International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 26, No. 8 (2012) 1260011 (22 pages)

DOI: 10.1142/S0218001412600117

World Scientific
www.worldscientific.com

ONLINE SPEAKER DIARIZATION FOR MULTIMEDIA DATA RETRIEVAL ON MOBILE DEVICES

KYUNG-MI PARK*, JEONG-SIK PARK†, JAE-HYUN BAE‡
and YUNG-HWAN OH§

Computer Science Department
Korea Advanced Institute of Science and Technology
Daejeon, South Korea
**kmpark@speech.kaist.ac.kr*
†parkjs@mokwon.ac.kr
‡jhbae@speech.kaist.ac.kr
§yhoh@speech.kaist.ac.kr

Received 30 November 2010
Accepted 23 July 2012
Published 9 January 2013

Speaker diarization detects speaker change points in spoken data and organizes speaker clusters so that each cluster contains one speaker's segments. This study aims to develop online speaker diarization for multimedia data retrieval on mobile devices. Researchers have proposed various methods of diarization, but most approaches thus far depend on an empirically determined threshold as a criterion or work in an offline manner that requires prior knowledge, such as the overall number of speakers. There are therefore clear drawbacks with mobile devices, on which various types of spoken data are frequently played and replaced. A new approach to online speaker segmentation and clustering is proposed for overcoming these drawbacks. The proposed segmentation method considers the temporal locality of an analysis window, assuming that each window contains only a small number of speakers. In accordance with this property, a local universal background model (UBM) is constructed in a window and the model is used to detect speaker change points. A cluster boundary-based dynamic decision criterion is proposed for speaker clustering. This approach estimates the internal characteristics of clusters and uses them to determine cluster boundaries. In experiments using a broadcast news corpus, our techniques exhibited superior performance compared to conventional approaches.

Keywords: Online speaker diarization; speaker segmentation; speaker clustering; multimedia data retrieval; universal background model.

1. Introduction

As mobile devices such as smart phones, tablet PCs, and personal memory aids have become hugely popular, users of the devices can easily download and enjoy various

†Corresponding author, J.-S. Park is currently with the Department of Intelligent Robot Engineering, Mokwon University, South Korea.

types of multimedia data, including movies, UCCs, and mobile TV, anytime and anywhere. The effortless access to a tremendous amount of data has encouraged the use of multimedia data retrieval techniques for managing and searching for data on the devices. Text-driven retrieval approaches have achieved satisfactory performance in these tasks, as users can search for data directly using specific text information such as subtitles. However, there are clear limitations when retrieving data for which no text information is provided.

Various technical approaches have recently been introduced in an effort to cope with the limitations of text-driven retrieval, such as scene analysis and image tracking/recognition. Such image processing techniques were reported to be capable of searching for video-format data, but they do not work with audio-format data. Consequently, speech-driven retrieval (known as "spoken document retrieval") is a preferred method when multimedia data contain a sufficient number of audio streams. In particular, speech information can provide a better understanding of multimedia data content and even the personalities of the characters. For this reason, its application covers most types of multimedia data, including broadcast news, talk shows, and movies, which typically contain speech sequences.

Spoken document retrieval extracts speaker information in conjunction with content information. Speaker information plays no less an important role than content information in the retrieval process. This information allows users to successfully search for data spoken by a specific speaker and facilitates the arbitrary editing of documents with respect to a speaker. Furthermore, the speaker information makes it possible to employ speaker adaptation techniques during the speech recognition procedures, thereby contributing to the acquisition of more accurate content information from speech. Document summarization and audio indexing also require speaker information. Accordingly, speaker diarization, in which speaker-related information is extracted from spoken data, is a core technology in the field of spoken document retrieval.

Speaker diarization, also designated as speaker indexing, detects speaker change points in a document (this task is commonly referred to as "speaker segmentation") and organizes speaker clusters so that each cluster contains a single speaker's segments (in a process known as "speaker clustering"). In short, speaker diarization concentrates on replying to the question, "Who spoke when?", whereas speech recognition obtains text information by answering the question, "What did the speaker say?"[8,10]

Both speaker segmentation and speaker clustering tasks greatly affect diarization performance. Most proposed approaches to these tasks use an empirically determined static threshold as a decision criterion and/or work on an offline system requiring prior information pertaining to the speakers, such as the number of speakers and the identities of those involved in the document. For these reasons, they perform poorly when working on mobile devices, in which an immense variety of spoken documents is played and then typically replaced with new ones. This study targets the

development of an online speaker diarization system that will work on mobile devices without any prior information, while reflecting dynamic changes in the acoustic characteristics.

This paper is organized as follows. Speaker diarization and its problems in mobile applications are introduced in Sec. 2. Next, the proposed online speaker segmentation and clustering approaches are described in Sec. 3. Experimental results are provided in Sec. 4. Finally, this paper is concluded in Sec. 5.

2. Previous Work on the Subject of Speaker Diarization

The main goal of speaker diarization is to split spoken documents into homogeneous segments and clusters so that every segment and cluster contains a single speaker's speech data. Thus, many researchers have concentrated their efforts on speaker segmentation and clustering.[5–7,20]

2.1. *Conventional speaker segmentation and clustering*

Speaker segmentation detects speaker change points from speech data in accordance with a decision criterion with which speaker change is determined. The most popular criterion used is the Bayesian Information Criterion (BIC), which has been characterized as better suited for the task of detecting speaker change points owing to a precise estimate of the acoustic dissimilarity between different speakers.[2,4,7,12] The conventional BIC-based speaker segmentation scheme initially splits specified multimedia data in a regular size (here, speech data of a regular size is referred to as an "analysis window") and constructs a single Gaussian model (SGM) for each of the two neighboring speech streams divided from an analysis window. The dissimilarity between the two models is estimated according to the BIC principle.

Speaker clustering serves to divide the respective speaker segments obtained from the segmentation procedure into corresponding speaker clusters, each of which contains the segments of a single speaker.[1] Hierarchical clustering has been broadly used to accomplish this, as it automatically generates clusters based on a distance matrix without predetermining the number of clusters.[5,16,20] This approach maintains a distance matrix by which the distance between all pairs of clusters is estimated. Relatively adjacent clusters are then merged into a single cluster. This procedure is repeated in an iterative manner until the distance between the closest clusters becomes larger than a predetermined threshold.[6,18]

2.2. *Problems of conventional approaches in mobile applications*

BIC-based speaker segmentation and hierarchical clustering have been successfully applied to speaker diarization. However, they do not necessarily assure the best performance for the variety of multimedia data played on mobile devices. The BIC-based segmentation scheme occasionally fails to detect speaker changes for short

speech segments, as the SGM may not correctly describe the acoustic speaker characteristics of its corresponding segment due to the lack of training data. The conventional hierarchical clustering method depends highly on empirically determined thresholds when estimating distances between clusters. Such a static threshold may undoubtedly lead to incorrect decisions during clustering if the application data is acoustically different from the training data used during the determination of the threshold. Moreover, several approaches have performed diarization procedures in an offline manner that requires prior knowledge, such as the overall number of speakers and even their identities.

To take mobile multimedia data tendencies into account, speaker diarization must be processed in an online manner. Online speaker diarization processes diarization procedures on respective sets of speech streams corresponding to an analysis window. On the other hand, an offline method runs through the entire diarization process at one time and starts the procedures only after all of the data is completely prepared.

