Speaker Diarization Exploiting the Eigengap Criterion and Cluster Ensembles

Nikoletta Bassiou, Vassiliki Moschou, and Constantine Kotropoulos, *Senior Member, IEEE*

***Abstract*—A novel system for speaker diarization is proposed that combines the eigengap criterion and cluster ensembles. No explicit assumptions on the number of speakers are made. Two variants of the system are developed. The first variant does not cluster the speech segments that are detected as outliers, while the second one does. The aforementioned system variants are assessed with respect to various metrics, such as the overall classification error, the average cluster purity, and the average speaker purity. Experiments are conducted on two-person dialogue scenes in movies as well as on news broadcasts from MDE RT-03 Training Data Speech Corpus released by the U.S. National Institute of Standards and Technology. In the latter case, the diarization error rate is also reported. It is demonstrated that the clustering performance does not degrade when outliers are present. Moreover, thanks to the eigengap criterion, the evaluation metrics are improved.**

***Index Terms*—Speaker diarization, speaker clustering, eigengap criterion, cluster ensembles, movie scene analysis, two-person dialogues, broadcasts.**

I. Introduction

NOWADAYS, a rapid increase in the volume of recorded speech is manifested. For example, archives of television and audio broadcasting, meeting recordings, and voice mails have become a commonplace. As a result, a growing need for automatically processing such archives has arisen [1]. However, their enormous size hinders content organization, navigation, browsing, and search. *Speaker segmentation* and *speaker clustering* alleviate the management of huge audio archives.

Speaker segmentation aims at splitting an audio stream into acoustically homogeneous segments, so as each segment is attributed ideally to only one speaker [2]. It is an integral part of the MPEG-7 standard developed by the Motion Picture Experts Group [3]. To model speakers, MPEG-7 low-level audio feature descriptors, such as AudioSpectrumProjection, AudioSpectrumEnvelope [4], [5], AudioSpectrumCentroid, AudioWaveformEnvelope [6], [7] could be used. MPEG-7 high-level tools, such as SpokenContent, that exploit speakers' word usage or prosodic features, could also be exploited for speaker segmentation.

Speaker clustering refers to the unsupervised classification of speech segments based on speaker voice characteristics [8]. That is, to identify all speech segments uttered by the same speaker in an audio episode and assign a unique label to them [9]. Many speaker clustering methods have been developed ranging from hierarchical ones, such as the bottom-up (or agglomerative) methods and the top-down (or divisive) ones, to optimization methods including the K-means algorithm [10] or the autoassociative neural networks [11] to mention a few. Speaker segmentation could precede speaker clustering. In such a case, the segmentation errors degrade clustering performance. Alternatively, speaker segmentation and clustering can be jointly optimized [12]–[15].

Speaker segmentation followed by speaker clustering is called *speaker diarization* [8], [14]–[16]. It has received much attention recently, as is manifested by the focused competitions on diarization conducted under the auspices of the U.S. National Institute of Standards and Technology (NIST) [2], [17]–[19]. Speaker diarization is the process of automatically splitting the audio recording into speech segments and determining which segments are uttered by the same speaker. It is used to answer the question "who spoke when?". Speaker diarization is also related to *speaker verification* and *speaker identification*. In automatic speaker verification, the claimed speaker identity is tested whether it is true or not [20], while no a priori speaker identity claims are made and the system decides who the speaker is in speaker identification [21].

Several applications of speaker segmentation and speaker clustering could be identified. The first application is *rich transcription* [17]. Rich transcription adds several metadata in a spoken document, such as speaker identity, sentence boundary detection, annotations for disfluency and so on. A second application is *movie analysis*. For example, *dialogue detection* determines whether a dialogue occurs in an audio recording or not. Further questions, such as who the interlocutors are, when the actors appear, could also be addressed in the framework of movie analysis.

Some interesting observations on speaker segmentation and clustering can be deduced. First, speaker segmentation significantly affects speaker clustering. In addition, the inclusion of speech that comes from two speakers in a single speech segment deteriorates speaker clustering performance. Moreover, it is crucial speech segments to be homogeneous. Thus, the research community is motivated to strongly prefer over-segmentation, since false alarms are considered to be less cumbersome than miss detections. Most speaker clustering algorithms set the number of clusters a priori. This fact usually leads to more clusters than those actually existing in the data to be referred as *natural clusters*. Of course, it is preferable to start with many clusters, that can be merged in a latter step, than under-estimating the number of clusters. It is worth mentioning that the complexity of speaker diarization depends on the population size, the duration of the speech segments, the signal bandwidth, the environmental noise, the recording

The authors are with the Department of Informatics, Aristotle University of Thessaloniki, Box 451 Thessaloniki 541 24, GREECE.

Corresponding author: C. Kotropoulos, e-mail: costas@aiia.csd.auth.gr.

equipment, and whether the task has to be performed in real-time or not [22].

In this paper, a novel speaker diarization approach is proposed that combines cluster ensembles and the eigengap criterion in order to improve clustering performance. Cluster ensembles are able to reveal the natural clusters, that actually exist in data, by combining a *randomly* chosen number of partitions, which are created by multiple clustering algorithms yielding a *random* number of clusters each. One of the main reasons to employ a random number of partitions and/or random number of clusters within cluster ensembles is to avoid the bias of the voting mechanism. However, the adjacency matrix, used by most clustering algorithms, strongly depends on vertex labeling. To remove such a dependence, we employ the graph spectrum, which is a graph invariant. Accordingly, the proposed method combines the main advantages of cluster ensembles and spectral graph clustering. We resort to the eigengap criterion in order to obtain a preliminary estimate of the number of clusters present in the data. Having estimated the number of clusters, there is no need to randomly set the number of clusters to be created in each partition, while the number of partitions can be kept equal to the number of the *different* clustering algorithms tested. Cluster merging, that relies on basic set theory operations, completes the procedure by merging the clusters which significantly overlap. The proposed method exploits concepts from spectral graph theory and cluster ensembles in speaker diarization. To the best of authors knowledge, no similar work has been reported for speaker diarization. The proposed approach is tested on two application scenarios. First, two-person dialogue scenes are considered within the context of movie scene analysis. In movies, the conversations are unconstrained allowing for concurrent speech and frequently take place in the presence of background music, clapping, and/or environmental sounds, that cause speaker segmentation errors, and consequently, harden speaker clustering. Second, news broadcasts from MDE RT-03 Training Data Speech Corpus of total duration exceeding 4 hours, released by the NIST and distributed by the Linguistic Data Consortium, are considered in order to test the performance of the proposed approach when the number of speakers is greater than 2. In the latter case, single microphone recordings have been used and the reference speech/non-speech segmentation has been exploited in order to focus on a single source of the diarization error rate, namely the speaker error, that is associated to the portion of the total length of the speech segments that are clustered into wrong speaker groups.

The outline of the paper is as follows. Related work on speaker diarization emphasizing on speaker clustering is reviewed in Section II. Section III briefly addresses the distance metric used for speaker clustering, while in Section IV the proposed speaker diarization system is described in detail. Experimental results are demonstrated in Section V. Conclusions are drawn and future work is identified in Section VI.

II. Related work

Several algorithms for speaker diarization have been proposed and tested on different applications. They either assume an a priori known speaker segmentation or apply speaker segmentation in order to extract speech segments. The most representative works are briefly presented next. Generally, speaker diarization algorithms can be classified into two main categories: *deterministic* and *probabilistic* ones [22]. The deterministic algorithms cluster together similar audio segments with respect to a metric, whereas the probabilistic ones use Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs) to model the clusters.

As far as deterministic algorithms are concerned, [23], [24] propose a SOM-based speaker clustering algorithm, that assumes an a priori number of speakers and uses a SOM to model each speaker. An alternative is the on-line hierarchical speaker clustering algorithm [9]. This method considers speech segments that have been extracted manually and makes no assumptions on the number of speakers. It employs the generalized likelihood ratio and the within-cluster dispersion to estimate the distances between the segments. A deterministic step-by-step speaker diarization system is proposed by Meignier et al., that is based on speaker turn detection followed by hierarchical clustering [14].

