International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.02, February 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

# The SAIL Speaker Diarization System for Analysis of Spontaneous Meetings

**Boddeda Ganesh[1], Nemalapuri Sai Kiran[2],Jarugulla Aswani[3]**

[1,2,3]*Assoc.Professor, Department Of CSE , Srivenkateswara College of Engg.,*

## ABSTRACT

*In this paper ,we propose a novel approach to speaker Diarization of spontaneous meetings in our own multimodal SmartRoom environment. The proposed speaker Diarization system first applies a sequential clustering concept to segmentation of a given audio data source, and then performs agglomerative hierarchical clustering for speaker-specific classification (or speaker clustering)of speech segments. The speaker clustering algorithm mutilizes an incremental Gaussian mixture cluster modelling strategy, and a stopping point estimation method based on information change rate. Through experiments on various meeting conversation data of approximately 200minutes total length, this system is demonstrated to provide diarization error rate of 18.90% on average.*

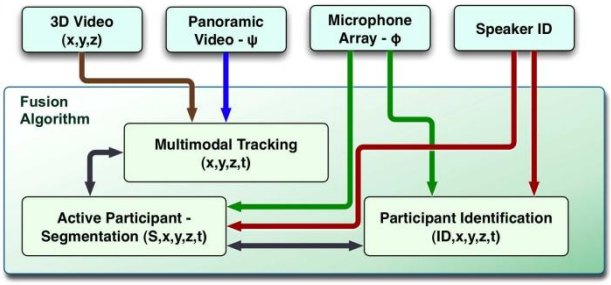
## I. INTODUCTION

Speech Analysis and Interpretation Laboratory(SAIL) at USC has been working on multimodal analysis of spontaneous meeting conversations since it presented its own Smart Room environment[1].The SAIL's Smart Room has four modalities, that is a tracking system using four CCD ceiling cameras, a face detection system by a full-circle Omni directional camera, a circular microphone array with 16 microphones, and a speaker identification system with one directional microphone, as shown in Fig.1.This multimodal conference-room setup was originally intended for real-time localization and tracking of meeting participants, but is also useful for offline post- analysis of the collected data, such as interaction patterns between the participants[2]. Such post-analysis for high-level understanding of given data could be applicable to summarization, classification, and retrieval of spontaneous meetings. From this post-analysis perspective, *speaker diarization*, which refers to the process of automatically transcribing a given audio data source interims of "who spoke when" [3], is important because it can improve the performance of the speaker identification modality in the Smart Room. In this paper, we propose an oval approach to speaker diarization of spontaneous meetings within the framework of SAIL's Smart Room environment.

A variety of state-of-the-art speaker diarization systems, e.g., [4]-[9], have been thus far developed by a number of leading research institutes.

Basic system structure in common: *segmentation*, followed by *speaker clustering*. The former step is to separate speech and non-speech parts in the entire data source given for speaker diarization (*speech*

(a)




(b)

Fig. 1.SmartRoom by SAIL at USC. (a) Video captures from the four ceiling cameras (upper-left side) and the Omni directional camera (upper-right side), and the panoramic transform of the capture from the omnidirectional camera (bottom). (b) Information exchange between all the modalities.

*Activity detection*) and further divide the speech parts into speaker-specific segments by detecting every potential speaker changing point (*speaker change detection*), while the latter step is to classify the resultant speech segments by speaker identity. Keeping this structure as well, our proposed speaker diarization system exploits an oval approach to each step in the structure. This paper is organized as follows. We introduce *a segmentation method based on a sequential clustering concept* in Section III, which follows description of the data sources and the experimental set up used forth experiments reported in the paper in Section II. This segmentation method is based on not only by our real time processing experiences in the Smart Room, but also by our previous work[10]which

TABLE I
TRAINING DATA SET.

| | Source | Name | Length (min:sec) | No. of Speakers |
|---|---|---|---|---|
| 1 | ICSI | Bmr018 | 20:01 | 7 |
| 2 | ICSI | Bro003 | 20:00 | 7 |
| 3 | ICSI | Bsr001 | 20:00 | 8 |
| 4 | NIST | 20020214 | 19:59 | 6 |
| 5 | NIST | 20030925 | 19:59 | 4 |
| 6 | USC | 200804011325 | 19:41 | 4 |

TABLE II
TESTING DATA SET.