Several studies dealt with online speaker diarization issues. An approach applied an unsupervised adaptive learning for the online dairization, but this technique is dependent upon the threshold in the detection of unregistered speakers.[13,14] Another studies introduced a new diarization approach employing multimodal knowledge sources.[15,19] However, this technique requires well-detected visual information and its performance needs to be more carefully analyzed using a sufficient number of audio-visual data. In particular, these studies targeted at meeting data in which all participants can be correctly detected. Thus, we wonder if the approaches are applicable to mobile mutimedia data including movies and broadcast news. A recent study combines a traditional offline system with an online speaker identification system.[22] Although this approach deals with some problems of the conventional online diarization approaches, the performance of the online system is likely to depend on the correctness of the speaker labels information obtained from the offline process. In addition, the system needs to be verified with various kinds of multimedia data in which a number of speakers are included.

These drawbacks of the conventional approaches may induce abnormal speaker diarization results on mobile devices. In these devices, various multimedia data deliver spoken documents while retaining the diverse characteristics of various speakers. In particular, most of them contain an undetermined number of speakers, except for a few categories of data such as a public address or a forum. For this reason, it is necessary to devise a more sophisticated diarization approach that considers the dynamic characteristics of multimedia data.

3. Online Speaker Diarization Based on the Dynamic Characteristics of Multimedia Data

In this section, we propose new approaches to online speaker segmentation and clustering that consider the dynamic characteristics of multimedia data.

3.1. *Online speaker segmentation based on local UBM adaptation*

The use of a Gaussian distribution is a typical strategy for describing the acoustic characteristics of a speaker in a segment. The Gaussian mixture model (GMM) framework can be a suitable substitute for the SGM, as it can more precisely represent the diverse characteristics of each speaker. However, constructing a reliable GMM from the short speech stream of an analysis window is not feasible.

An adaptation technique can be utilized to construct a GMM from a small amount of speaker data, as this technique has been successfully applied to many tasks, including speaker identification.[12,16] The first step of this approach is to construct a generalized GMM designated as the Universal Background Model (UBM), from a sufficient amount of data (known as "UBM data"), including the speech of various speakers. A relatively small amount of data (termed "adaptation data") spoken by a single speaker is then adapted to the UBM in accordance with an adaptation algorithm. The adapted model constructed in this manner is described as a GMM distribution. It is known as an "adapted GMM". In particular, it reflects the acoustic characteristics of the speaker corresponding to the adaptation data in a more sophisticated manner than the use of a GMM directly constructed from a small amount of data.

UBM-based GMM adaptation will greatly contribute to speaker segmentation on mobile devices by describing the characteristics of the rapidly changing speakers in multimedia data. However, if the types of speakers or the environmental characteristics of the UBM data are completely irrelevant to those of the adaptation data, the adapted GMM may be excessively subordinate to the UBM data and ignore the characteristics of the adaptation data. The ideal approach is to organize the respective UBM using similar types of adaptation data. However, this is not practical for mobile multimedia data that are continuously updated. In this study, we propose an approach to constructing more reliable UBM data and using them during the speaker segmentation procedure.

3.1.1. *Local UBM adaptation based on temporal locality*

Every short part of an audio stream in multimedia data contains only a small number of speakers. The speakers have their own speech boundaries inside the part, assuming that speakers do not speak simultaneously. Based on this type of temporal locality of the speakers, a GMM is constructed from each analysis window and the model is used in the GMM adaptation process in place of the conventional UBM. Therefore, we designate this GMM as a "local UBM". Speech data used in the construction of a local UBM are directly employed in the adaptation process; thus, UBM data retain directly relevant characteristics of adaptation data. As a result, the model adapted from the local UBM more correctly describes the acoustic characteristics of short speech streams as a GMM distribution.

Figure 1 illustrates the proposed GMM adaptation using the *local UBM*. First, we construct a local UBM from speech data of an analysis window. Assume that each

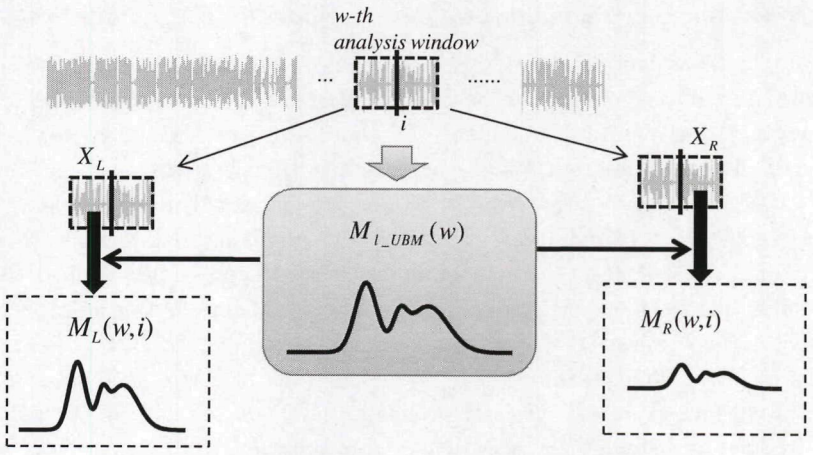


Fig. 1. Adapted GMM construction from local UBM (color online).

analysis window consists of N speech frames. Let us denote $M_{l\_\mathrm{UBM}}(w)$ as the wth local UBM constructed in the wth analysis window. Then, we divide the speech data of the window into two sets of speech streams, X_L $(= x_k|k = 1, \ldots, i)$ and X_R $(= x_k|k = i+1, \ldots, N)$, on the basis of a candidate change point, i. Once each of the pair of speech streams is adapted to $M_{l\_\mathrm{UBM}}(w)$, two adapted GMMs, $M_L(w,i)$ and $M_R(w,i)$, are obtained, respectively.

3.1.2. *Online speaker segmentation using BIC*

The ultimate goal of our approach is to apply the adapted GMM constructed from the local UBM to the conventional BIC procedure and perform online speaker segmentation. Figure 2 describes the proposed online speaker segmentation conducted in the wth analysis window. This procedure is repeatedly performed on respective analysis windows, which are regularly split and sequentially shifted over audio streams of given multimedia data.

Firstly, a local UBM is constructed from speech data in a given window. The second step is to construct a pair of adapted GMMs on the basis of a given hypothesized speaker change point i, as described in Fig. 1. Next, a BIC value for a given i, $\mathrm{BIC}_w(i)$, is estimated using the following equations:

$$\mathrm{BIC}_w(i) = D_w(i) - \frac{\lambda}{2}\left(p + \frac{1}{2}p(p+1)\right) \times \log N, \tag{1}$$

$$D_w(i) = \sum_{k=1}^{i} \log P(x_{w,k}|M_L(w,i)) + \sum_{k=i+1}^{N} \log P(x_{w,k}|M_R(w,i)) - \sum_{k=1}^{N} \log P(x_{w,k}|M_{l\_UBM}(w,i)), \tag{2}$$

