# ROBUST SPEAKER CLUSTERING STRATEGIES TO DATA SOURCE VARIATION FOR IMPROVED SPEAKER DIARIZATION

*Kyu J. Han, Samuel Kim, and Shrikanth S. Narayanan*

Speech Analysis and Interpretation Laboratory (SAIL)
Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA, USA
Emails: {kyuhan, kimsamue}@usc.edu and shri@sipi.usc.edu

## ABSTRACT

Agglomerative hierarchical clustering (AHC) has been widely used in speaker diarization systems to classify speech segments in a given data source by speaker identity, but is known to be not robust to data source variation. In this paper, we identify one of the key potential sources of this variability that negatively affects clustering error rate (CER), namely short speech segments, and propose three solutions to tackle this issue. Through experiments on various meeting conversation excerpts, the proposed methods are shown to outperform simple AHC in terms of relative CER improvements in the range of 17-32%.

***Index Terms***— Speaker diarization, agglomerative hierarchical clustering (AHC), data source variation, clustering error rate (CER)

## 1. INTRODUCTION

*Speaker diarization* refers to the process that automatically transcribes a given audio data in terms of "who spoke when" [1]. This process can help provide speaker-perspective statistics for the data, such as frequency of speaking turn change, average speaking time per turn, number of speakers, speaking time distribution over speakers, and so on. It also enables selecting the speaker-specific data that can be utilized for unsupervised speaker adaptation. Because of its broad significance, speaker diarization is one of the main categories evaluated in the Rich Transcription Evaluation led by the National Institute of Standards and Technology (NIST).

A speaker diarization system basically consists of three main steps following audio feature extraction. The first step is *speech/non-speech detection*, which separates target speech regions from a given audio data. Next, *speaker change detection* identifies potential speaker changing points in each speech region, and further divides the separated target speech regions into speaker-specific segments. Lastly, *speaker clustering* classifies the resultant segments by speaker identity to append a unique label to the segments belonging to the same speaker. The present paper focuses on aspects of speaker clustering, specifically, in addressing robustness issues due to data source variation. It has been shown that data source variation causes significant performance problems in current speaker diarization systems [1][2].

Agglomerative hierarchical clustering (AHC) [3] has been widely used as a speaker clustering strategy in many of the speaker diarization systems that have been developed by a variety of research institutes [4]-[8], due to its simple structure and acceptable level of performance. Algorithm 1 (inset, next page) shows how AHC works within the framework of speaker diarization. Regarding the speech segments given by the speaker change detection step as initial clusters, AHC recursively merges the closest pair of clusters until clustering error rate (CER) reaches the lowest level. For AHC to work properly, two critical questions need to be answered:

- How to estimate when CER reaches the lowest level?
- How to achieve the minimum possible level of CER?

To address these questions in the state of the art, a stopping method based on Bayesian information criterion (BIC) [9] and a merging-cluster selection scheme based on generalized likelihood ratio (GLR) have been widely used [10][11].

Robustness problems in AHC are faced by both the BIC-based stopping method and the GLR-based merging-cluster selection scheme in the presence of data source variation. The BIC-based stopping method leads to unreliable estimation of determining when CER reaches the lowest level, while the GLR-based merging-cluster selection scheme results in severe variability in the minimum achievable CER. In order to tackle the robustness problem in the BIC-based stopping method, we previously proposed a novel stopping method using information change rate (ICR) in [12], and showed experimental results of improved CER across data sources. In this paper, we tackle the robustness problem in the GLR-based merging-cluster selection scheme.

This paper is organized as follows. In Section 2, the data

THIS WORK WAS SUPPORTED BY *NSF* AND *U.S. ARMY*.

**Algorithm 1** Agglomerative Hierarchical Clustering (AHC)

**Require:** $\{\mathbf{x}_i\}, i = 1, ..., \hat{n}$: speech segments
$\hat{C}_i, i = 1, ..., \hat{n}$: initial clusters
**Ensure:** $C_i, i = 1, ..., n$: finally remaining clusters
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** CER reaches the lowest level
7: **return** $C_i, i = 1, ..., n$

sources and the setup used for experiments in the paper are described. The relationship between data source characteristics and clustering error is investigated in Section 3. Based on this analysis, we note that one of the major factors contributing to the high levels of clustering error is the presence of a large number of short speech segments in a data source. In Section 4, we propose three modified versions of AHC to minimize the effect of such short speech segments on the GLR-based merging-cluster selection scheme. The experimental results comparing the proposed methods on a variety of meeting corpus excerpts are also presented. In Section 5, we conclude the paper with comments on future work.