| | Source | Name | Length (min:sec) | No. of Speakers |
|---|---|---|---|---|
| 1 | ICSI | Bdb001 | 19:57 | 5 |
| 2 | ICSI | Bed015 | 20:00 | 6 |
| 3 | ICSI | Bmr013 | 20:01 | 7 |
| 4 | ICSI | Bro028 | 20:00 | 4 |
| 5 | ICSI | Buw001 | 20:02 | 8 |
| 6 | NIST | 20011115 | 17:52 | 4 |
| 7 | NIST | 20030702 | 20:00 | 4 |
| 8 | NIST | 20031215 | 19:57 | 5 |
| 9 | USC | 200804011207 | 17:23 | 5 |
| 10 | USC | 200804011233 | 13:01 | 4 |
| 11 | USC | 200804011259 | 6:28 | 4 |

Verified that a sequential process prior to speaker clustering could result in overall clustering performance improvement. In Section IV, we present an ovel approach to speaker clustering within the frame work of conventional agglomerative hierarchical clustering (AHC) by utilizing our previous work [11]-[13] on more reliable AHC performance. The proposed speaker clustering method is based on *incremental Gaussian mixture cluster model inl*and *stopping point estimation based on information change rate* (ICR). The former is a statistical cluster modelling method using a Gaussian mixture model (GMM) whose mixture components increase in proportion to cluster size, and the latter is a better solution to estimating the optimal stopping point(where the owest diarization error rate (DER)would be achieved during AHC)than the conventional one[14]based on Bayesian information criterion(BIC).Ex- per mental results and discussions are given in Section V, and concluding remarks and future work are provided in Section VI.

## II. DATADESCRIPTIONANDEXPERIMENTALSETUP

Tables I and II present the two data sets (training and testing) used forth experiments reported in this paper. The training data set is used fortuning the whole speaker diarization system, while the testing dataset is used for performance evaluation. All the data sources in the data sets were chosen from ICSI, NIST, and USC

Meeting speech corpora, and are distinct from one another in terms of number of speakers and meeting topics (not given in the tables).

In order to measure DER, we use a scoring tool distributed by NIST, i.e., md-eval-v21.pl[1]. This tool calculates DER as the Sum of false-alarm rate, missed-detection rate, and speaker- error-time rate. Each error rate is defined in the evaluation plans[2] released by the RTE valuation thus far.

Mel-frequency cepstral coefficients (MFCCs)are used as acoustic features in this paper. Through 23mel-scaled filter banks,a12-dimensional MFCC vector is generated for every

20ms-long frame of a given data source. Every frame is shifted with the fixed rate of 10mssothat there can be an overlap between two adjacent frames.

**Algorithm 1** Leader-Follower Clustering (LFC)

**Require:** $\{\mathbf{x}_i\}, i = 1, ..., \hat{n}$: data sequentially incoming
$\theta$: threshold
**Ensure:** $C_i, i = 1, ..., n$: clusters finally remaining
1: $C_1 \leftarrow \{\mathbf{x}_1\}, n \leftarrow 1, m \leftarrow 1$
2: **do** $m \leftarrow m + 1$
3: $\hat{C} \leftarrow \{\mathbf{x}_m\}$
4: $i \leftarrow \arg\min d(C_j, \hat{C}), j = 1, ..., n$
5: **if** $d(C_i, \hat{C}) > \theta$
6: $n \leftarrow n + 1$
7: $C_n \leftarrow \hat{C}$
8: **else**
9: merge $\hat{C}$ into $C_i$
10: **until** $m = \hat{n}$
11: **return** $C_i, i = 1, ..., n$

## III. EGMENTATION

In this section, we introduce an ovel segmentation method based on leader-follower clustering(LFC)[15],which is a well-known sequential clustering strategy. As shown in Al- gorithm1, LFC sequentially classifies in coming data, either by having them merged to existing clusters or by generating new clusters for them. The decision is made by comparing the minimum distance between the in coming data and the existing clusters with a preset threshold, and continues until there are no more data.

***A .Speech Activity Detection***

Our proposed segmentation method utilizes this sequential process of LFC for speech activity detection, as follows:

1. We divide the data source given for speaker diarization into 2s-long frames[3] without overlap, and perform LFC (for speech activity detection)on all the frames.

2. LFC decides which cluster every incoming frame is the closest to, choosing from1) the silence cluster, 2)the universal background cluster, and3)one of the existing speaker clusters.