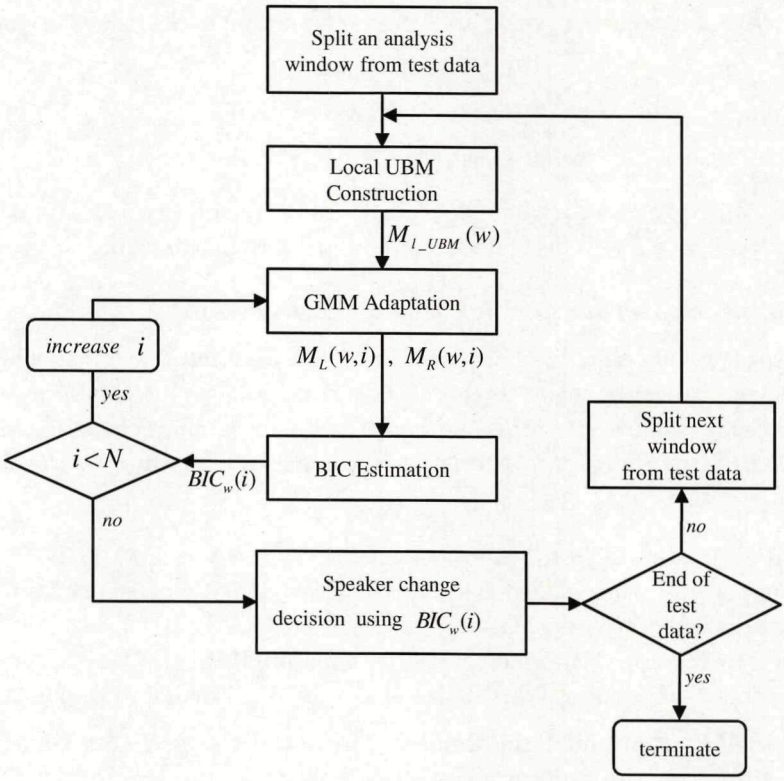


Fig. 2. Online speaker segmentation based on local UBM.

where λ is a penalty factor and p indicates the feature vector dimensions. Here, each of two equations was derived from the conventional BIC principle.[7] The process is repeated from the second step and respective BIC values, $\mathrm{BIC}_w(i)$, are estimated while sequentially shifting i towards the last frame in the analysis window. If the point i reaches the last frame in its corresponding window, a speaker change point is finally determined using $\mathrm{BIC}_w(i)$, where $i = 1, \ldots, N-1$.

In Eq. (2), $D_w(i)$ is the dissimilarity between the log-likelihoods of each of the two divided speech streams on their respective adapted GMMs and that of the overall speech streams on the local UBM. If both of the adapted models demonstrate similar distributions to the local UBM, $D_w(i)$ is close to 0, indicating that no speaker change occurs at a point i; otherwise, $D_w(i)$ becomes far from 0 and that increases the value of $\mathrm{BIC}_w(i)$.

The conventional BIC-based criterion detects at most one change point in an analysis window.[7] This detection may result in the loss of some change points in several types of multimedia data in which speakers are rapidly changed and their speaking duration is relatively short. In contrast, we take more generalized cases of mobile multimedia data into account. Without restrictions on the number of change

points, zero or more change points are detected according to the following conditions:

$$\begin{cases} \mathrm{BIC}_w(i) > 0 \\ \mathrm{BIC}_w(i) - \mathrm{BIC}_w(i-1) > 0 \\ \mathrm{BIC}_w(i+1) - \mathrm{BIC}_w(i) < 0. \end{cases} \tag{3}$$

In Eq. (3), every point, i, where the value of $\mathrm{BIC}_w(i)$ indicates a local maximum as well as a positive value is determined to be a speaker change point.

3.2. *Online speaker clustering based on relative GLR*

Online speaker clustering provides an answer to a question that arises whenever a new speaker segment appears: "Is this new segment included in an existing cluster or in a new cluster?" To address this question, a well-known online clustering technique (known as "leader-follower clustering") is conducted according to the following procedures.[11]

(1) Initialize a threshold to determine whether to merge or not.
(2) Compute the Generalized Likelihood Ratio (GLR) between each of existing clusters and a new segment.
(3) Find the closest cluster indicating the smallest GLR.
(4) Compare the smallest GLR with the threshold calculated in step (1).
    - If GLR $<$ threshold, the segment is merged with the closest cluster.
    - Otherwise, the segment creates a new cluster.
(5) Repeat steps (2)–(4) until all segments find their positions.

As described in this procedure, most speaker clustering approaches mainly depend on a predetermined threshold during segment categorization. Although such a mechanism is simple to operate and requires a small amount of computation, the threshold must be re-estimated for different types of multimedia data. The determination of a correct threshold value requires a sufficient amount of speech data. For this reason, this is certainly not a practical approach in online speaker diarization, in which segmentation/clustering procedures are sequentially applied to each analysis window. In order to solve this problem, we propose a new online clustering mechanism based on GLR.

3.2.1. *Problems associated with the conventional GLR-based decision criterion*

GLR is a well-known criterion that is used in the estimation of acoustic dissimilarity between two speech segments. The speech segments of the ith cluster and the jth cluster are denoted as X_{c_i} and X_{c_j}, respectively. The GLR between these clusters is computed as follows:

$$\mathrm{GLR}(X_{c_i}, X_{c_j}) = \frac{P(X_{c_i}|M_{c_i})P(X_{c_j}|M_{c_j})}{P(X_{c_{i+j}}|M_{c_{i+j}})}. \tag{4}$$

Here, M_{c_i}, M_{c_j}, and $M_{c_{i+j}}$ refer to GMMs constructed from X_{c_i}, X_{c_j} and the overall set of segments, respectively. In general, $\mathrm{GLR}(X_{c_i}, X_{c_j})$ has a higher value, as the dissimilarity between two clusters becomes larger. The conventional GLR-based clustering approach estimates the GLR between a cluster and a new segment and uses it as an implicit criterion.

Speaker clustering aims to maintain some distance between each cluster while preserving consistency within each cluster. In other words, the segments included in a cluster should retain characteristics that are acoustically similar to each other. Although the conventional GLR effectively estimates the dissimilarity between a cluster and a segment, the use of this criterion can lead to incorrect decisions regarding segment clustering, as it ignores the internal consistency of a cluster. An example is given below. If a new segment s acoustically preserves more similar characteristics from segments in C_k than those in other clusters, $\mathrm{GLR}(X_{c_k}, s)$ indicates a lower value. In this case, this segment can be a member of C_k in accordance with the conventional criterion. However, this criterion may confirm that s definitely retains the corresponding characteristics for inclusion in C_k, despite the fact that the segment may damage the internal characteristics of C_k compared to all of the segments in the cluster, as shown in Fig. 3. This undesirable addition may worsen the consistency of this cluster, thus generating incorrect clustering results. For this reason, we carefully expand the decision criterion to consider the internal characteristics of clusters.

3.2.2. *Internal GLR and relative GLR: New decision criteria for online speaker clustering*

To measure the internal consistency of a cluster, the average GLR between the segments belonging to the cluster is calculated and this value is designated as the "*internal GLR*". Figure 4 illustrates the fundamental internal GLR concept. It is assumed here that a cluster, C_k, contains N_k segments. The internal GLR in this cluster is obtained using the following equation:

$$\begin{aligned}\mathrm{GLR}_{\mathrm{int}}(C_k) &= \frac{1}{N_k}\sum_{i=1}^{N_k}\mathrm{GLR}(X^c_{c_{k,i}}, x_{c_{k,i}}) \\ &= \frac{1}{N_k}\sum_{i=1}^{N_k}\log\frac{P(X^c_{c_{k,i}}|M_{X^c_{c_{k,i}}})P(x_{c_{k,i}}|M_{x_{c_{k,i}}})}{P(X_{c_k}|M_{c_k})},\end{aligned} \tag{5}$$

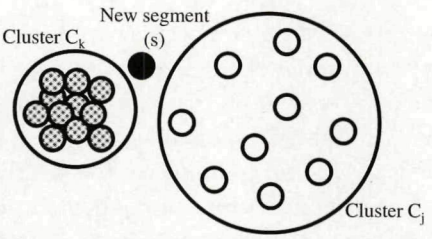


Fig. 3. Problems associated with the conventional GLR-based decision criterion.