Concerning GMM-based algorithms, Solomonoff et al. develop a method, that creates a cluster dendrogram according to the generalized likelihood ratio and the cross entropy [1]. The core of the clustering method relies on dendrogram cutting, that yields the final partition. Tsai et al. employ GMMs and propose a speaker clustering method, which is based on a voice characteristic reference space in [10]. Another method is described in [25], that is based on the maximum purity estimation. It aims to maximize the total number of within-cluster speech segments uttered by the same speaker and employs a genetic algorithm to determine the cluster where each segment should be assigned to. Jin et al. propose an automated speaker clustering algorithm [26], that builds a tree of clusters according to the distance measure introduced by Gish et al. [27]. The tree can be pruned for any given number of clusters, so that the optimal partition is produced. Lu and Zhang present an unsupervised speaker segmentation and tracking algorithm in real-time audio content analysis [28]. No prior knowledge of the number of speakers and their identity is assumed. The method first finds speaker change points, that are then validated, and afterwards speaker models are built. [15], [29] describe three variants of a speaker clustering algorithm, that use GMMs to model the clusters.

Many HMM-based algorithms have also been proposed. Ajmera et al. propose an HMM-based speaker clustering algorithm [13], where each HMM state represents a cluster and the probability density function (pdf) of each cluster is modeled by a GMM. The HMM is trained using the Expectation Maximization (EM) algorithm. A robust speaker clustering algorithm, that automatically performs both speaker segmentation and clustering without any prior knowledge of the speaker identities or the numbers of speakers, is presented by Ajmera and Wooters [30]. The algorithm uses HMMs, agglomerative clustering, and the Bayesian Information Criterion (BIC) [31], [32]. Meignier et al. also propose an integrated speaker diarization system, that generates an HMM, which detects and adds a new speaker [14]. For a more detailed

survey, the interested reader may refer to [22].

III. Distance used for speaker clustering

Speaker clusters are usually modeled as multivariate Gaussian distributions. Typical distances between two multivariate Gaussian distributions are the Kullback-Liebler (KL) [33], the Bhattacharyya (Bha) [34], the generalized likelihood ratio (GLR) [14], the cross entropy [1], the $d_{COVMEAN}$ [27] etc. It must be mentioned that although the KL, the Bha, and the $d_{COVMEAN}$ distances always admit positive values, they do not satisfy the triangular inequality. All the aforementioned distances have been tested within the clustering algorithm detailed in Section IV in order to assess their influence on the clustering evaluation measures defined in Section V-B. However, it has been found by experiments that the $d_{COVMEAN}$ performs better than the KL and the Bha. Thus, our discussion is confined to this distance only. Let $\mathcal{N}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$ be a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ modeling each class. Let also N_i, $i = 1, 2$ be the number of feature vectors assigned to each class. The $d_{COVMEAN}$ distance between the aforementioned Gaussian distributions is estimated as [27]:

$$d_{COVMEAN} = d_{COV} \cdot d_{MEAN} \tag{1}$$

where

$$d_{COV} = \left(\frac{|\boldsymbol{\Sigma}_1|^a \, |\boldsymbol{\Sigma}_2|^{1-a}}{|\mathbf{W}|} \right)^{\frac{N_T}{2}} \tag{2}$$

$$d_{MEAN} = \left(1 + \frac{N_1 \, N_2}{N_T^2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{W}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right)^{-\frac{N_T}{2}} \tag{3}$$

with $N_T = N_1 + N_2$, $a = \frac{N_1}{N_T}$, and $\mathbf{W} = a\boldsymbol{\Sigma}_1 + (1-a)\boldsymbol{\Sigma}_2$.

IV. Proposed diarization system

The proposed speaker diarization system is composed of four serially connected modules, namely: the speaker segmentation module, the speaker modeling module, the clustering module, and the cluster merging module. Voiced speech is input to the system. Voiced/unvoiced (V/U) segmentation is performed by PRAAT [35]. PRAAT's algorithm performs acoustic periodicity detection on the basis of an accurate autocorrelation method, as described in [36].

A. Speaker segmentation module

A BIC-based speaker segmentation algorithm is applied. BIC is a robust, well-founded statistical criterion, widely used by the research community [31]–[33]. According to the BIC, speaker segmentation is formulated as a two hypothesis testing problem. Let $\mathcal{X} = \{\mathbf{x}_i, \quad i = 1, 2, \ldots, N_{\mathcal{X}}\}$ be the set of feature vectors extracted from an acoustic chunk (i.e. acoustic segment of minimum duration 0.5 s in our case) before time t_j. Here, $\mathbf{x}_i$ denotes the vector of 24 Mel Frequency Cepstral Coefficients (MFCCs), which are extracted every 10 ms for 20 ms long frames determined by Hamming windowing within the acoustic chunk under discussion. MFFCs are used, because they proved to work better in noisy environments than other features, e.g. Linear Prediction Coefficients [30]. Let $\mathcal{Y} = \{\mathbf{y}_i, \quad i = 1, 2, \ldots, N_{\mathcal{Y}}\}$ be the set of feature vectors extracted from a neighboring acoustic chunk of duration 0.5 s just after t_j. The problem is to decide whether a speaker change point occurs at t_j or not. Let $\mathcal{Z} = \mathcal{X} \bigcup \mathcal{Y}$.

Under the null hypothesis, H_0, there is no speaker change at time t_j. The feature vectors in $\mathcal{Z}$ are then modeled by a single multivariate Gaussian density with parameters (e.g. the mean vector $\boldsymbol{\mu}_{\mathcal{Z}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\mathcal{Z}}$) stacked into the vector $\boldsymbol{\theta}_{\mathcal{Z}}$. The log likelihood under H_0, L_0, is calculated as [32]:

$$L_0 = \sum_{i=1}^{N_{\mathcal{X}}} \log p(\mathbf{x}_i | \boldsymbol{\theta}_{\mathcal{Z}}) + \sum_{i=1}^{N_{\mathcal{Y}}} \log p(\mathbf{y}_i | \boldsymbol{\theta}_{\mathcal{Z}}). \tag{4}$$

Under the alternative hypothesis, H_1, a speaker change occurs at time t_j. Accordingly, the feature vectors in $\mathcal{X}$ and $\mathcal{Y}$ are modeled by separate multivariate Gaussian densities, whose parameters are denoted by $\boldsymbol{\theta}_{\mathcal{X}}$ and $\boldsymbol{\theta}_{\mathcal{Y}}$, respectively. The corresponding log likelihood L_1 is given by [32]:

$$L_1 = \sum_{i=1}^{N_{\mathcal{X}}} \log p(\mathbf{x}_i | \boldsymbol{\theta}_{\mathcal{X}}) + \sum_{i=1}^{N_{\mathcal{Y}}} \log p(\mathbf{y}_i | \boldsymbol{\theta}_{\mathcal{Y}}). \tag{5}$$

The dissimilarity between L_1 and L_0 according to BIC is calculated as [32]:

$$\Delta = L_1 - L_0 - \frac{\lambda}{2} \left(K + \frac{K(K+1)}{2} \right) \log N_{\mathcal{Z}} \tag{6}$$

where $N_{\mathcal{Z}} = N_{\mathcal{X}} + N_{\mathcal{Y}}$ is the total number of feature vectors in $\mathcal{Z}$, λ is a penalty factor, and K is the dimension of the extracted feature vector (e.g. $K = 24$). When the covariance matrix of each Gaussian pdf is estimated by the sample dispersion matrix, (6) takes the form:

$$\begin{aligned} \Delta = {} & N_{\mathcal{Z}} \log |\boldsymbol{\Sigma}_{\mathcal{Z}}| - N_{\mathcal{X}} \log |\boldsymbol{\Sigma}_{\mathcal{X}}| - N_{\mathcal{Y}} \log |\boldsymbol{\Sigma}_{\mathcal{Y}}| \\ & - \frac{\lambda}{2} \left(K + \frac{K(K+1)}{2} \right) \log N_{\mathcal{Z}}. \end{aligned} \tag{7}$$

However, a more generic BIC formulation has been derived in [37], which allows for alternative estimates of the covariance matrix to be employed, such as the minimum covariance determinant estimate [38]. If $\Delta > 0$, t_j is declared to be a speaker change point. Otherwise, it is decided that there is no speaker change point at time t_j. The next test is applied at $t_j = t_j + 0.5$ s. Obviously, the next chunk $\mathcal{X}$ will be equal to the previous $\mathcal{Y}$, when a speaker change point is found. Otherwise, it will be equal to the previous $\mathcal{Z}$ (i.e. when there is no speaker change point). The penalty factor λ is frequently tuned by employing development data despite the fact that its ideal value equals 1.0. In our experiments, the penalty factor λ is set to 1.0. Obviously, the resolution of the segmentation is 0.5 s. Over-segmentation is strongly preferred against the risk of not detecting true speaker change points in order to ensure the homogeneity of speech segments. The speaker segmentation module outputs N speech segments $\mathcal{S}_i$, $i = 1, 2, \ldots, N$.