•If1)is selected, the frame considered is labelled

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED SPEECH ACTIVITY DETECTION PROCESS WITH AND WITHOUT UPDATING THE SILENCE CLUSTER, IN TERMS OF THE TWO DETECTION ERROR RATES FOR THE TRAINING DATA SET.

| | Without Update | With Update |
|---|---|---|
| False-Alarm Rate | 2.90% | 2.70% |
| Missed-Detection Rate | 3.05% | 4.03% |
| Total Detection Error Rate | **5.95%** | 6.73% |

•If2)is chosen, a new speaker cluster for the frame is generated.(The frame is newly labelled as well.)

•Incase3),the frame is merged to the corresponding speaker cluster. (It comes to have the same label as the other frames in the cluster.)

3. The previous step is repeated until there remain no more incoming frames.

Forth is process, the silence and the universal background cluster should be generated prior to LFC.(For reference, there is no speaker cluster initially other than these two clusters. Speaker clusters are generated during LFC.)For the silence cluster, we gather atotalof15s of 25ms-long audio chunks with the lowest energy from the entire data source given for speaker diarization, as summing that silence spreads over the given data source with various lengths at least longer than 25ms, and that the total length of such silence chunks in the data source is at least longer than15soverall. Empirically, 15s is considered as enough amount to fully represent the spectral characteristics of silence. For the universal background cluster, we use the given data source entirely. This huge cluster works as if it is a source-dependent threshold for LFC, and thus we donot need to tune such a certain threshold value prior to the process as shown(as $\theta$)inAlgorithm1in the previous page. Note that the silence cluster is not updated during the proposed sequential process for speech activity detection, while the speaker clusters keep being updated through merging. This is to preserve the initial purity of the silence cluster, which might be damaged by incorrectly merging it with speech frames. Such contamination in the silence cluster could be propagated over the whole process and thus result in a lower rate of speech detection. As shown in Table III, the proposed speech activity Detection process with updating the silence cluster would reduce the false-alarm rate at the relatively high cost of the missed-detection rate. As a result, the sum of the two error rates would increase the overall in this case. In the proposed process, distance between the frame considered and all the clusters is measured by generalized likelihood ratio(GLR)[18].For the frame *F* and one of the clusters *C*, GLR for the two

$$\mathrm{GLR}(F,C) = \frac{p(F|\Theta_F)\cdot p(C|\Theta_C)}{p(F\cup C|\Theta_{F\cup C})}. \qquad (1)$$

Each object and the union of the objects are modelled by single Gaussian distributions with full covariance matrices to compute the likelihoods in the equation above, and Θ is a set of parameters in each normal distribution and is estimated toward maximizing the likelihoods of the data(or acoustic feature vectors) in *F*,*C*, and $F \cup C$ for the respective model distributions. Since single Gaussian models are used for representing the objects, Eq.(1)can be simplified as follows:

$$\mathrm{GLR}(F,C) = \frac{|\Sigma_{F\cup C}|^{\frac{N_F+N_C}{2}}}{|\Sigma_F|^{\frac{N_F}{2}}\cdot|\Sigma_C|^{\frac{N_C}{2}}}, \qquad (2)$$

Where Σ is a covariance matrix for a normal distribution,

$|\cdot|$ is the determinant of a matrix, and*N* is the cardinality of the objects considered.

***B.Speaker Change Detection***

For speaker change detection, we use the result of the previous process for speech activity detection. As shown in the previous sub section, every 2s-long incoming frame to LFC is labelled as silence or one of the speaker tags assigned to the speaker clusters, respectively. In other words, all the frames except silence frames have the respective speaker tags, which means that we already have the boundary information of potential speaker changing points in the given data source. Therefore, using this information, we can further divide the data source into speaker-specific segments, each of

which is surrounded by two consecutive boundaries. Every resultant segment becomes an initial cluster for AHC in then ext step, i.e. ,speaker clustering.

## IV. SPEAKERCLUSTERING

In this section, we apply our recent work [11]-[13]for enhancing the reliability of AHC performance under the framework of speaker diarization. In the previous work, we assumed perfect speech activity and speaker change detection to concentrate on AHC aspects. Although this assumption was reasonable in that errors from the two detection steps are usually not that significant incurrent state-of-the-art speaker diarization systems, it is still obvious that such errors exist any how and could give a negative effect to AHC performance to some degree. It would be interesting to see if our previous research results with such as sum option can also be applied to the end-to-end speaker diarization system. Let us start this section by

Briefly investigating how AHC works in speaker diarization systems. As shown in Algorithm2, AHC considers the speaker-specific segments given from speaker change detection as individual initial clusters, and recursively merges the closest pair of clusters in terms of speaker characteristics. Its recursive process continues until it is decided that extra cluster merging would not improve speaker clustering performance any further, i.e. until DER is estimated to reach its lowest level. All these gments in each of the clusters finally remaining are identically labelled ,and every cluster label is unique.