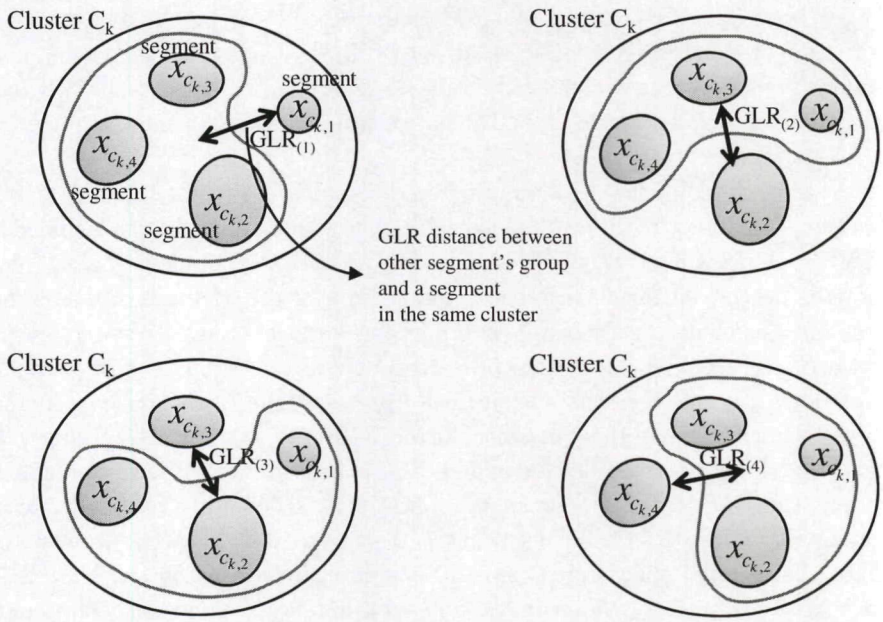


Fig. 4. Internal GLR to represent intra-cluster characteristics (color online).

where $x_{c_{k,i}}$ and $X^c_{c_{k,i}}$ denote the ith segment of C_k and a set of the other segments except for $x_{c_{k,i}}$, respectively.

$\mathrm{GLR}_{\mathrm{int}}(C_k)$ represents the consistency of the cluster C_k. The GLR between respective pairs of segments has a higher value when respective segments in a cluster retain inconsistent characteristics relative to those of other segments; otherwise, a lower value of $\mathrm{GLR}_{\mathrm{int}}(C_k)$ results.

The use of internal GLR serves to estimate the dissimilarity between a cluster and a new segment. To do this, a new criterion called "relative GLR" is proposed. This is explained below:

$$\mathrm{GLR}_{\mathrm{rel}}(C_k, s) = \frac{\mathrm{GLR}(C_k, s) - \mathrm{GLR}_{\mathrm{int}}(C_k)}{\mathrm{GLR}_{\mathrm{int}}(C_k)}. \tag{6}$$

Relative GLR refers to the relative difference between conventional GLR ($\mathrm{GLR}(C_k, s)$) and internal GLR ($\mathrm{GLR}_{\mathrm{int}}(C_k)$). The relative GLR of the segment is estimated for each cluster whenever a new segment, s, seeks an appropriate cluster. This criterion considers the internal characteristics of the cluster C_k as well as the dissimilarity between the new segment and the cluster in order to determine whether the addition of s is actually suitable for C_k, while preserving the internal consistency of the cluster. If a segment satisfactorily preserves the discriminative characteristics of a cluster in comparison to the overall set of segments in the cluster, the conventional GLR takes on a low value and becomes close to or lower than the internal GLR value of the cluster, decreasing the relative GLR value. On the other hand, if the

segment retains characteristics that are different from those in the cluster, the difference between the conventional GLR and the internal GLR increases. These relative GLR tendencies overcome the drawbacks of the conventional GLR criterion. Let us assume that a segment, s, has the highest value in the conventional GLR of the cluster C_k. In accordance with our criterion, even if s scarcely retains the internal characteristics of a cluster compared to all of the segments in the cluster, the relative GLR becomes higher on C_k and prevents the addition of the segment to C_k.

3.2.3. *Dynamic decision criterion considering the cluster boundaries*

Once a respective value of the relative GLR is estimated from each cluster, a cluster indicating the smallest value becomes a candidate cluster of a new segment. The next procedure is to decide whether to merge the segment into the cluster or to create a new cluster. We propose a new decision criterion for this decision, instead of using a predetermined threshold.

Let assume that s and C_k refer to the new segment and its candidate cluster, respectively, and x is the nearest segment to C_k among all segments in other clusters. If the distance between s and C_k is smaller than that between x and C_k, s can be a member of C_k; otherwise, the segment is encouraged to become the first member of a new cluster. According to this assumption, $\mathrm{GLR}_{\mathrm{rel}}(C_k, x)$ is used as a decision criterion. In short, if $\mathrm{GLR}_{\mathrm{rel}}(C_k, x)$ is larger than $\mathrm{GLR}_{\mathrm{rel}}(C_k, s)$, s is included in C_k. Otherwise, a new cluster is created to include s. In this approach, $\mathrm{GLR}_{\mathrm{rel}}(C_k, x)$ is performed as an extended boundary of C_k, in which C_k is able to contain segments. In other words, if s is located within this boundary, this segment becomes a member of C_k.

This mechanism is expected to facilitate the correct decision on whether or not to merge the segment into the cluster. However, this decision may unconditionally include the new segments in an existing cluster without creating new clusters if a small number of segments and clusters exist, especially in the beginning of clustering. To resolve this problem, the scale of the extended boundary of C_k is reduced by multiplying $\mathrm{GLR}_{\mathrm{rel}}(C_k, x)$ by a value between 0 and 1. This reduction prevents a segment regarded as an outlier from belonging to the cluster. This segment is included in a new cluster instead.

Based on this concept, the proposed criterion is summarized as follow:

$$\mathrm{boundary}(C_k) = w \cdot \min\{\mathrm{GLR}_{\mathrm{rel}}(C_k, x_i)\} \quad (i = 1, \ldots, N), \tag{7}$$

where x_i refers to the ith segment among N segments included in clusters except for C_k and w is a scale factor ranging from 0 to 1.