B. Speaker modeling module

The N resulted speech segments are fed as input to the speaker modeling module. Let us assume that the speech segments are homogeneous. Accordingly, if the feature vectors within each speech segment are treated as independent identically distributed (i.i.d.) Gaussian random vectors, then each speech segment can be modeled by the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2, \ldots, N$. It must be noted that the number of feature vectors modeled by each multivariate Gaussian distribution depends on the duration of the corresponding speech segment. Indeed, the parameters of the multivariate Gaussian pdf are more accurately estimated when more feature vectors are available, as is the case for long homogeneous segments.

C. Clustering module

The third module of the algorithm, the clustering module, deals with the task of revealing the natural clusters hidden in the data. The arbitrary shape of the clusters makes the choice of the suitable clustering algorithm a challenging issue. Furthermore, when simultaneous speech from 2 or more speakers occurs, implying that clusters overlap, revealing the natural clusters of the data becomes extremely difficult.

Generally, a single clustering algorithm usually underperforms on datasets that contain clusters of arbitrary shapes/sizes or clusters that are nested within one another. For example, k-means or k-medoids [39] may fail to detect clusters of non-spherical shape. Thus, it is necessary to use multiple algorithms in order to reveal the natural groupings of the data [40]. Cluster ensembles are collections of clusterings, which are of the same "kind", e.g., collections of partitions or collections of hierarchies [41]. Here, we are interested in collections of partitions. They combine the different partitions in order to improve the clustering performance, as is assessed in Section V-B, and increase the robustness to outliers. Such objectives are challenging, because a) speaker clustering is strongly affected by the presence of noise in the speech signal and b) the partitions produced by different clustering algorithms, when they are applied to the same dataset, may vary [42]. A cluster ensemble produces N_P different partitions and then combines them using a consensus function in order to reveal the natural clusters of the data. The consensus function can be derived from the co-association matrix [40] and/or voting [44], [45]. Evidence accumulation is a voting mechanism for the combination of N_P partitions. The co-occurrences of speech segment pairs in the same cluster are considered as votes for their association. Thus, the N_P partitions of N speech segments are mapped into an $N \times N$ co-association matrix with entries

$$CM(i,j) = \frac{v_{ij}}{N_P} \tag{8}$$

where v_{ij} is the number of times the segments $\mathcal{S}_i$ and $\mathcal{S}_j$ are assigned to the same cluster among the N_P partitions [40].

In the proposed system, three hierarchical algorithms build the cluster ensemble, namely the average group linkage, the weighted average group linkage, and Ward's hierarchical clustering method [46]. The use of hierarchical algorithms in speaker clustering is motivated by their popularity and the progressive manner the clusters are created. In addition, by examining the height of the dendrogram links one may determine whether a grouping is natural or forced [46].

Our first consideration has to do with setting the number of partitions and the number of clusters produced in each partition. The typical choice is to combine a randomly chosen number of partitions where each partition creates a randomly chosen number of clusters. It is rational though, to upper limit the number of partitions by the number of clustering algorithms to be applied, i.e. 3. However, since each clustering algorithm ends up with one partition and performs well for different data shapes and sizes, the problem of setting the number of clusters in each partition becomes crucial. Instead of randomly setting the number of clusters, it would be less risky to initially predicting this number.

The number of clusters can be quickly and efficiently predicted using the eigengap criterion, borrowed from spectral graph theory [47]. Spectral graph theory studies the properties of a graph with respect to the eigenvalues and associated eigenvectors of its adjacency matrix or its Laplacian matrix, as is detailed next. It determines how combinatorial features of a graph can be revealed by its spectrum, which is a graph invariant. Clustering the speech segments can be considered as a weighted graph partitioning problem. The $N \times N$ adjacency matrix $\mathbf{A}$ is built by treating the speech segments as graph vertices. The element $A(i,j)$ corresponds to the weight between speech segments $\mathcal{S}_i$ and $\mathcal{S}_j$, which is defined as the distance $d_{COVMEAN}$ between the multivariate Gaussian pdfs modeling the aforementioned speech segments. The adjacency matrix is symmetric. The un-normalized Laplacian of the graph is given by [47]

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{9}$$

where $\mathbf{D}$ is the $N \times N$ diagonal matrix with $D(i,i) = \sum_j A(i,j)$. That is, $D(i,i)$ is the sum of the weights of the edges that are incident to vertex i. The normalized Laplacian of the graph can be derived as [47]

$$\mathbf{L}_{norm} = \mathbf{D}^{-\frac{1}{2}}\, \mathbf{L}\, \mathbf{D}^{-\frac{1}{2}}. \tag{10}$$

Since $\mathbf{L}_{norm}$ is symmetric, its eigenvalues are all real and non-negative. The set of the eigenvalues $\beta_0 \leq \beta_1 \leq \ldots \leq \beta_{N-1}$ is called the spectrum of the graph [47]. In the ideal case of N_c completely disconnected clusters, the eigenvalue 0 has multiplicity N_c, and then there is a gap to the $(N_c + 1)$th eigenvalue $\beta_{N_c+1} > 0$. This is the so-called eigengap criterion. Accordingly, the most stable clustering is generally obtained for k that maximizes the difference $(\beta_k - \beta_{k-1})$. The eigengap criterion can be applied to $\mathbf{L}_{norm}$ in order to predict the number of clusters for each partition. The criterion usually works well if the data contain very well pronounced clusters, but in ambiguous cases it also returns ambiguous results [48]. A similar heuristic was also applied in [49].

After having determined the number of clusters to be formed, the three clustering methods yield three partitions and the co-association matrix (8) is computed. A speech segment $\mathcal{S}_i$ is assigned to the same cluster with segment $\mathcal{S}_j$, when

$$CM(i,j) \geq \vartheta \quad i \neq j \tag{11}$$

where ϑ is a threshold. It is expected that the larger the value of ϑ becomes, the higher is the similarity between the speech segments assigned to the same cluster. ϑ values $\in [0.5, 0.9]$ are tested in order to find the value that improves the performance most. It has been found that $\vartheta = 0.7$ suffices for good clustering.

Finally, N sets $\mathcal{S}_i$, $i = 1, 2, \ldots, N$ are created, where $\mathcal{S}_i$ contains the indices of the speech segments that are clustered together with segment $\mathcal{S}_i$. It is expected that the overlap between the sets, which consist of segments belonging to the same (natural) cluster, will be large.

