted comparatively with the classic statistical one (taken as reference) on the same stereophonic dataset (DB15), which consists of 15 different scenarios.

This corpus consists in meeting recordings of several speakers, who where seated at different positions in the meeting-room.

The corresponding results are summarized by the following "Good Clustering" scores (GC):

- A GC score of 95.11% for the MGSC clustering method and a score of 100 % for the proposed EDSC technique, in case of scenarios containing 2 speakers;
- A GC score of 86.36% for the MGSC clustering method and a score of 96,38 % for the proposed EDSC technique, in case of scenarios containing 3 speakers;
- A GC score of 92.61% for the MGSC clustering method and a score of 98.97 % for the proposed EDSC technique, for all the scenarios;

In the overall, we can notice that both techniques show good performances on the DB15 speech corpus when the duration of the homogeneous speech segments exceeds 4s.

However, when the audio recording contains several speech segments that are shorter than 3s, the sequential method falls down and presents many errors of clustering.

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 experiments, especially when the scenarios contain short homogeneous segments: the proposed method seems to be not affected by the speech utterance durations at all. Differently, the sequential algorithm presents a failure: this is due to the use of the similarity measure (based on mono-gaussian model), which is not able to group short segments since the statistical properties do not respect the Gaussian model.

References

[1] 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.

[2] E. Singer, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, D. Sturim, "The MITLL NIST LRE 2011 Language Recognition System", Odyssey Workshop on Speaker and Language Recognition, Singapore, 26 June 2012.

[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, ISSN: 1744-2400, July 2009.

[4] J. Ajmera and G. Lathoud and I. McCowan, "Clustering And Segmenting Speakers And Their Locations In Meetings", ICASSP, Volume 1, 17-21 May 2004, pp. 605-608.

[5] 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.

[6] 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.

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

[8] H. Sayoud, S. Ouamour and S. Khennouf, "Automatic Speaker Tracking by Camera Using Two-Channel-Based Sound Source Localization", International Journal of Intelligent Computing and Cybernetics, Volume 4 (1), pp. 40- 60, 2011.

[9] P. Delacourt, C.J. Wellekens, "DISTBIC: A speaker-based segmentation for audio data indexing", Speech Communication, Volume 32, pp. 111-126, 2000.

[10] I. Magrin-Chagnolleau, J.F. Bonastre, and F. Bimbot, "Effect of Utterance Duration and Phonetic Content on Speaker Identification using Second-Order Statistical Methods", Proceedings of EUROSPEECH 95, Vol.1, pp. 337-340, Madrid, Spain, September 1995.

[11] J. F. Bonastre, and L. Besacier, "Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur", Actes du 4ème Congrès Français d'Acoustique, pp. 357-360, Marseille, France, 14-18 April 1997.

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

[13] H. Sayoud, S. Ouamour, and M. Boudraa, "'ASTRA' An Automatic Speaker Tracking System based on SOSM measures and an Interlaced Indexation", Acta Acustica, Volume 89, number 4, 2003, pp. 702-710.