The first step is to estimate the relative GLR between C_k and each segment of x. A segment indicating the smallest relative GLR corresponds to the nearest segment to C_k. Let us denote x_n as the nearest segment. As illustrated in Fig. 5, $\mathrm{GLR}_{\mathrm{rel}}(C_k, x_n)$ represents the maximum boundary of C_k. But, we use $w \cdot \mathrm{GLR}_{\mathrm{rel}}(C_k, x_n)$ as an extended boundary. A decision on whether or not a new segment is included in C_k is

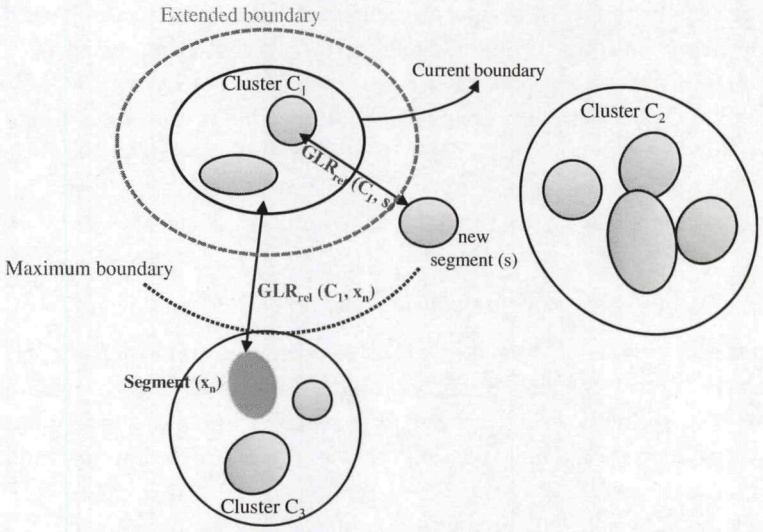


Fig. 5. Dynamic threshold estimation based on relative GLR (color online).

made on the basis of boundary (C_k). The conventional threshold is determined only once from a certain amount of speech data. On the other hand, this boundary-based criterion is continuously updated whenever a new segment appears, therefore contributing to making a correct decision on the feasibility of merging the segment into a cluster.

3.2.4. *Online speaker clustering procedure*

Figure 6 describes the proposed online speaker clustering procedure. The following steps are repeatedly processed for respective speaker segments detected from the online speaker segmentation process based on our mechanism.

Step 1. A relative GLR between a new segment (s) and each of existing clusters is estimated (relevant to Eq. (6)).
Step 2. A cluster (C_k) indicating the smallest relative GLR is decided as a candidate cluster of s.
Step 3. A dynamic decision criterion (boundary(C_k)) is determined from each of segments included in clusters except for C_k (relevant to Eq. (7)).
Step 4. Compare the smallest relative GLR ($\mathrm{GLR}_{\mathrm{rel}}(C_k, s)$) with boundary($C_k$). If $\mathrm{GLR}_{\mathrm{rel}}(C_k, s)$ is smaller than boundary(C_k), s is included in C_k. Otherwise, a new cluster is created and s becomes the first member of the cluster.

3.3. *Contributions of proposed approaches in online speaker diarization*

The purpose of the proposed online speaker diarization process is to relevantly reflect the dynamic characteristics of various multimedia data played on mobile devices. To

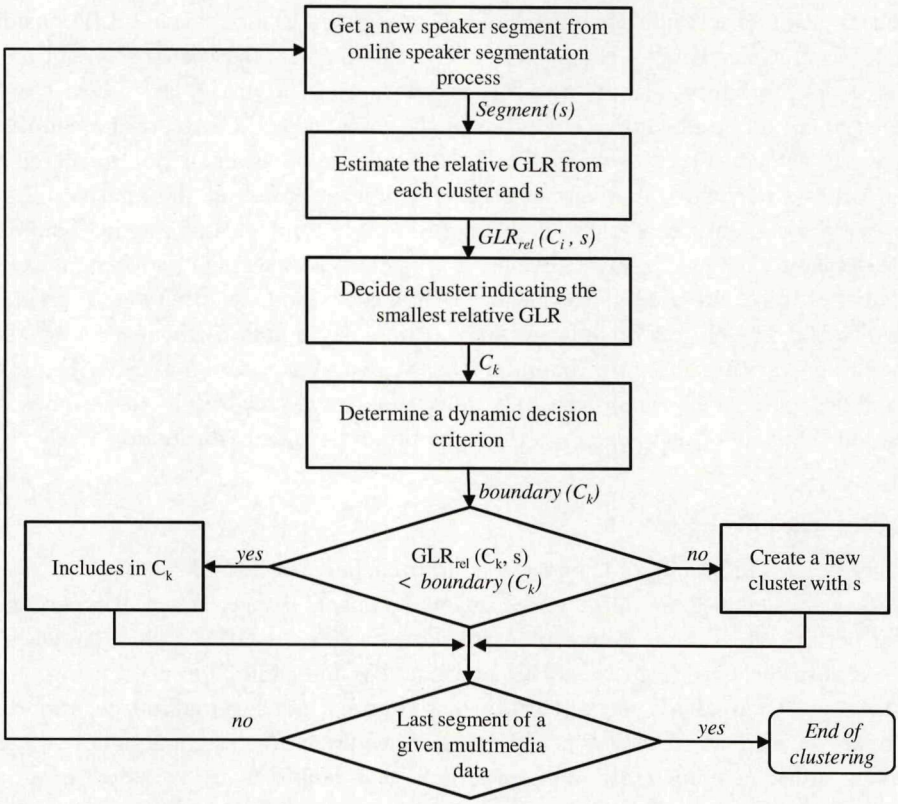


Fig. 6. Online speaker clustering procedure based on relative GLR.

satisfy this requirement, we apply a local UBM-based adaptation and a dynamic decision criterion to online speaker segmentation and clustering, respectively, which are representative and essential procedures in speaker diarization tasks.

In the online speaker segmentation process, a local UBM is constructed in each analysis window and two adapted models are then obtained from the UBM and two neighboring speech streams divided in the window. The local UBM sufficiently retains the acoustic characteristics of speakers included in the corresponding window, thus more precisely representing the diverse characteristics of each speaker. Therefore, when two neighboring streams are speech data derived respectively from two different speakers, each of the adapted models represents more discriminative characteristics in a Gaussian distribution than those of the models derived from conventional approaches.

The proposed online speaker clustering process utilizes a dynamic decision criterion estimated by the internal GLR and the relative GLR. The internal GLR represents the internal consistency of a cluster, which is derived from the average GLR between the segments belonging to a cluster. The relative GLR is estimated from the internal GLR and the conventional GLR. It plays a role as a criterion in the

determination of a candidate cluster of a new segment. The relative GLR considers the internal consistency of clusters as well as the distance between a segment and a cluster, so it produces a more accurate estimation of the dissimilarity between the new segment and each cluster than that of the conventional GLR. After a candidate cluster is determined, a decision should be made on whether or not to merge the segment in the cluster. A dynamic decision criterion based on the relative GLR is employed for making this decision. This criterion determines an extended boundary of the cluster, using the nearest segment to the cluster. A segment can become a new member of the cluster only if it is located within the boundary; otherwise, it creates a new cluster. The cluster boundaries are continuously updated whenever a new segment appears; therefore, the boundary-based decision criterion reflects the characteristics of clusters more precisely and thus more accurately determines the position (cluster) of new segments than the pre-determined threshold.

4. Experimental Results

To verify the efficiencies of the proposed approaches, we carried out several experiments on speaker segmentation and speaker clustering, respectively. All experiments were performed on a well-known audio data corpus, "HUB4", which consists of speech data recorded from broadcast news on TV and radio.[9]

Like most broadcast news, this corpus contains background music and commercials as well as newscasters' speeches. Severely noisy streams such as background music, commercials, and environmental noises were excluded from the corpus in order to concentrate our research on the speaker segmentation and clustering task.

Evaluation data are approximately 1.2 h in duration and consist of speech data spoken by 50 male and female speakers. In addition, they have 335 speaker change points, which refer to an exact number of speaker segments. The length of segments varies from 0.44 s to 67.7 s. To describe speaker characteristics from each speech frame, we extracted a feature vector, which are configured as 12-dimensional MFCCs and their derivatives. All vectors were computed within 25 ms frames.