D. Cluster merging module

The most critical step of the algorithm is cluster merging. Since we strongly prefer over-segmentation, it is possible to end up with more speech segments than those ideally should be produced by choosing a small chunk duration. This fact, combined with the use of threshold ϑ might cause speech segments from the same speaker to be split into different clusters. Such clusters should be merged. The cluster merging algorithm is applied recursively to $\mathcal{S}_i$. The iterative procedure stops, when no more mergers occur. At each iteration, the intersection between $\mathcal{S}_i$ and $\mathcal{S}_j$ is calculated and the following heuristic rule is applied: *$\mathcal{S}_i$ and $\mathcal{S}_j$ are merged together into a single cluster, if the cardinality of the intersection $\mathcal{S}_i \bigcap \mathcal{S}_j$ is greater than or equal a threshold set equal to a portion of the smallest cardinality of the individual sets $\mathcal{S}_i$ and $\mathcal{S}_j$*. The larger the value of the threshold is, the more similar clusters are merged in the sense that the number of common speech segments shared between the two sets is larger. For example, the threshold is set to 4/5 in two-person dialogue movie scenes and 6/10 in MDE RT-03 news broadcasts to avoid the risk of merging clusters not really belonging together. Next, the union $\mathcal{S}_i \bigcup \mathcal{S}_j$ is determined, that replaces the individual sets. At the end of the recursion, N_c' sets are produced.

A usual phenomenon during cluster merging is the creation of singleton clusters, i.e. clusters that contain only one speech segment. This fact practically means that the specific speech segment can not be clustered together with any other segment, implying that such a speech segment might be an outlier. Outliers usually appear in speaker clustering applications, due to the following reasons.

1) Speaker segmentation algorithms are not error-free, leading to non-homogeneous speech segments, that contain data from more than one speaker.
2) Speaker segmentation algorithms can not deal with concurrent speech.
3) Due to noise attributed to either the environmental/recording conditions or speakers' affective condition, there exist speech segments that are not similar to each other.

Generally, it is not desirable to have "small" clusters. It is proposed that a speech segment that cannot be assigned to any cluster to be treated as an outlier. In order to study the effect of outliers in speaker diarization, two variants of the diarization system are studied. The first variant excludes the outliers from the clustering procedure, while the second variant does not. In both variants, a cluster $\mathcal{C}_\ell$ is modeled by a GMM with k_ℓ components each, where k_ℓ is the number of speech segments assigned to it.

The first variant ends up with N_c compact clusters and N_o outlier segments, which are singleton clusters. Only the N_c compact clusters resulted after merging are considered. It is expected that the performance of the first variant will be better, since the outliers, that generally deteriorate clustering performance, are excluded. By removing the outliers, the speaker time to be clustered will be reduced as well, which is not generally desirable.

In the second variant, the N_o outlier segments (singleton clusters) are forced to be merged with the other speech segments resulting again to N_c compact clusters at the end. The outlier segments are assigned to the closest *compact* cluster. It it reminded that a speech segment is modeled by a Gaussian pdf. Thus, the problem of determining the distance between a cluster $\mathcal{C}_\ell$, $\ell = 1, 2, \ldots, N_c$ and an outlier speech segment $\mathcal{S}_j$, $j = 1, 2, \ldots, N_o$ is reduced to calculating the distance between a GMM and a single Gaussian pdf. The distance between a Gaussian component $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ and the ℓth Gaussian mixture of k_ℓ Gaussian components $\mathcal{N}(\boldsymbol{\mu}_{m\ell}, \boldsymbol{\Sigma}_{m\ell})$, $m = 1, 2, \ldots, k_\ell$ is given by

$$d(\{\boldsymbol{\Sigma}_{m\ell},\ \ m = 1, 2, \ldots, k_\ell\}, \boldsymbol{\Sigma}_j) = \sum_{m=1}^{k_\ell} \gamma_{m\ell}\, d_{COVMEAN}(\boldsymbol{\Sigma}_{m\ell}, \boldsymbol{\Sigma}_j) \tag{12}$$

where $\gamma_{m\ell} = \frac{N_{m\ell}}{N_{\mathcal{C}_\ell}}$, $N_{m\ell}$ is the number of speech frames in the mth segment, and $N_{\mathcal{C}_\ell} = \sum_{m=1}^{k_\ell} N_{m\ell}$ is the total number of speech frames assigned to the speech segments, which belong to the ℓth cluster $\mathcal{C}_\ell$. Each speech segment $\mathcal{S}_j$ is assigned to the GMM $\mathcal{C}_\ell$, $\ell = 1, 2, \ldots, N_c$ that minimizes (12).

V. Experiments and Results

This section describes the data used in the experiments, the evaluation measures employed, the baseline system tested against the proposed speaker diarization system, and the results obtained.

A. Data

Two different datasets have been used in the experiments. The first dataset comprises 15 scenes of two-person dialogues extracted from 5 movies, while the second dataset is a subset consisting of 43 single microphone recordings extracted from the *MDE RT-03 Training Data Speech corpus* [50].

1) Two-person dialogue movie scenes: 15 scenes extracted from *Analyze That, Cold Mountain, Jackie Brown, Lord of the Rings I,* and *Secret Window* were used in the experiments. Their audio track is digitized in PCM at a sampling rate of 48 kHz. Each sample is quantized in 16 bit two-channel. The total duration of the scenes is 16 min and 16 s. All scenes contain two-person dialogues implying that ideally the speaker diarization system should yield 2 natural clusters. Based on the background noise, the dialogue scenes can be either *clean dialogue scenes (CD)*, i.e. scenes with low-level audio background or *background dialogue scenes (BD)* where

a noisy background, such as music, clapping, or environmental noise is present. Table I summarizes the scene details, e.g. the movie title they are extracted from, their duration in seconds, the actors' genders (F for female, M for male), and the dialogue type. In addition, ground truth information related to exact timing information for each actor appearance is available for all movie scenes [51][1]. To provide an estimate of the noise power we selected 4 pairs of BD and CD scenes extracted from the same movie and estimated the ratio of the average signal power in the BD scenes over the average signal power in the CD scenes. This ratio was found to be 7.34 (or 8.657 dB).

TABLE I
DETAILS ON THE MOVIE SCENES.

Scene	Movie title	Duration (s)	Actors' genders	Dialogue type
AT1	Analyze That	63	M-F	CD
AT2	Analyze That	57	M-M	CD
CM1	Cold Mountain	35	M-F	CD
CM2	Cold Mountain	71	F-F	BD
CM3	Cold Mountain	76	M-M	CD
CM4	Cold Mountain	91	M-F	BD
CM5	Cold Mountain	69	M-F	CD
JB1	Jackie Brown	65	M-M	CD
JB2	Jackie Brown	94	M-M	CD
JB3	Jackie Brown	95	M-F	CD
LOTR1	Lord of the Rings I	41	M-M	BD
LOTR2	Lord of the Rings I	30	M-M	BD
LOTR3	Lord of the Rings I	56	M-M	CD
SW1	Secret Window	45	M-M	BD
SW2	Secret Window	88	M-M	CD

2) MDE RT-03 Training Data Speech Corpus subset: It comprises 43 sound files of *BN speech data* that are encoded in 16-bit PCM with a sampling rate of 16kHz. In Table II, the sound files that were used, their duration in seconds, and the actual number of speakers are summarized. The total duration of the sound files is 4 hr and 10 min.

B. Evaluation measures

To evaluate the performance of a speaker clustering algorithm, several measures are used: the average classification error; the cluster purity and its average value; the speaker purity and its average value; and the diarization error rate.

Let

n_{ij} be the total number of audio segments in cluster i uttered by actor j;
N_a be the total number of actors;
N_c be the total number of clusters;
N be the total number of audio segments;
$n_{.j}$ be the total number of audio segments uttered by actor j;
$n_{i.}$ be the total number of audio segments in cluster i.

Relationships between the aforementioned variables are as follows:

$$n_{i.} = \sum_{j=1}^{N_a} n_{ij}, \quad n_{.j} = \sum_{i=1}^{N_c} n_{ij}, \quad N = \sum_{i=1}^{N_c}\sum_{j=1}^{N_a} n_{ij}. \tag{13}$$

[1]Data can be shared with interested researchers upon request.

TABLE II
DETAILS ON THE SUBSET OF THE MDE RT-03 TRAINING DATA SPEECH (N_a DENOTES THE NUMBER OF SPEAKERS.