## 2. DATA SOURCES AND EXPERIMENTAL SETUP

Table I presents the data sources used for the experiments reported in this paper, which include 5 different meeting conversation excerpts (of total length approximately 1 hour). The data sources are chosen from ICSI, NIST, and ISL meeting speech corpora[1], and are distinct from one another in terms of number of speakers, gender distribution over speakers, total speaking time, number of speaking turn changes, and average speaking time per turn.

For the experiments in this paper, we assume that both speech/non-speech detection and speaker change detection are perfectly done so that we can concentrate on AHC issues. To enable this, we manually segmented each data source according to a reference transcription prior to the experiments. In order to avoid the potential confusion (in performance analysis) that might result from overlaps between segments, we excluded all the segments involved in any overlap during data preparation.

Mel-frequency cepstral coefficients (MFCCs) are used as general acoustic features in this paper. Through 23 mel-scaled filter banks, a 12-dimensional MFCC vector is generated for every 20ms-long frame of speech regions. Every frame is shifted with the fixed rate of 10ms so that there can be an overlap between two adjacent frames. In order to measure CER, we used a scoring tool, i.e., md-eval-v21.pl, distributed by NIST [http://www.nist.gov/speech/tests/rt/rt2007].

[1]LDC2004S02, LDC2004S09, and LDC2004S05, respectively.

**Table 1**. Data sources. $N_s$: # of speakers (male:female), $T_s$: total speaking time (sec.), $N_t$: # of speaking turn changes, and $T_a$: average speaking time per turn (sec.).

| | Data Sources | | | | |
|---|---|---|---|---|---|
| | ICSI-I | ICSI-II | NIST-I | NIST-II | ISL-I |
| $N_s$ | 7 (5:2) | 6 (4:2) | 4 (3:1) | 6 (4:2) | 4 (2:2) |
| $T_s$ | 931.3 | 1148.5 | 443.4 | 624.1 | 477.7 |
| $N_t$ | 278 | 243 | 74 | 143 | 118 |
| $T_a$ | 3.3 | 4.7 | 5.9 | 4.0 | 4.0 |

**Table 2**. Lowest levels of CER for data sources.

| | ICSI-I | ICSI-II | NIST-I | NIST-II | ISL-I |
|---|---|---|---|---|---|
| CER | 19.29% | 2.65% | 7.63% | 9.72% | 27.00% |

## 3. ROBUSTNESS PROBLEM IN AHC CAUSED BY GLR-BASED MERGING-CLUSTER SELECTION

The GLR-based merging-cluster selection scheme chooses the pair having the smallest GLR value among all pairs of (remaining) clusters as the closest pair for merging. For a certain pair of clusters $C_X$ and $C_Y$ consisting of feature samples $X = \{x_1, x_2, \cdots, x_M\}$ and $Y = \{y_1, y_2, \cdots, y_N\}$ respectively, GLR is computed as follows:

$$\begin{aligned} \text{GLR}(C_X, C_Y) &= \frac{P(X \cup Y|H_0)}{P(X \cup Y|H_A)} \\ &= \frac{P(X|\theta_X)}{P(X|\theta_{X\cup Y})} \times \frac{P(Y|\theta_Y)}{P(Y|\theta_{X\cup Y})}, \end{aligned} \quad (1)$$

where

- $H_0$: $C_X$ and $C_Y$ are left unmerged. The clusters are modeled by two normal distributions $\theta_X$ and $\theta_Y$, whose model parameters are estimated by way of maximizing the likelihoods of $X$ and $Y$ respectively.
- $H_A$: $C_X$ and $C_Y$ are merged. A newly merged cluster is modeled by one normal distribution $\theta_{X\cup Y}$, whose model parameters are estimated by way of maximizing the likelihood of $X \cup Y$.

Since $\theta_X$, $\theta_Y$, and $\theta_{X\cup Y}$ are all normal distributions, the above equation can be simplified [10] as

$$\text{GLR}(C_X, C_Y) = \frac{|\Sigma_{\theta_{X\cup Y}}|^{\frac{M+N}{2}}}{|\Sigma_{\theta_X}|^{\frac{M}{2}} |\Sigma_{\theta_Y}|^{\frac{N}{2}}}, \quad (2)$$

where $\Sigma_{\theta_X}$, $\Sigma_{\theta_Y}$, and $\Sigma_{\theta_{X\cup Y}}$ are sample covariance matrices for $\theta_X$, $\theta_Y$, and $\theta_{X\cup Y}$ respectively, and $|\cdot|$ is determinant. For reference, $\Sigma_{\theta_{X\cup Y}}$ has the following relation with $\Sigma_{\theta_X}$ and