4.1. *Speaker segmentation experiments*

4.1.1. *Performance measures*

In the standard speaker segmentation tasks, the numbers of change points that were correctly detected (or accepted) and missed (or rejected) is necessarily investigated to evaluate the performance. With respect to this number, several measures are used for the performance evaluation: false alarm rate (FAR), false rejection rate (FRR), and F-measure.[2,4] Each of the measures is calculated as follows:

$$\text{FAR (false alarm rate)} = \frac{\text{number of false acceptance}}{\text{number of total found points}}, \tag{8}$$

$$\text{FRR (false rejection rate)} = \frac{\text{number of false rejection}}{\text{number of total true change}}, \tag{9}$$

$$\text{F-measure} = \frac{2 \times \text{PRC} \times \text{RCL}}{\text{PRC} + \text{RCL}}, \tag{10}$$

where PRC and RCL refer to the precision rate and the recall rate, and they are calculated by (1-FAR) and (1-FRR), respectively. The lower the value of FAR and FRR is, the better the system performance is. The results of FAR tend to be inversely proportional to those of FRR. The F-measure is used to evaluate FAR in company with FRR. The F-measure value ranges from 0 to 1, with a higher value indicating better performance.

4.1.2. *Results and discussions*

Firstly, we investigated the segmentation performance, varying the length of an analysis window, in which a local UBM is constructed, from 10 s to 10 m. Figure 7 illustrates the results. FAR and FRR offered conflicting results with each other. In the length where FAR showed the best performance, FAR presented the worst performance, and vice versa. According to the F-measure, our system achieved the best performance in the length of 15 s. This result explains that in the analysis window of a relatively short length, speech data is so insufficient that a local UBM hardly provides a precise description of the speaker's characteristics. On the other hand, the window of the large size loses the temporal locality. As a result, characteristics of additional speakers are included in a local UBM, thus indicating a relatively non-discriminative UBM.

In the standard speaker segmentation tasks based on BIC, the penalty factor (λ) used in a calculation of BIC influences the segmentation performance. In addition, the number of Gaussian mixtures, which controls the degree of precision in GMM distribution, also affects the accuracy. In the next experiments, we investigated the

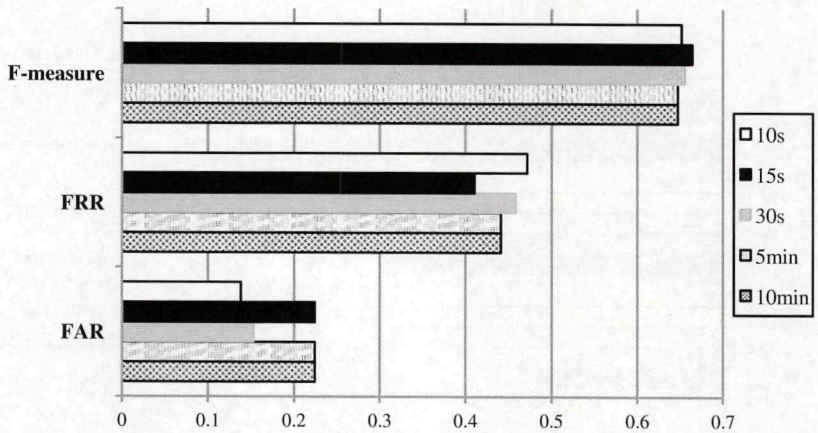


Fig. 7. Speaker segmentation performance according to the length of an analysis window.

performance, while varying the value of λ and the number of mixtures. For the purpose of comparison, we observed the segmentation results according to several types of models describing a speech stream: a SGM, the standard GMM (GMM), the adapted GMM constructed by the conventional UBM adaptation approach (GMM + UBM), and the adapted GMM constructed by the proposed approach (GMM + local UBM). Among them, "SGM" is the standard model framework in the BIC-based segmentation tasks. For the conventional adaptation approach, we constructed a UBM from about 2.3 h of broadcast news including additional data obtained from NIST corpus.[3] In our experiments, we confirmed that the change in the number of mixtures greatly affects the performance. In the standard GMM, two mixtures provided the best result. In contrast, the two kinds of UBM adaptation approaches reached their best performance in eight mixtures. We consider that the reason why the accuracy is not improved in more number of mixtures is related with the model capacity. Provided that the number of mixtures is large, GMM may be trained insufficiently with a small amount of data. In another experiments, we confirmed that the penalty factor also affects the performance. Each approach provides its best performance in respectively different value of λ, since λ is closely associated with the decision criterion. In Fig. 8, we described the performance, when both the number of mixtures and the value of penalty factor provide the best result in the respective approaches. As shown in this figure, the UBM adaptation approaches presented better performance than "SGM" and "GMM". This result explains that the adapted GMMs represent speaker characteristics for a short speech stream more accurately than the SGM or the standard GMM. The reason why the result of "GMM" deteriorated more than that of "SGM" is that GMM is not sufficiently trained with a small amount of speech data. In this figure, our proposed approach yielded superior performance compared to other approaches. More specifically, it demonstrated 31.0% and 8.3% relative improvement on FAR and FRR, respectively, over "SGM". Compared to "GMM + UBM", the "GMM + local UBM" exhibited better performances on FAR and FRR. In addition, the proposed "GMM + local

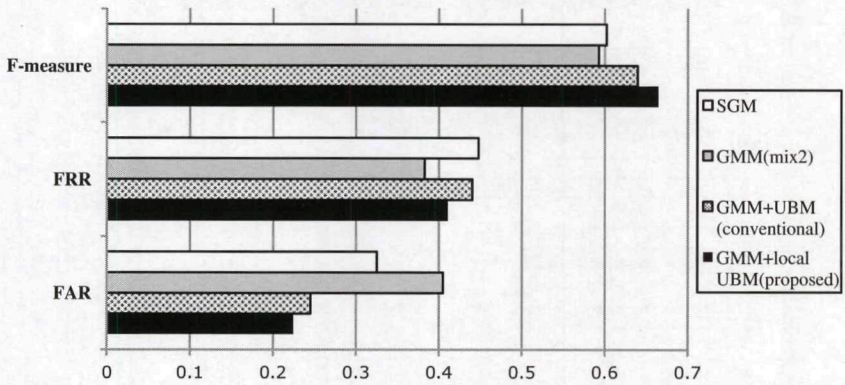


Fig. 8. Speaker segmentation performance according to types of Gaussian models.

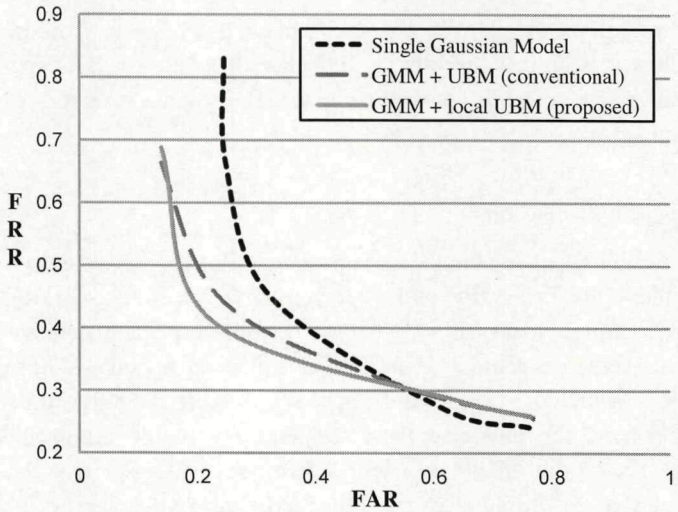


Fig. 9. Speaker segmentation performance via DET curve (color online).