File No	Sound file	Duration (s)	N_a
1	ea980107-split003	360.801	4
2	ea980108-split002	400.745	6
3	ea980109-split003	376.828	7
4	ea980110-split001	369.88	1
5	ea980110-split002	295.759	5
6	ea980112-split003	349.783	2
7	ea980113-split002	343.783	8
8	ea980114-split002	347.734	10
9	ea980120-split002	379.755	12
10	ea980122-split004	295.804	4
11	ea980123-split001	422.477	9
12	ea980123-split002	363.76	9
13	ea980123-split003	388.086	13
14	ea980124-split004	264.551	4
15	ea980126-split004	301.242	7
16	ea980127-split001	411.793	9
17	ea980128-split002	407.199	7
18	ea980130-split003	432.749	12
19	ed980106-split003	416.045	3
20	ed980106-split004	347.918	3
21	ed980106-split005	218.773	9
22	ed980108-split001	335.964	3
23	ed980108-split005	336.448	8
24	ed980111-split003	353.55	5
25	ed980112-split001	393.373	6
26	ed980121-split003	334.303	3
27	ed980122-split001	339.621	4
28	ee970625-split004	311.388	3
29	ee970625-split007	481.43	7
30	ee970626-split005	190.793	3
31	ee970626-split007	269.768	7
32	ee970627-split005	326.187	3
33	ee970627-split006	369.379	10
34	ee970627-split009	359.83	5
35	ee970630-split004	319.577	4
36	ee970630-split007	411.063	11
37	ee970701-split003	577.378	4
38	ee970701-split005	342.602	3
39	ee970702-split004	216.034	3
40	ee970703-split006	360.992	14
41	ee970722-split003	243.975	6
42	ee970723-split005	343.583	4
43	ee970807-split002	333.597	5

1) Average classification error: The classification error is defined as the percentage of time not attributed correctly to a reference speaker [30]. The error for cluster i, CE_i, is defined as the percentage of the total time spoken by actor whose speech segments appear in majority in cluster i, that has not been clustered to this cluster. The average classification error, ace, is defined as [30]:

$$ace = \frac{1}{N_c}\sum_{i=1}^{N_c} CE_i. \tag{14}$$

ace admits values between 0 and 1. The smaller the ace value is, the better performance is achieved.

2) Average cluster purity: The purity of cluster i, is defined as [1]

$$\pi_{i.} = \sum_{j=1}^{N_a} n_{ij}^2 / n_{i.}^2. \tag{15}$$

The average cluster purity is given by [10]

$$acp = \frac{1}{N} \sum_{i=1}^{N_c} \pi_{i.}\ n_{i.}. \tag{16}$$

acp provides a measure of how well a cluster is limited to only one speaker. It admits values between 0 and 1. The higher the acp value is, the more homogeneous the clusters are.

3) Average speaker purity: The purity of speaker (actor) j, is defined as [13]

$$\pi_{.j} = \sum_{i=1}^{N_c} n_{ij}^2 / n_{.j}^2. \tag{17}$$

The average speaker purity is given by [13]

$$asp = \frac{1}{N} \sum_{j=1}^{N_a} \pi_{.j}\ n_{.j}. \tag{18}$$

asp provides a measure of how well a speaker is limited to only one cluster. It admits values between 0 and 1. The higher the asp value is, the better the speaker is limited to one cluster. The calculation of asp is required, since acp alone can be misleading.

4) Diarization error rate: It is defined by the NIST Rich Transcription Evaluation [52] as

$$derr = \frac{T_{FA} + T_{MS} + T_{wrong}}{T_{total}} \tag{19}$$

where T_{FA} is the total duration of the non-speech segments that were classified as speech, T_{MS} is the total duration of the speech segments that were classified as either non-speech or silence, T_{wrong} is the total duration of speech segments that were correctly classified as speech, but that were clustered into wrong speaker groups, and T_{total} is the total duration of all the speech segments. This measure is reported for the speaker diarization experiments conducted on the MDE RT-03 Training Data Speech Corpus subset.

C. Baseline system

The baseline system consists of the same four serially connected modules: the speaker segmentation module, the speaker modeling module, the clustering module, and the cluster merging module. All modules are left intact, as they were described in Section IV, *except the clustering module*. The differentiation lies in the way the cluster ensemble is built by the clustering module. In the baseline system, the number of clusters is *not* estimated by the eigengap criterion. The average group linkage, the weighted average group linkage, and Ward's hierarchical clustering methods are employed again to create 30 different partitions of varying numbers of clusters. Such a number of partitions is experimentally found to be adequate for the ensemble voting mechanisms to yield good clustering results. For each partition, the clustering algorithm to be applied is randomly selected among the three hierarchical algorithms. Furthermore, the number of clusters to be created is randomly chosen in the range [2, 20]. The consensus function (8) and the criterion (11) are employed in order to derive the index sets $\mathcal{S}_i$, $i = 1, 2, \ldots, N$ and their merger, as described in Section IV-D, yields systematically (i.e. not randomly) the baseline clusters. Parameter ϑ is set equal to 0.7.

D. Results

1) Two-person dialogue movie scenes: In order to evaluate the speaker diarization performance, experiments are carried out, when the outliers are either excluded from or included in the clustering procedure.

Tables III and IV show the number of clusters created, and the values of ace, acp, and asp measured in each movie scene for the proposed system and the baseline one, when outliers are either excluded from or included in the clustering procedure, respectively. Table V summarizes the cumulative figures of merit (i.e. the mean value and standard deviation of the number of clusters, the ace, the acp, and the asp) for the proposed system and the baseline one. The first number in parentheses corresponds to the mean value, while the second one is the standard deviation. The best figures of merit are indicated in bold.

TABLE III
FIGURES OF MERIT FOR THE PROPOSED SPEAKER DIARIZATION SYSTEM, WHEN OUTLIERS ARE EITHER EXCLUDED FROM OR INCLUDED IN CLUSTERING.

Scene	**Number of clusters and evaluation measures**							
	Outliers excluded				Outliers included			
	N_c	ace(%)	acp	asp	N_c	ace(%)	acp	asp
AT1	2	7.57	0.92	0.93	2	13.33	0.89	0.90
AT2	4	21.74	0.68	0.53	4	21.28	0.69	0.53
CM1	3	9.21	0.85	0.55	3	9.22	0.85	0.55
CM2	2	32.43	0.58	0.87	2	33.11	0.57	0.85
CM3	2	30.56	0.61	0.51	2	31.14	0.60	0.52
CM4	2	26.23	0.72	0.72	2	25.87	0.73	0.70
CM5	2	6.36	0.83	0.83	2	5.64	0.84	0.84
JB1	3	16.62	0.65	0.47	3	16.39	0.66	0.47
JB2	3	25.24	0.66	0.51	3	24.54	0.66	0.51
JB3	2	8.01	0.84	0.84	2	12.81	0.74	0.74
LOTR1	2	44.44	0.52	0.60	2	37.02	0.52	0.60
LOTR2	2	13.76	0.87	0.81	2	13.76	0.87	0.82
LOTR3	2	11.12	0.83	0.84	2	8.52	0.84	0.85
SW1	2	3.16	0.93	0.93	2	4.16	0.91	0.91
SW2	3	38.92	0.55	0.50	3	38.93	0.56	0.50

The inspection of Table III gives rise to several interesting observations. 1) The proposed system predicts the correct number of clusters (e.g. 2) in most cases. It is seen that in 10 out of the 15 movie scenes the number of clusters is accurately predicted (i.e. success 66.6%). It is encouraging that the proposed system succeeds to predict the correct number of clusters for *all* background dialogue scenes. 2) In the case of clean dialogue scenes, the values measured are significantly better than those for background dialogues, as it was expected. However, there are cases that deserve further attention. In AT2 and SW2 scenes, the system not only fails to predict the correct number of clusters, but the ace is large, too. It must be mentioned that in these scenes, one or both actors are yelling, respectively. The system did not detect that the actor yelling was the same with that previously speaking and had clustered the corresponding speech segments into different clusters. Concerning JB2, this scene contains simultaneous speech from two actors. Thus, the proposed system formed 3

TABLE IV
FIGURES OF MERIT FOR THE BASELINE SPEAKER DIARIZATION SYSTEM, WHEN OUTLIERS ARE EITHER EXCLUDED FROM OR INCLUDED IN CLUSTERING.