In order for AHC to achieve reliable performance, two critical questions need to be answered properly: 1) how to select the closest pair of clusters for merging at every recursion step and 2)how to decide the optimal(recursion) stopping point where the lowest DER would be achieved. In this context, our proposed speaker clustering methodutilizes

**Algorithm 2** Agglomerative Hierarchical Clustering (AHC)

**Require:** $\{\mathbf{x}_i\}, i = 1, ..., \hat{n}$: speaker-specific segments
$\hat{C}_i, i = 1, ..., \hat{n}$: initial clusters
**Ensure:** $C_i, i = 1, ..., n$: clusters finally remaining
1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\}, i = 1, ..., \hat{n}$
2: **do**
3: $i, j \leftarrow \arg\min d(\hat{C}_k, \hat{C}_l), k, l = 1, ..., \hat{n}, k \neq l$
4: merge $\hat{C}_i$ and $\hat{C}_j$
5: $\hat{n} \leftarrow \hat{n} - 1$
6: **until** DER is estimated to reach the lowest level
7: **return** $C_i, i = 1, ..., n$

Two novel approaches to address the questions ,respectively: incremental Gaussian mixture cluster modelling and ICR-based stopping point estimation.

***A. Incremental Gaussian Mixture Cluster***

***Modeling***

The inter-cluster distance measurement to select the closest pair of clusters at every recursion step of AHC is typically done by comparing ΔBIC values for all possible cluster pairs [14].(Once such comparison is done, the cluster pair having the smallest ΔBIC value is picked for merging.)For two clusters $C_x$and $C_y$, ΔBIC is presented as follows:

$$\Delta\text{BIC}\,(C_x, C_y) = \ln \text{GLR}\,(C_x, C_y) - \frac{\lambda}{2}(m_1 - m_2)\ln N_{\text{total}}, \quad (3)$$

Where $\lambda$(equalto1,ideally[19])is a tuning parameter ,$m_1$ is the sum of the numbers of parameters in the statistical distributions representing $C_x$ and $C_y$,and $m_2$ is the number of parameters in

The distribution representing the union of the two clusters.(These distributions are the ones used as cluster models for GLR computation.) In addition, $N_{\text{total}}$ is thesum of the cardinalities of all the initial clusters for the given data source. Therefore, Eq.(3)would be written as below with single Gaussian cluster modelling as in Eq.(2),

would not improve speaker clustering performance any further, i.e. until DER is estimated to reach its lowest level. All these gments in each of the clusters finally remaining are identically labelled ,and every cluster label is unique.

In order for AHC to achieve reliable performance, two critical questions need to be answered properly: 1) how to select the closest pair of clusters for merging at every recursion step and 2)how to decide the optimal(recursion) stopping point where the lowest DER would be achieved. In this context, our proposed speaker clustering methodutilizes

**Algorithm 2** Agglomerative Hierarchical Clustering (AHC)

**Require:** $\{\mathbf{x}_i\}, i = 1, ..., \hat{n}$: speaker-specific segments
$\hat{C}_i, i = 1, ..., \hat{n}$: initial clusters
**Ensure:** $C_i, i = 1, ..., n$: clusters finally remaining
1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\}, i = 1, ..., \hat{n}$
2: **do**
3: $i, j \leftarrow \arg\min d(\hat{C}_k, \hat{C}_l), k, l = 1, ..., \hat{n}, k \neq l$
4: merge $\hat{C}_i$ and $\hat{C}_j$
5: $\hat{n} \leftarrow \hat{n} - 1$
6: **until** DER is estimated to reach the lowest level
7: **return** $C_i, i = 1, ..., n$

Two novel approaches to address the questions ,respectively: incremental Gaussian mixture cluster modelling and ICR-based stopping point estimation.