UBM" demonstrated 10.3% and 3.9% relative improvement on F-measure over "SGM" and "GMM + UBM". Although our approach showed slightly inferior than "GMM" on the FRR measure, it presented 44.6% relative improvement on FAR. These results demonstrate that the local UBM-based adaptation provides a more accurate description of speaker characteristics in Gaussian model, thus contributing to more correct detection of speaker changes.

In order to analyze the overall segmentation performance, we investigated the change of FRR according to FAR. This result is given as a DET curve shown in Fig. 9. The proposed approach presented the best performance in the DET curve, demonstrating lower equal error rate (EER) compared to the conventional approach.

4.2. *Speaker clustering experiments*

4.2.1. *Performance measures*

In the standard speaker clustering tasks, the clustering performance is evaluated by the cluster purity and the speaker purity.[10,23] Assume that S speakers contained in a multimedia data is clustered into C groups (clusters). Let us denote n_{ij} and N_{ij} as the number of speaker segments corresponding to a speaker j that are labeled as a cluster i and the number of speech frames contained in n_{ij}, respectively. The two kinds of measures are defined as:

$$\text{Cluster purity} = \frac{\sum_{i=1,j\in[1,S]}^{C} \max(N_{ij})}{\sum_{i=1}^{C}\sum_{j=1}^{S} N_{ij}}, \tag{11}$$

$$\text{Speaker purity} = \frac{\sum_{j=1,i\in[1,C]}^{S} \max(N_{ij})}{\sum_{i=1}^{C}\sum_{j=1}^{S} N_{ij}}, \tag{12}$$

The cluster purity is relevant to the purity of each cluster. It provides a measure of how well a cluster is limited to only one speaker. In contrast, The speaker purity is relevant to the purity of each speaker's cluster. It provides a measure of how well a speaker is limited to only one cluster.

4.2.2. *Results and discussions*

Most of the online speaker clustering approaches demonstrated lower performance than hierarchical clustering that is a representative approach working in an offline manner.[11,23] Although a decision-tree based online approach achieved better performance than the hierarchical approach, it failed to achieve a notable improvement.[23] More specifically, this online approach exhibited 88.9% and 88.5% in the speaker purity and the cluster purity, respectively, while hierarchical clustering demonstrated 88.0% and 87.4% in each of two measures. For this reason, it was a challenge to devise an online clustering method that gives even better results than the representative offline approach.

In Fig. 10, the performance of our online approach employing the relative GLR (RGLR) is compared with that of the conventional hierarchical clustering (HC). We investigated the clustering results according to the number of Gaussian mixtures. In this figure, the results in three or more mixtures were excluded, as the performance of both approaches significantly deteriorated. As shown in this figure, the cluster purity decreases, when the speaker purity increases. Approaches getting nearer to the top right in this figure provides better performance. Our proposed online clustering exhibited outstanding performance in each number of mixtures compared to the

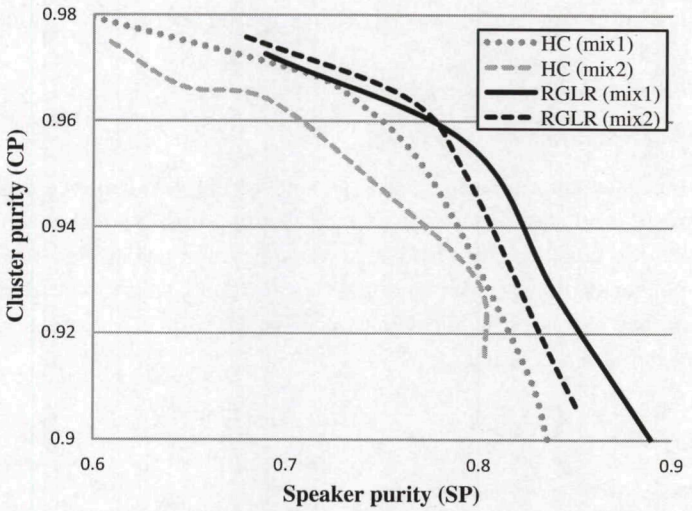


Fig. 10. Speaker clustering performance: Comparison between hierarchical approach (HC) and relative GLR based approach (RGLR; proposed).

Table 1. Comparison proposed clustering with HC.

Clustering Method	Speaker Purity (%)	Cluster Purity (%)
HC	83.5	90.0
RGLR	89.2	89.9

conventional approach, although the proposed algorithm required less information than hierarchical clustering. This result explains that the relative GLR estimates the dissimilarity between a segment and a cluster more accurately than the conventional GLR and this dynamic criterion successfully overcomes the limitations of the predetermined threshold. It is interesting to observe that a single Gaussian distribution provided better clustering results than GMM with two mixtures. This result is related with model capacity, which means relatively short segments cannot construct reliable GMMs preserving two or more mixtures in Gaussian distribution, rather they are more suitable for a SGM.

To observe the clustering performance more explicitly, we investigated the speaker and cluster purities when both measures demonstrate the best performance. As presented in Table 1, our online approach exhibited 6.8% relative improvement in speaker purity over hierarchical clustering. This improvement is an outstanding result compared to the decision-tree based conventional approach, which exhibited a relative improvement of only one percent over HC.[23]

In the next experiment, we verified the performance of online speaker diarization. While the first experiment described in Fig. 10 was carried out using the speaker segments that are correctly divided according to speakers, the next experiment was performed using the segments obtained from our online speaker segmentation process. The segments used are relevant to the segmentation results presenting the highest F-measure described in Fig. 7.

As demonstrated in Fig. 11, the clustering performance slightly deteriorated due to segmentation error. Nevertheless, the proposed approach provided a stable performance in the speaker diarization. However, the cluster purity of our approach seemed to deteriorate more rapidly than that of the conventional approach. This result is owing to characteristics of the proposed dynamic decision criterion. When the segmentation process produces a number of errors in speaker change detection, the dynamic criterion cannot estimate correct boundaries of clusters, thus worsening the consistency of clusters. In contrast, a static threshold used in HC has less effects on clustering of segments.

Our experimental results demonstrate that the proposed online speaker segmentation and clustering approaches solve two important problems that should be addressed when online speaker diarization process is operated on mobile devices: dynamic characteristics of a broad variety of multimedia data and irregular characteristics of various speakers. To relevantly reflect such characteristics of mobile multimedia, we applied a local UBM-based adaptation and a dynamic decision criterion to online speaker segmentation and clustering, respectively. As a result, the

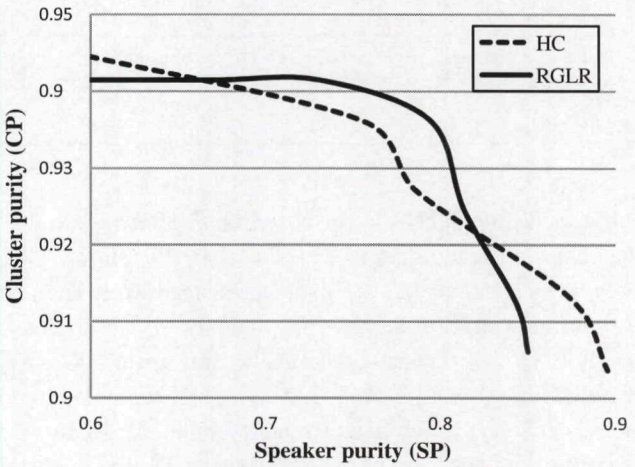


Fig. 11. Speaker segmentation and clustering results: Comparison hierarchical approach (HC) and relative GLR-based approach (RGLR; proposed).

performance of our diarization approach was successfully improved in comparison with that of the conventional approaches.