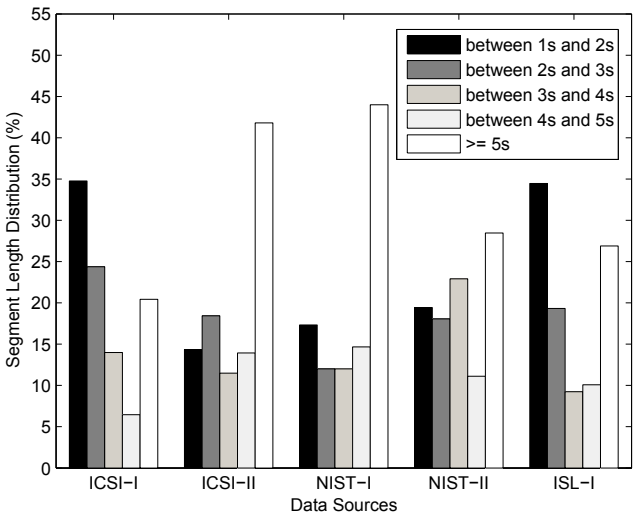


**Fig. 1**. Segment length distributions for data sources.

**Table 3**. Lowest levels of CER when short speech segments are excluded from each data source.

| | ICSI-I | ICSI-II | NIST-I | NIST-II | ISL-I |
|---|---|---|---|---|---|
| CER | 5.36% | 0.47% | 0.99% | 8.94% | 16.22% |

$\Sigma_{\theta_Y}$:

$$\Sigma_{\theta_{X\cup Y}} = \frac{M\Sigma_{\theta_X}+N\Sigma_{\theta_Y}}{M+N}+\frac{M\mu_{\theta_X}\mu_{\theta_X}^T+N\mu_{\theta_Y}\mu_{\theta_Y}^T}{M+N} - \frac{M\mu_{\theta_X}+N\mu_{\theta_Y}}{M+N}\cdot\left(\frac{M\mu_{\theta_X}+N\mu_{\theta_Y}}{M+N}\right)^T, \quad (3)$$

where $\mu_{\theta_X}$ and $\mu_{\theta_Y}$ are sample means for $\theta_X$ and $\theta_Y$ respectively.

Table 2 shows the minimum achievable CER for each of the data sources described in Section 2. The large variability in the results of the table demonstrate the robustness problem in AHC due to the GLR-based merging-cluster selection scheme. Specifically, the levels of CER for ICSI-I and ISL-I are distinctly high compared to those for the other data sources considered. In order to investigate the relationship between the lowest possible level of CER and a data source, we analyzed the properties of the data sources and found significant differences in constituent speech segment length distributions. Fig. 1 shows the distributions of segment lengths for each of the data sources considered. The interesting observation found in this figure is that ICSI-I and ISL-I consist of a large number of speech segments that are shorter than 3 seconds[2]. The proportions of such segments in these data sources exceed 50%. This led us to hypothesize a negative

[2]Let us call these segments *short speech segments* in the rest of this paper. In contrast, we call the speech segments longer than or equal to 3 seconds *long speech segments*.

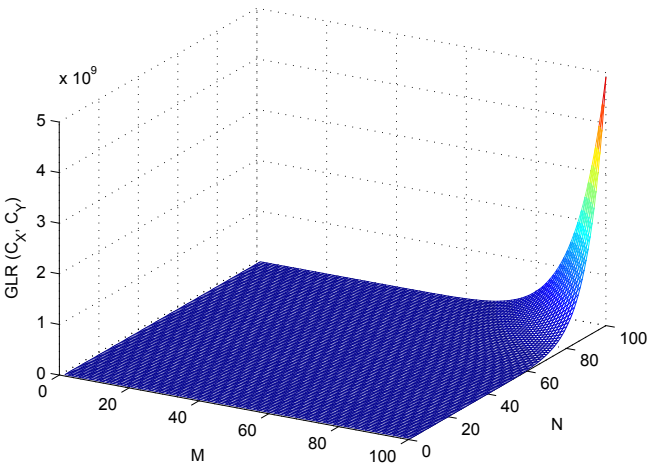


**Fig. 2**. GLR versus the number of feature samples in each cluster with the fixed second order statistics: $\mu_{\theta_X}=0$, $\mu_{\theta_Y}=1$, and $\Sigma_{\theta_X}=\Sigma_{\theta_Y}=1$.

relation between the portion of short speech segment in a data source and achievable CER.

To further confirm the effect of short speech segments on CER, we re-calculated CERs for the experiments presented in Table 2 by excluding short speech segments and report them in Table 3. Note that the lowest levels of CER for all the data sources in this table are noticeably improved compared to those in Table 2. Specifically, the improvements for ICSI-I and ISL-I are considerable (19.29%→5.36% and 27.00%→16.22%, respectively).

We hence claim that the *large portion of short speech segments* in a given data source is a significant factor to negatively affect CER. Short speech segments can arise out of two causes, one of which is due to the inherent nature of interactions as to how many short speaking turns are contained in a data source and the other is technological, depending on how speaker change detection is tuned (Note that in speaker diarization systems speaker change detection is usually tuned not to miss any speaker changing points at the cost of false alarms, which could generate a large number of short speech segments.) In this work, our focus is on the former since we assume speaker change detection is done perfectly.