***B. Incremental Gaussian Mixture Cluster Modeling***

The inter-cluster distance measurement to select the closest pair of clusters at every recursion step of AHC is typically done by comparing ΔBIC values for all possible cluster pairs [14].(Once such comparison is done, the cluster pair having the smallest ΔBIC value is picked for merging.)For two clusters $C_x$and $C_y$, ΔBIC is presented as follows:

$$\Delta\mathrm{BIC}\,(C_x, C_y) = \ln \mathrm{GLR}\,(C_x, C_y) - \frac{\lambda}{2}(m_1 - m_2)\ln N_{\mathrm{total}}, \quad (3)$$

Where $\lambda$(equalto1,ideally[19])is a tuning parameter ,$m_1$ is the sum of the numbers of parameters in the statistical distributions representing $C_x$ and $C_y$,and $m_2$ is the number of parameters in

The distribution representing the union of the two clusters.(These distributions are the ones used as cluster models for GLR computation.) In addition, $N_{\mathrm{total}}$ is thesum of the cardinalities of all the initial clusters for the given data source. Therefore, Eq.(3)would be written as below with single Gaussian cluster modelling as in Eq.(2),

$$\Delta\mathrm{BIC}\,(C_x, C_y) = \frac{N_{C_x} + N_{C_y}}{2}\ln\left|\Sigma_{C_x \cup C_y}\right| - \frac{N_{C_x}}{2}\ln\left|\Sigma_{C_x}\right| - \frac{N_{C_y}}{2}\ln\left|\Sigma_{C_y}\right| - \frac{\lambda}{2}\left\{d + \frac{1}{2}d(d+1)\right\}\ln N_{\mathrm{total}}, \quad (4)$$

where *d* is the dimension of the acoustic feature vectors. Unlike speech activity detection in SectionII. A, however, single Gaussian cluster modelling for inter-cluster distance measurement has a critical issue in AHC, i.e., the average size of the clusters handled increases as merging recursions in AHC continue, where as a single Gaussian distribution has a limited capability in representing clusters of large data size, especially in terms of speaker characteristics [20]. In general, speaker characteristics are known to be better represented by a complex distribution with multiple modes, e.g., a GMM, than by as imple distribution with only one mode. In this sense, single Gaussian cluster modelling could degenerate dis-cernibility between heterogeneous clusters in terms of speaker characteristics at the late recursion steps of AHC, and hence cause speaker clustering performance to degrade severely.

To tackle this issue, we utilize the incremental Gaussian mixture cluster modelling method[11] that we recently pro-posed, which works as follows:

1. Each initial cluster is modelled by a normal distribution.
2. For GLR computation in Eq.(3), the union of the clusters considered is modelled by the distribution[4] whose pdf is the weighted sum of the pdf soft he distributions representing the clusters, respectively, and
3. Any newly merged cluster is modelled by the distribution whose pdf is the weighted sum of the pdf soft he respective distributions representing merging-involved clusters, for GLR computation with other clusters at

the sub sequent recursion steps of AHC.

This approach during AHC enables not only the smooth transition of cluster models from single Gaussian distributions to GMMs, but also the gradual increase in the complexity of GMMs(or the number of mixture components in GMMs).In this cluster modelling method, Eq.(3)is thus written as below:

$$\Delta\mathrm{BIC}\left(C_x, C_y\right) = \ln \frac{p\left(C_x|\Lambda_{C_x}\right) p\left(C_y|\Lambda_{C_y}\right)}{p\left(C_x \cup C_y|\Lambda_{C_x \cup C_y}\right)}, \quad (5)$$

Where $\Lambda_{C_X}$ ,$\Lambda_{C_Y}$, and $\Lambda_{C_X \cup C_Y}$are sets of parameters in the incremental Gaussian mixture distributions representing $C_X \cup C_Y$ is simply determined as follows:

$$f_{\Lambda_{C_x \cup C_y}} = \frac{N_{C_x}}{N_{C_x} + N_{C_y}} f_{\Lambda_{C_x}} + \frac{N_{C_y}}{N_{C_x} + N_{C_y}} f_{\Lambda_{C_y}}. \quad (6)$$

In the above equation, *N* is the cardinality of the clusters, and $F$ is the pdf of a model distribution with .

It is not able that the expectation-maximization (EM)pro cedure, which is normally applied to GMM training for better representation of given observations, is not applied to any GMM in this cluster modeling method, because it was demonstrated in[11]not to significantly improved is cernibility between clusters. This enables GMMs with a considerable number of mixture components to be used as cluster model distributions during AHC with feasible computational com-plexity.

***B.ICR-based Stopping Point Estimation***

A conventional stopping point estimation method, which is based on BIC, checks if ΔBIC for the closest pair of clusters is greater than using Eq.(4) at every recursion step of AHC [14].However, this method is known to be unreliable (across data sources)interms of estimation accuracy [12]-[13]. In order to over come such unreliability, we previously proposed a new stopping point estimation method[12]-[13], namely ICR- based stopping point estimation. In this subsection, we apply it to our speaker diarization system.