5. Conclusion

This paper proposed an online speaker diarization approach for multimedia data retrieval on mobile devices. In speaker segmentation process, a local UBM is used in the construction of the adapted GMMs. Since this adapted GMMs retains quite discriminative characteristics of a short speech stream, they operate well in the BIC-based segmentation. In speaker clustering process, the relative GLR-based dynamic decision criterion plays an important role in the estimation of the dissimilarity between a segment and a cluster in conjunction with the estimation of cluster boundaries. Since this criterion is frequently updated, considering the dynamic characteristics of multimedia data, it provides stable clustering performance on mobile devices. In various experiments conducted on a broadcast news corpus, our approaches exhibited superior performance compared to the conventional approaches. In particular, the proposed online clustering technique demonstrated remarkable performance improvement over a representative offline clustering approach. For further verification, we will apply our online approaches to other kinds of multimedia data such as movies, talk shows, and UCCs.

Acknowledgments

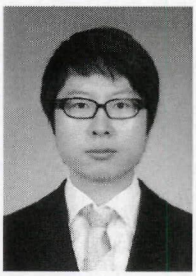
This study was financially supported by the NAP (National Agenda Project) of the Korea Research Council of Fundamental Science & Technology and Defense Acquisition Program Administration and Agency for Defense Development under the contract.

References

1. J. Ajmera and C. Woosters, A robust speaker clustering algorithm, in *Proc. IEEE ASRU* (2003), pp. 411–416.
2. J. Ajmera, I. McCowan and H. Bourlard, Robust speaker change detection, *IEEE Signal Proc. Lett.* **11** (2004) 649–651.
3. M. P. Alvin and A. Martin, NIST speaker recognition evaluation chronicles, in *Proc. Odyssey, The Speaker and Language Recognition Workshop* (2004), pp. 12–22.
4. X. Anguera, XBIC: Real-time cross probabilities measure for speaker segmentation, *ICSI Berkeley Tech. Rep.* TR-05-008 (2005) 1–10.
5. C. Barras, X. Zhu, S. Meignier and J. Gauvain, Improving speaker diarization, in *Proc. DARPA Rich Transcription Workshop* (2004).
6. S. Cassidy, The Macquarie speaker diarization system for RT04S, in *Proc. ICASSP 2004 Meeting Recognition Workshop* (2004).
7. S. S. Chen and P. S. Gopalakrishnan, Speaker, environment and channel change detection and clustering via the Bayesian information criterion, in *Proc. DARPA Broadcast News Transcription and Understanding* (1998), pp. 127–132.
8. P. Delacourt and C. J. Wellekens, DISTBIC: A speaker-based segmentation for audio data indexing, *Speech Commun.* **32** (2000) 111–126.
9. J. S. Garofolo, J. G. Fiscus and W. M. Fisher, Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora, in *Proc. DARPA Speech Recognition Workshop* (1997), pp. 15–21.
10. M. Kotti, V. Moschou and C. Kotropoulos, Review: Speaker segmentation and clustering, *Signal Proc.* **88** (2008) 1091–1124.
11. D. Liu and F. Kubala, Online speaker clustering, in *Proc. IEEE ICASSP* (2004), pp. 333–336.
12. A. S. Malegaonkar, A. M. Ariyaeeinia and P. Sivakumaran, Efficient speaker change detection using adapted Gaussian mixture models, *IEEE Trans. Audio, Speech, Lang. Proc.* **15** (2007) 1859–1869.
13. K. Markov and S. Nakamura, Improved novelty detection for online GMM based speaker diarization, in *Proc. Interspeech* (2008), pp. 363–366.
14. K. Markov and S. Nakamura, Never-ending learning system for on-line speaker diarization, in *Proc. IEEE ASRU* (2007), pp. 699–704.
15. A. K. Noulas and B. J. Krose, On-line multi-modal speaker diarization, in *Proc. ICMI* (2007), pp. 350–357.
16. D. Reynolds, E. Singer, B. Carlson, J. O'Leary, J. McLaughlin and M. Zissman, Blind clustering of speech utterances based on speaker and language characteristics, in *Proc. ICSLP* (1998), pp. 3193–3196.
17. D. A. Reynolds, T. F. Quatieri and R. B. Dunn, Speaker verification using adapted Gaussian mixture models, *Dig. Signal Proc.* **10** (2000) 19–41.
18. H. Sayoud and S. Ouamour, Speaker clustering of stereo audio documents based on sequential gathering process, *J. Inf. Hiding Multimedia Signal Proc.* **1** (2010) 344–360.
19. J. Schmalenstroeer and H. U. Reinhold, Fusing audio and video information for online speaker diarization, in *Proc. Interspeech* (2009), pp. 1163–1166.
20. R. Sinha, S. E. Tranter, M. J. F. Gales and P. C. Woodl, The Cambridge university March 2005 speaker diarisation system, in *Proc. InterSpeech* (2005), pp. 2437–2440.
21. S. Tranter and D. Reynolds, An overview of automatic speaker diarization systems, *IEEE Trans. Audio, Speech, Lang. Proc.* **14** (2006) 1557–1565.
22. C. Vaquero, O. Vinyals and G. Friedland, A hybrid approach to online speaker diarization, in *Proc. Interspeech* (2010), pp. 2638–2641.

23. W. Wang, P. Lv, Q. Zhao and Y. Yan, A decision-tree-based online speaker clustering, *Lecture Notes in Comput. Sci.* **4477** (2007) 555–562.

Kyung-Mi Park received her B.E. degree and Ph.D. in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2011. She is currently a senior engineer at Samsung Electronics. Her research interests include speech recognition, speaker segmentation, speaker diarization, spoken document retrieval, and speaker identification.



Jeong-Sik Park received his B.E. degree in Computer Science from Ajou University, South Korea in 2001 and his M.E. degree and Ph.D. in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2003 and 2010, respectively. From 2010 to 2011, he was a postdoctoral researcher in the Computer Science Department, KAIST. He is now an Assistant Professor in the Department of Intelligent Robot Engineering, Mokwon University. His research interests include speech emotion recognition, speech recognition, speech enhancement, and voice interface for human-computer interaction.

Jae-Hyun Bae received his B.E. degree in Computer Engineering from Kyungpook National University, South Korea in 1998 and his M.E. degree and Ph.D. in Computer Science from KAIST (Korea Advanced Institute of Science and Technology) in 2000 and 2011. He is currently a senior engineer at Samsung Electronics. His research interests include speech synthesis, intention expression, and singing voice generation.

Yung-Hwan Oh received his B.S. and M.S. degrees from Seoul National University, South Korea and his Ph.D. from Tokyo Institute of Technology, Japan, in 1972, 1974, and 1980, respectively. From 1981 to 1985, he was an Assistant Professor in the Computer Engineering Department of Chungbuk National University. He was a visiting research staff at the University of California, Davis, from 1983 to 1984, and a visiting research professor at Carnegie-Mellon University from 1995 to 1996. He is now a professor in the Computer Science Department of KAIST, Daejeon, South Korea. His research interests include speech recognition, speech synthesis, speech coding, and speech enhancement.