Scene	Number of clusters and evaluation measures							
	Outliers excluded				Outliers included			
	N_c	ace(%)	acp	asp	N_c	ace(%)	acp	asp
AT1	4	9.99	0.79	0.53	3	11.83	0.75	0.51
AT2	2	46.59	0.54	0.90	5	52.42	0.67	0.38
CM1	1	4.17	0.55	1.00	1	23.20	0.52	1.00
CM2	7	20.97	0.69	0.46	6	16.13	0.78	0.32
CM3	1	32.81	0.58	1.00	1	31.14	0.60	1.00
CM4	1	29.03	0.69	1.00	5	25.87	0.74	0.37
CM5	1	16.35	0.85	1.00	1	31.02	0.56	1.00
JB1	4	17.86	0.66	0.60	4	16.39	0.68	0.61
JB2	2	26.69	0.62	0.92	3	28.27	0.61	0.76
JB3	2	24.11	0.67	0.80	3	33.40	0.64	0.44
LOTR1	4	37.28	0.64	0.58	4	37.02	0.58	0.54
LOTR2	2	14.64	0.87	0.83	2	13.76	0.74	0.52
LOTR3	2	14.64	0.87	0.83	1	48.98	0.50	1.00
SW1	4	16.46	0.75	0.43	5	22.89	0.64	0.37
SW2	5	38.87	0.64	0.60	2	41.43	0.53	0.91

clusters with high ace. Scene CM3, although it is characterized as clean dialogue, includes actors under stress who sometimes are whispering. This fact implies that the emotional state of the speakers plays an important role in speaker clustering. 3) BD scenes generally demonstrate a high ace as well as low acp and asp values. This is not the case for LOTR2 and SW1 scenes, where noise appears when no actor talks. Accordingly, the proposed system is able to discriminate speech segments and assign them to the appropriate cluster. In the remaining background dialogues noise co-occurs while some speaker talks and the system is thus confused. 4) When outliers are included in the clustering, the system performance generally deteriorates than when they are excluded. Exceptions are the scenes CM4, CM5, LOTR1, and LOTR3 for which the measured figures of merit are better when outliers are included than when they are excluded. The latter observation could be attributed to the fact that some singleton speech segments have mistakenly been considered as outliers, and, thus, their inclusion in the closest cluster improves performance. It must be mentioned that the time reduction caused, when outliers are excluded, equals 0.9% of the total time, implying that our algorithm efficiently discriminates and handles the outliers.

To assess the impact of the eigengap criterion exploited within cluster ensemble, Table IV summarizes the evaluation measures for the baseline system on the same movie scenes. It is seen that the baseline system ends up with a different number of clusters when outliers are excluded from than when they are included in the clustering for the *same* scene. The baseline system predicts the correct number of clusters only in 5 out of the 15 scenes in the absence of outliers, and only in 2 out of the 15 scenes, when outliers are included. Additionally, there exist scenes for which the algorithm fails to predict at least 2 clusters and creates only 1 cluster. Concerning the presence of outliers or not, the influence of outliers is more evident in the baseline system than the proposed one that utilizes the eigengap criterion. When the outliers are included, the evaluation measures values are significantly worse than those when the outliers are excluded. In all cases except AT1 and CM2, the baseline system performs worse when the outliers are included than when they are omitted. The exceptions refer to two scenes where 3 and 6 clusters are created, respectively, a fact that reduces the error. When outliers are excluded, there are 3 scenes, namely CM1, CM2, and LOTR1, where the baseline system yields better figures of merit than the proposed system. For scenes CM2 and LOTR1, this is attributed to the large number of the clusters created. In the case of CM1, the small error is explained by the fact that only 12 out of 35 s are clustered by the baseline system. For the sake of completeness, it is reminded that 31.57 s are clustered by the proposed system when the outliers are excluded, which is more desirable.

TABLE V
CUMULATIVE FIGURES OF MERIT (I.E., MEAN AND STANDARD DEVIATION) FOR THE PROPOSED SPEAKER DIARIZATION SYSTEM AND THE BASELINE ONE, WHEN THE OUTLIERS ARE EITHER EXCLUDED FROM OR INCLUDED IN CLUSTERING.

Figure of merit	Proposed system		Baseline system	
	Outliers excluded	Outliers included	Outliers excluded	Outliers included
N_c	2	2	3	3
$ace(\%)$	(**19.69, 12.82**)	(19.71, 11.50)	(25.30, 12.59)	(28.57, 12.33)
acp	(**0.74, 0.14**)	(0.73, 0.13)	(0.67, 0.11)	(0.64, 0.08)
asp	(0.70, 0.17)	(0.69, 0.10)	(**0.77, 0.22**)	(0.66, 0.27)

Table V accumulates the results of all figures of merit for both the proposed and the baseline systems. The proposed system predicts the correct number of clusters in either the presence or the absence of outliers. The best ace and acp values measured are 19.69% and 0.74, respectively, when outliers are excluded from clustering. However, the best asp of 0.77 is returned by the baseline system, when the outliers are excluded. The latter value is a consequence of the existence of only 1 cluster in 4 scenes leading to a unit asp value. It is obvious that the objective figures of merit for the proposed system vary slightly with the inclusion or exclusion of outliers. This is not the case with the baseline system. This fact in addition to the very low time reduction, when outliers are included, supports the argument that the proposed system handles efficiently the outliers. That is, it does not isolate speech segments as outliers, if they are not actually such.

Let us finally address the discriminative power of the proposed system. This term refers to the ability of the system to assign the clusters created to different speakers. For example, if all clusters were assigned to only a single speaker in each two-person dialogue scene, the system would obviously fail. On the opposite, when different speakers are assigned to the clusters created in the two-person dialogue scenes, the system succeeds to discriminate between the speakers. The discriminative power of the proposed system reaches 60%, because there are 6 movie scenes (i.e. CM2, CM3, JB1, LOTR1, LOTR2, and SW1) for which the system fails to discriminate among the speakers.

2) MDE RT-03 Training Data Speech Corpus subset: Since the proposed system is able to handle the outliers efficiently, we will not exclude the outliers from the clustering here. In

addition, the reference speech/non-speech segmentation has been exploited in order to focus on a single source of the diarization error rate, namely the speaker error T_{wrong}, that is associated to the portion of the total length of the speech segments that are clustered into wrong speaker groups.

Table VI shows the actual number of clusters, N_a, the number of clusters created, N_c, as well as the values of der, acp, and asp measured in each sound file for the proposed system and the baseline one. The performance of the proposed system with respect to the diarization error is improved by 8.1% on average, since the average diarization error of the proposed system is measured to be 20.2%, while that of the baseline system is 28.3%. The average acp and the average asp of proposed system is 0.81 and 0.59, respectively. The corresponding measures of the baseline system are 0.72 and 0.52, respectively.

TABLE VI
EVALUATION MEASURES FOR THE MDE RT-03 TRAINING DATA SPEECH CORPUS SUBSET.