According to [12]-13],ICR for two clusters $C_X$and $C_Y$ is defined as

$$\mathrm{ICR}\left(C_x, C_y\right) \triangleq \frac{1}{N_{C_x} + N_{C_y}} \ln \mathrm{GLR}\left(C_x, C_y\right). \quad (7)$$

From the information theory viewpoint, this statistical measure between clusters represents how much entropy would be increased by merging the clusters considered. Thus, it is natural to expect ICR to be small when the clusters considered are homogeneous in terms of speaker characteristics and each cluster is large enough to fully cover the intra-speaker variance of the corresponding speaker identity. In other words, ICR would be small when the clusters considered have the same speaker identity source and do not need additional information in representing full speaker characteristics. On the contrary, ICR would be relatively large when the clusters considered are heterogeneous, or when they are homogeneous but contains mall size data to cover only a part of the whole speaker characteristics. As a consequence, ICR could properly work as a measure to decide homogeneity for clusters if every cluster considered were large enough to fully represent the characteristics of

the corresponding speaker identity.

Based on this, the ICR-based stopping point estimation Method.

1. Waits until AHC reaches the end of its process,i.e.,

Until all the initial clusters are merged to one big cluster.

2. For the pair of clusters merged at the last recursionstep of AHC,$C_X$ and $C_Y$,computes ICR.

3. Compares ICR with a pre-set threshold $\eta$. If ICR($C_X$,$C_Y$) >$\eta$, decides that $C_X$ and $C_Y$ are heterogeneous and considers the pair of clusters merged at then extlatest recursion step. Otherwise, stops considering the merged clusters and selects the recursion step previously considered as the final stopping point.

Like the conventional BIC-base done, this method depends upon the reasoning that every merging after the optimal stopping point would occur only between heterogeneous clusters. There as on why its consideration of the merged clusters starts from the pair of clusters merged at the last recursion step of AHC(i.e., the opposite direction to the one used in the BIC-based method) is because such a strategy can make ICR properly work as a homogeneity measure by handling large clusters only.

### C. *Comparison*

Table IV shows comparison of our proposed approaches versus the conventional ones to cluster modeling and stopping point estimation for AHC. The proposed techniques resulted in improvement of 10.05% (absolute)in terms of speaker clustering performance in the end-to-end speaker diarization system. This means that our previous work[11]-[13]with the assumption of perfect speech activity and speaker change detection is applicable as well without such assumption.

## V. EXPERIMENTALRESULTS

Figure1presents the overall performance of the proposed speaker diarization system on non-overlapped speech in the

TABLE IV
COMPARISON OF 1) INCREMENTAL GAUSSIAN MIXTURE CLUSTER MODELING + ICR-BASED STOPPING POINT ESTIMATION, AND 2) SINGLE GAUSSIAN CLUSTER MODELING + BIC-BASED STOPPING POINT ESTIMATION, IN TERMS OF SPEAKER-ERROR-TIME RATE FOR THE TESTING DATA SET. $\lambda = 25.0$ AND $\eta = 0.225$, WHICH ARE TUNED BASED ON THE TRAINING DATA SET.

| | 1) | 2) |
|---|---|---|
| Speaker-Error-Time Rate | **14.22%** | 24.27% |

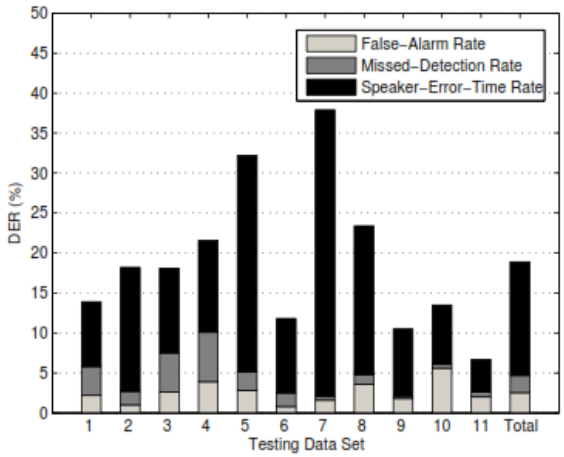



Fig.2. Performance of the proposed speaker diarization system on non- overlapped dspeech in the testing dataset, in terms of DER.

Testing dataset, in terms of DER. The lowest DER(6.77%) was achieved for the test data source 11while the highest one(37.93%) was obtained for the test data source 7. Average DER is 18.90%. These results are quite comparable with those in the recent RT evaluations[21]-[23].(However, fair comparison with other state-of-the-art speaker diarization systems is impossible in this paper because system performance varies across data sources, and the best way for such comparison would be to join in the next RT evaluation possiblyin2009 and compete with the other systems.)

One interesting observation is that a main reason for such relatively bad results at the test data sources 5 and 7 was a lot of wrong merging between heterogeneous clusters during AHC. Furthermore, this also caused mismatch between the optimal and the estimated stopping point, which led to severe DER degradation over all compared to DERs for the other test data sources. (For reference, the ICR-based stopping point estimation method correctly detected the point where DER reaches the lowest level, for the rest o f the test data sources. From these data sources, we were able to find out that there existed relatively small portion of wrong merging during AHC.) This observation support sourer centre search focus on enhancement of inter-cluster distance measurement rather than stopping point estimation[10],[11],[24], with the reasoning that good inter-cluster distance measurement should be a desirable prior condition for reliable estimation of the optimal stopping point.

TABLE V
PERFORMANCE COMPARISON OF THE PROPOSED SPEAKER DIARIZATION SYSTEM ON NON-OVERLAPPED SPEECH ONLY AND SPEECH INCLUDING OVERLAPS IN THE TESTING DATA SET, IN TERMS OF DER AND ITS THREE CONSTITUENT ERROR RATES.

| | No Overlap | Overlap |
|---|---|---|
| False-Alarm Rate | 2.54% | 2.10% |
| Missed-Detection Rate | 2.14% | 13.00% |
| Speaker-Error-Time Rate | 14.22% | 12.77% |
| Total (or DER) | 18.90% | 27.87% |

## VI. CONCLUSION

In this paper, we introduced the SAIL speaker diarization system for analysis of spontaneous meetings, utilizing various novel approaches to segmentation and speaker clustering of a given audio data source. For instance, one of those approaches was to apply LFC to segmentation so that speech activity and speaker change detection can be performed simultaneously. Through the proposed process, we were able to obtain comparable DER of less than 20% on average over11 meeting conversation excerpts of approximately 200 minutes total length. This result was possible due to dynamic cluster representation during AHC by the incremental Gaussian mixture cluster modelling strategy and reliable estimation by the ICR-based stopping point estimation method.

As clearly shown in Table V, overlapped speech detection and classification would be one future research direction toward reliable speaker diarization. In the table, speaker diarization performance gets degraded severely due to the abrupt increase(from2.14%to13.00%)of the missed-detection rate in the case of speech with overlaps. This research field now seems to be in the initial stages with in the community of speaker diarization, and thus there have not been many published works besides[25]. One possible way to overcome this issue would be to utilize diversity using multiple microphonesor even microphone arrays[23],[26].

Another interesting future research direction would be to identify factors to contribute mediocre inter-cluster distance measurement during AHC. The factors can be categorized into two parts, one of which comes from data source characteristics themselves and the other from signal processing/pattern recognition approaches. A more interesting category is the former, because there are many possibilities that should be considered, e.g., inherent discernibility between speakers in a feature space. From this perspective, our previous work has discovered one factor, i.e., the portion of short turns between speakers in a given audio data source[10] However, there is along way to go with this issue.

## REFERENCES

[1] Carlos Busso, SergiHernanz, Chi-Wei Chu,Soon-IlKwon, SungLee, Panayiotis G.Georgiou, Isaac Cohen, and Shrikanth S.Narayanan, "Smartroom: Participantand speaker localization and identification," *Proc.ICASSP2005*,vol.2,pp.1117-1120,Mar.2005.

[2] Carlos Busso, Panayiotis G.Georgiou, and Shrikanth S.Narayanan, "Real-time monitoring of participants interaction in a meeting using audio-visualsensors,"*Proc.ICASSP2007*,vol.2,pp.685-688,Apr.

2007.

[3] Sue E.Tranter and Douglas A.Reynolds,"An overview of automatic speaker diarization systems," *IEEETrans. Audio SpeechLang. Process.*, vol.14(5),pp.1557-1565,Sept.2006.

[4] Douglas A.Reynoldsand Pedro A.Torres-Carrasquillo, "TheMITLin- coln laboratory RT-04 Fdiarizationsystems :Applications to broadcast news and telephone conversations, "*Proc.Fall2004RTWorkshop*,Nov.

2004.