File No	N_a	**Proposed system**				**Baseline system**			
		N_c	der	acp	asp	N_c	der	acp	asp
1	4	5	0.10	0.95	0.44	7	0.22	0.79	0.44
2	6	9	0.24	0.83	0.60	6	0.34	0.65	0.66
3	7	10	0.22	0.80	0.87	5	0.48	0.49	0.59
4	1	9	0.33	0.75	0.69	3	0.53	0.46	0.60
5	5	11	0.13	0.89	0.40	5	0.31	0.69	0.58
6	2	7	0.00	1.00	0.44	6	0.02	0.98	0.38
7	8	10	0.31	0.77	0.72	6	0.38	0.62	0.54
8	10	8	0.33	0.73	0.87	4	0.52	0.55	0.38
9	12	15	0.23	0.73	0.51	4	0.39	0.50	0.66
10	4	9	0.24	0.73	0.70	5	0.45	0.51	0.46
11	9	11	0.22	0.73	0.54	8	0.27	0.68	0.53
12	9	9	0.36	0.61	0.58	6	0.40	0.50	0.56
13	13	14	0.36	0.68	0.89	7	0.39	0.60	0.60
14	4	4	0.11	0.89	0.86	7	0.15	0.84	0.75
15	7	10	0.22	0.76	0.38	8	0.24	0.72	0.40
16	9	8	0.28	0.75	0.69	6	0.30	0.63	0.54
17	7	11	0.13	0.81	0.53	9	0.18	0.84	0.64
18	12	8	0.24	0.75	0.68	8	0.25	0.76	0.55
19	3	3	0.14	0.94	0.42	6	0.20	0.89	0.42
20	3	4	0.09	0.90	0.76	8	0.07	0.92	0.24
21	9	9	0.30	0.75	0.44	6	0.33	0.73	0.46
22	3	4	0.08	0.87	0.47	2	0.14	0.70	0.89
23	8	8	0.25	0.70	0.71	7	0.30	0.65	0.48
24	5	6	0.15	0.85	0.38	8	0.17	0.73	0.44
25	6	10	0.27	0.84	0.73	7	0.37	0.75	0.61
26	3	5	0.19	0.95	0.61	8	0.19	0.84	0.30
27	4	6	0.10	0.97	0.43	5	0.15	0.94	0.87
28	3	5	0.06	0.94	0.40	8	0.12	0.86	0.27
29	7	11	0.35	0.66	0.40	8	0.39	0.61	0.48
30	3	9	0.06	0.94	0.21	7	0.06	0.94	0.26
31	7	15	0.24	0.77	0.55	6	0.37	0.68	0.63
32	3	3	0.16	0.81	0.79	8	0.17	0.87	0.30
33	10	10	0.29	0.72	0.71	5	0.40	0.68	0.81
34	5	8	0.20	0.91	0.81	4	0.28	0.88	0.64
35	4	10	0.04	0.98	0.50	7	0.16	0.85	0.59
36	11	12	0.31	0.72	0.60	8	0.30	0.70	0.75
37	4	7	0.26	0.72	0.39	7	0.40	0.70	0.35
38	3	8	0.14	0.87	0.50	5	0.27	0.71	0.36
39	3	3	0.11	0.83	0.62	7	0.11	0.87	0.24
40	14	6	0.39	0.46	0.68	5	0.42	0.48	0.74
41	6	9	0.24	0.73	0.50	4	0.33	0.59	0.76
42	4	8	0.15	0.86	0.49	6	0.28	0.82	0.46
43	5	7	0.16	0.82	0.54	5	0.26	0.74	0.54

By examining Tables VI and II, it can be verified that the proposed system creates as many clusters as the number of speakers in 8 out of 43 cases, while the baseline system does the same only in 3 cases. In addition, the proposed system produces more clusters than the number of speakers in 30 out of the 43 cases, and less clusters than the number of speakers in 5 cases. The baseline system creates more clusters than the speakers in 21 out of the 43 cases, and delivers less clusters than the speakers in 19 cases. Clustering quality is expected to deteriorate when less clusters than the actual ones are created. On the contrary, clustering quality is expected to improve, when the resulting clusters are equal or a little bit more than the actual ones.

It is worth noting that the eigengap criterion succeeds in predicting the correct number of clusters in more than 8 cases, but the number of the resulting clusters deviates from the predicted number of clusters by the eigengap due to the cluster merging module. Figure 1 depicts the eigenvalue distribution of $\mathbf{I} - \mathbf{L}_{norm}$ for two sound files of 4 and 5 speakers.

VI. DISCUSSION AND CONCLUSION

A novel acoustic speaker diarization system for movie scenes that combines cluster ensembles and the eigengap criterion has been proposed. The system has been evaluated first on 15 movie scenes extracted from 5 movies. Clustering quality has been assessed with respect to the overall classification error, the average cluster purity, and the average speaker purity. The influence of outliers in clustering quality has been studied. The best average classification error approximately equal to 19.7% is not influenced by either the inclusion or the exclusion of the outliers. Accordingly, the experimental findings suggest that the proposed system is able to handle outliers efficiently. In addition, the number of speakers who participate in dialogues is correctly predicted with 66.6% success, while the discriminative power of the system is found to be 60%. The system has been further tested on broadcast news of total duration exceeding the 4 hours from the MDE RT-03 Training Data Speech Corpus. A diarization error rate of 20.2% is reported on average in which the contribution of T_{wrong} is found to be 14.18%. The proposed system performance has been compared to that of a baseline system, which does not exploit the eigengap criterion in cluster ensembles and found to be superior in both cases.

However, there is still room to improve the proposed approach. First, a pre-clustering step that performs gender classification could improve clustering. Second, environmental and channel variations, which affect speaker diarization, deserve further consideration, because, it has been observed that whispers and background noise may deteriorate performance.

ACKNOWLEDGMENT

The authors would like to acknowledge Mr. Alexandros Psaltis' effort in voiced/unvoiced segmentation of the audio files using PRAAT. They would also like to thank the anonymous Reviewers for their constructive criticism, which enabled them to improve the quality of their manuscript, as well as the Associate Editor for having coordinated the revision of the manuscript.

REFERENCES

[1] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in Proc. *1998 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 75-760, Seattle, USA, May 1998.

[2] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," in Proc. *European Conf. Speech Communication and Technology*, pp. 2437-2440, Lisbon, Portugal, September 2005.

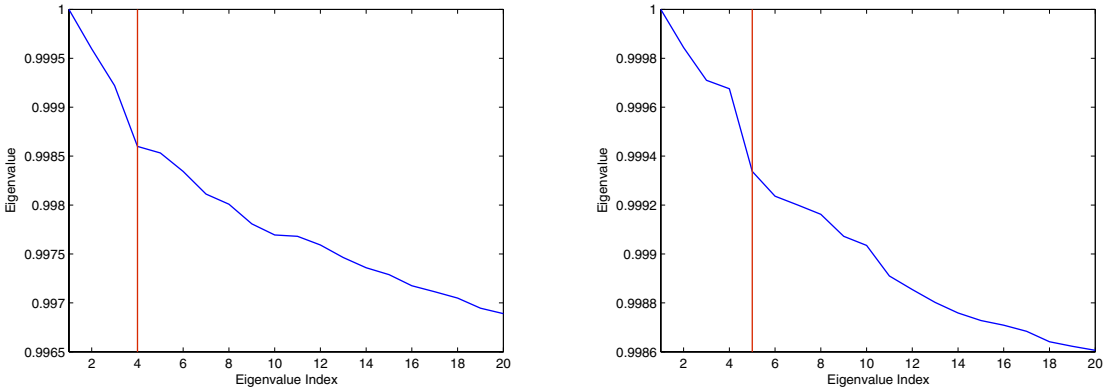

Fig. 1. The top 20 eigenvalues of $\mathbf{I} - \mathbf{L}_{norm}$ for two sound files. The vertical line indicates the actual number of speakers, which coincides with the drastic drop in the magnitude of the eigenvalues.

[3] H. G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond Audio Content Indexing and Retrieval*. Chichester, England: J. Wiley & Sons, 2005.

[4] H. G. Kim and T. Sikora, "Comparison of MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation," in Proc. *2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, pp. 925-928, Montreal, Canada, May 2004.

[5] H. G. Kim and T. Sikora, "Audio spectrum projection based on several basis decomposition algorithms applied togeneral sound recognition and audio segmentation," in Proc. *12th European Signal Processing Conf.*, pp. 1047-1050, Vienna, Austria, September 2004.

[6] M. Kotti, E. Benetos, and C. Kotropoulos, "Automatic speaker change detection with the Bayesian information criterion using MPEG-7 features and a fusion scheme," in Proc. *2006 IEEE Int. Symp. Circuits and Systems*, Kos, Greece, May 2006.

[7] M. Kotti, L. G. P. M. Martins, E. Benetos, J. S. Cardoso, and C. Kotropoulos, "Automatic speaker segmentation using multiple features and distance measures: A comparison of three approaches," in Proc. *2006 IEEE Int. Conf. Multimedia and Expo*, pp. 1101-1104, Toronto, Canada, July 2006.

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, and Language Processing*, pp. 1557-1565, vol. 14, no. 5, September 2006.

[9] D. Liu and F. Kubala, "Online speaker clustering," in Proc. *2004 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, pp. 333-336, Montreal, Canada, May 2004.

[10] W. H. Tsai, S. S. Cheng, and H. M. Wang, "Speaker clustering of speech utterances using a voice characteristic reference space," in Proc. *8th Int. Conf. Spoken Language Processing*, Jeju Island, Korea, October 2004.

[11] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using autoassociative neural networks," *Engineering Applications of Artificial Intelligence*, vol. 22, pp. 667-675, 2009.

[12] S. Meignier, J. F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in Proc. *Odyssey Speaker and Language Recognition Workshop*, pp. 175-180, Crete, Greece, June 2001.

[13] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," in Proc. *7th Int. Conf. Spoken Language Processing*, pp. 573-576, Colorado, USA, September 2002.

[14] S. Meignier, D. Moraru, C. Fredouille, J. F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 303-330, April-July 2006.

[15] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio, Speech, and Language Processing*, pp. 1505-1512, vol. 14, no. 5, September 2006.

[16] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Trans. Speech and Audio Processing*, vol. 14, no 5, pp. 1513-1525, September 2006.

[17] http://www.itl.nist.gov/iad/mig/tests/rt/

[18] D. van Leeuwen, "The TNO speaker diarization system for NIST RT05s for meeting data," in *NIST 2005 Spring Rich Transcrition Evaluation Workshop*, Edinburgh, UK, July 2005.

[19] C. Wooters and M. Huijbregts, "The ICSI RT07S speaker diarization system," in *Multimodal Technologies for Perception of Humans*, vol. LNCS 4625, pp. 509-519, Berlin, Germany: Springer, 2009.

[20] D. A. Reynolds, T. F. Quatiery, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, October 2000.

[21] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no 9, pp. 1437-1462, September 1997.

[22] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Processing*, vol. 88, no. 5, pp. 1091-1124, May 2008.

[23] I. Voitovetsky, H. Guterman, and A. Cohen, "Unsupervised speaker classification using self-organizing maps," in Proc. *IEEE Workshop Neural Networks for Signal Processing*, pp. 578-587, Amelia Island, USA, September 1997.

[24] I. Lapidot, H. Guterman, and A. Cohen, "Unsupervised speaker recognition based on competition between self-organizing maps," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 877-887, July 2002.

[25] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in Proc. *European Conf. Speech Communication and Technology*, pp. 2441-2444, Lisbon, Portugal, September 2005.

[26] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," in Proc. *Speech Recognition Workshop*, pp. 108-111, Chantilly, Virginia, 1997.

[27] H. Gish, M. H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in Proc. *1991 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 873-876, Toronto, Canada, April 1991.

[28] L. Lu and H. Zhang, "Unsupervised speaker segmentation and tracking in real-time audio content analysis," *Multimedia Systems*, vol. 10, no. 4, pp. 332-343, April 2005.

[29] J. -L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in Proc. *5th Int. Conf. Spoken Language Processing*, pp. 1335-1338, Sydney, Australia, December 1998.

[30] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in Proc. *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 411-416, Virgin Islands, November 2003.

[31] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 645 - 648, Seattle, USA, May 1998.

[32] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in Proc. *6th European Conf. Speech Communication and Techology*, pp. 679-682, September 1999.

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

[34] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2/e. San Diego, CA: Academic Press, 1990.

[35] http://www.praat.org

[36] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Institute of Phonetic Sciences*, vol. 17, pp. 97-110, 1993.

[37] M. Kotti, E. Benetos, and C. Kotropoulos, "Computationally efficient and robust BIC-based speaker segmentation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 920-933, July 2008.

[38] G. Almpanidis, M. Kotti, and C. Kotropoulos, "Robust detection of phone segments in continuous speech using model selection criteria with few observations," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 287-298, February 2009.

[39] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, , NY.: John Wiley & Sons, 1990.

[40] L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 6, pp. 835-850, June 2005.

[41] K. Hornik, "A CLUE for CLUster Ensembles," *J. Statistical Software*, vol. 14, no. 12, September 2005.

[42] X. Hu and I. Yoo, "Cluster ensemble and its applications in gene expression analysis," in *ACM Int. Conf. Proc. Series*, vol. 55, pp. 297-302, 2004.

[43] A. Fred, "Finding consistent clusters in data partitions," in *Multiple Classifier Systems*, vol. LNCS 2096, pp. 309-318, Springer, 2001.

[44] S. Dudoit and J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure," *BioInformatics*, vol. 19, no. 9, pp. 1090-1099, 2003.

[45] B. Fischer and J. M. Buhmann, "Bagging for path-based clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1411-1415, 2003.

[46] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2/e. New York, NY.: John Wiley & Sons, 2001.

[47] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI: American Mathematical Society, 1997.

[48] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395-416, 2007.

[49] H. Ning, M. Liu, H. Tang, and T. S. Huang, "A spectral clustering approach to speaker diarization", in Proc. *9th Int. Conf. Spoken Language Processing (ICSLP)*, Pittsburgh, Pennsylvania, 2006.

[50] S. Strassel, C. Walker, and H. Lee, "2004 RT-03 MDE Training Data Speech", Linguistic Data Consortium, Philadelphia, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004S08.

[51] D. Spachos, A. Zlantintsi, V. Moschou, P. Antonopoulos, K. Tzimouli, E. Benetos, M. Kotti, C. Kotropoulos, N. Nikolaidis, P. Maragos, and I. Pitas, "MUSCLE movie-database: A multimodal corpus with rich annotation for dialogue and Saliency Detection," in *Proc. LREC 2008 Workshop on Multimodal Corpora*, Marrakech, Morocco, May 26-27, 2008.

[52] NIST, "Rich Transcription 2004 Spring Meeting Recognition Evaluation Plan", http://www.itl.nist.gov/iad/mig/tests/rt/2004-spring/.

Nikoletta Bassiou was born in Larissa, Greece in 1977. She received the B.Sc. degree in Informatics from the Aristotle University of Thessaloniki, Greece in 2000.

Currently, she is pursuing the Ph.D. degree at the Artificial Intelligence and Information Analysis Lab of the Department of Informatics at the Aristotle University of Thessaloniki, Greece. Her current research interests lie in the areas of audio, speech and language processing, signal processing, pattern recognition, and information retrieval.

Vassiliki Moschou received the B.Sc. degree in Informatics in 2005 and the M.Sc. degree in Digital Media in 2007, both from the Aristotle University of Thessaloniki. Her research interests include audio and speech processing, speaker recognition, and movie analysis.

Constantine Kotropoulos was born in Kavala, Greece in 1965. He received the Diploma degree with honors in Electrical Engineering in 1988 and the PhD degree in Electrical & Computer Engineering in 1993, both from the Aristotle University of Thessaloniki.

He is currently an Associate Professor in the Department of Informatics at the Aristotle University of Thessaloniki. From 1989 to 1993 he was a research and teaching assistant in the Department of Electrical & Computer Engineering at the same university. In 1995, he joined the Department of Informatics at the Aristotle University of Thessaloniki as a senior researcher and served then as a Lecturer from 1997 to 2001 and as an Assistant Professor from 2002 to 2007. He was a visiting research scholar in the Department of Electrical and Computer Engineering at the University of Delaware, USA during the academic year 2008-2009 and he conducted research in the Signal Processing Laboratory at Tampere University of Technology, Finland during the summer of 1993. He has co-authored 42 journal papers, 147 conference papers, and contributed 6 chapters to edited books in his areas of expertise. He is co-editor of the book "Nonlinear Model-Based Image/Video Processing and Analysis" (J. Wiley and Sons, 2001). His current research interests include audio, speech, and language processing; signal processing; pattern recognition; multimedia information retrieval; biometric authentication techniques, and human-centered multimodal computer interaction.

Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a senior member of the IEEE and a member of EURASIP, IAPR, and the Technical Chamber of Greece. He is a member of the Editorial Board of Advances in Multimedia journal and serves as a EURASIP local liaison officer for Greece.