[5] RohitSinha, Sue E.Tranter, Mark J.F.Gales, and PhilC .Woodland, "The Cambridge university March2005 speakerdiarisation system," *Proc.Interspeech2005-Eurospeech*,pp.2437-2440,Mar.2005.

[6] Xavier Anguera, Chuck Wooters, Barbara Peskin, and Mateu Aguilo, "Robust speaker diarization formeetings: ICSIRT06 Smeetings evaluation system, "*Proc.MLMI2006*, pp.346-358,May2006.

[7]David A.vanLeeuwen and Marijn Huijbregts, "The AMI speaker diarizationsystem for NISTRT06 Smeetingdata, "*Proc.MLMI2006*, pp.371-384,May2006.

[8] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-FrancoisBonastre, and LaurentBesacier, "Step-by-step and integrated approaches in broadcast news speaker diarization, "*Computer Speec hLang.*,vol.

20(2-3),pp.303-330,July2006.

[9] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-LucGauvain, "Multistage speaker diarization of broadcast news, "*IEEE Trans. Audio Speech Lang.Process.*, vol.14(5), pp.1505-1512,Sept.2006.

[10] Kyu J.Han,Samuel Kim, and Shrikanth S.Narayanan, "Robust speaker clustering strategies to datasource variation for improved speaker diarization, "*Proc.ASRU2007* ,pp.262-267,Dec.2007.

[11] KyuJ. Hanand Shrikanth S.Narayanan, "Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling,"*Interspeech2008-ICSLP*, underreview.

[12] KyuJ. Hanand Shrikanth S.Narayanan, "A robust

Stopping criterion for agglomerative hierarchical custering in a speaker diarization system,"

*Proc.Inter speech 2007- Euro speech*, pp.1853-1856, Aug. 2007.

[13] Kyu J.Han, Samuel Kim,and Shrikanth S.Narayanan," Robust agglom-erative hierarchical clustering for reliable speaker diarization unde rdata source variation ,"*IEEETrans. AudioSpeechLang.Process.*,inpress.

[14] Scott S.Chenand Panani S.Gopalakrishnan, "Speaker, environment and Channel change detection and clustering via the Bayesian information criterion,"*Proc.1998 DARPABNTU Workshop*, pp.127-132,Feb. 1998.

[15] Richard O.Duda, Peter E.Hart, and David G.Stork ,*Patternclassifica-tion*. 2ndedition,JohnWiley&Sons,2001.

[16] AlainTritschlerand Ramesh Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion, "*Proc. Interspeech1999-Eurospeech*,vol.2,pp.679-682,Sept.1999.

[17] Perrine Dela court and Christian J.Wellekens, "DISTBIC:Aspeaker-Based segmentation for audio data indexing, "*Speech Communication*, vol.32(1-2),pp.111-126,Sept.2000.

[18] Herbert Gish, Man-HungSiu, and Robin Rohlicek, "Segregation of speakers for speech recognition and Speaker identification," *Proc. ICASSP1991*,vol.2,pp.873-876,May1991.

[19] Gideon Schwarz, "Estimating the dimension of a model, "*The Annals ofStatistics*,vol.6(2),pp.461-464,Mar.1978.

[20] Douglas A.Reynolds and Richard C.Rose, "Robust text-independent speaker identification using Gaussian mixture models,"*IEEE Trans. SpeechA udio Process.*, vol.3(1),pp.72-83,Jan.1995.

[21] Jonathan G. Fiscus, Nicolas Radde, John S. Garofolo, Audrey Le, Jerome Ajot ,and Christophe Laprun, "The Rich Transcription 2005 spring meeting recognition evaluation,"*MLMI2005*, pp.369-389,July2005.

[22] RT-06s speake rdiarization results and speech activity detection results.NIST. [Online] .http://www.nist.gov/speech/tests/rt/2006-spring/pdfs/rt 06s-SPKR-SAD-results-v5.pdf

[23] RT-07 speaker diarization results. NIST. [Online]. http://www.nist.gov/speech/tests/rt/2007/workshop/RT07-SPKR-v7.pdf

[24] K y uJ .Hanand Shrikanth S.Narayanan, "Anovelinter-clusterdistanceMeasure combining GLR and ICR for improved agglomerative hierarchical speaker clustering,"*Proc.ICASSP2008*,pp.4373-4376,Mar.2008.

[25] Kofi Boakye, Beatriz Trueba - Hornero, Oriol Vinyals, and Gerald Fried- land, "Overlapped speech detection for improved speaker diarization in multiparty meetings,"*Proc.ICASSP2008*,pp.4353-4356,Mar.2008.