In order to mitigate the negative effect of short speech segments on CER, it is necessary to examine the specific way that the GLR-based merging-cluster selection scheme is affected when a given data source for AHC contains a large number of short speech segments. According to [12], GLR gets larger as the total number of feature samples within a pair of clusters under consideration increases. This can be easily confirmed by Fig. 2, which shows GLRs between two clusters $C_X$ and $C_Y$ consisting of feature samples $X=\{x_1,x_2,\cdots,x_M\}$ and $Y=\{y_1,y_2,\cdots,y_N\}$ respectively along with the number of feature samples in each cluster. In order to observe the

**Table 4**. Distribution of types for the first quarter of the whole merging processes during AHC for each data source. $M_{ss}$: merging process between short speech segments, $M_{sl}$: merging process between a short and a long speech segment, and $M_{ll}$: merging process between long speech segments.

| | ICSI-I | ICSI-II | NIST-I | NIST-II | ISL-I |
|---|---|---|---|---|---|
| $M_{ss}$ | 52.86% | 16.39% | 21.05% | 22.22% | 50.00% |
| $M_{sl}$ | 35.71% | 44.26% | 47.37% | 47.22% | 40.00% |
| sub_total | 88.57% | 60.65% | 68.42% | 69.44% | 90.00% |
| $M_{ll}$ | 11.43% | 39.35% | 31.58% | 30.56% | 10.00% |

**Table 5**. Reliability of GLR-based merging-cluster selection scheme. The convention is same as that in Table 4.

| | $M_{ss}$ | $M_{sl}$ | $M_{ll}$ |
|---|---|---|---|
| Reliability | 80.22% | 93.58% | 98.17% |

effect of the number of the feature samples, we fix the second order statistics of $\theta_X$ and $\theta_Y$ arbitrarily. (In this case, $\mu_{\theta_X} = 0$, $\mu_{\theta_Y} = 1$, and $\Sigma_{\theta_X} = \Sigma_{\theta_Y} = 1$.) This figure clearly illustrates the abrupt increase of GLR as the number of the feature samples grows. Consequently, it shows that a pair of homogeneous clusters consisting of a small number of feature samples are likely to have smaller GLR values and will be regarded as closer than those consisting of a large number of feature samples.

This dependency of GLR on the total number of feature samples within a pair of clusters under consideration results in the tendency of the GLR-based merging-cluster selection scheme to preferentially select short speech segments as the closest for merging in the early stages of AHC. The tendency is well noticed in Table 4, where the distribution of the first quarter of the whole merging processes during AHC for each data source is given in terms of the length of the speech segments selected for merging. From the third row of this table, we can observe that short speech segments are involved in at least 60% of the first quarter of the whole merging processes during AHC for all the data sources. This tendency is particularly distinct for ICSI-I and ISL-I, which seems reasonable because these data sources contain a large number of short speech segments. Note that the proportion of merging processes between short speech segments ($M_{ss}$) is about 50% for both ICSI-I and ISL-I.

A problem is that the GLR-based merging-cluster selection scheme is not reliable when two short speech segments are selected to be the closest for merging. Table 5[3] clearly demonstrates this problem, indicating that approximately 20% of merging processes between short speech segments are likely to occur erroneously, i.e., occur between heterogeneous ones. Noting that over 50% of the first quarter of the whole merging processes occur between short speech segments for ICSI-I and ISL-I compared to below 25% for the other data sources, we can conclude that erroneous merging processes between short speech segments occur more frequently for data sources containing a large number of short speech segments. Considering that AHC has a recursive structure and thus any erroneous merging process during AHC becomes a potential seed for other erroneous merging processes in subsequent stages, frequent erroneous merging during AHC due to a large number of short speech segments can be regarded as a direct cause for the high levels of CER.

[3]For computing this reliability, we separated merging between homogeneous speech segments and merging between heterogeneous ones, and classified all of them by the length of the speech segments involved in merging.

**Algorithm 2** Modified Version 1 of AHC

**Require:** $\{\mathbf{x}_i\}, i = 1, ..., \hat{n}$: speech segments
$\hat{C}_i, i = 1, ..., \hat{n}$: initial clusters
**Ensure:** $C_i, i = 1, ..., n$: finally remaining clusters
1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\}, i = 1, ..., \hat{n}$
2: **do**
3: $i, j \leftarrow \arg\min \text{GLR}(\hat{C}_k, \hat{C}_l)$ such that either $\{\mathbf{x}_k\}$ or $\{\mathbf{x}_l\}$ is a long speech segment $\geq 3$ sec., $k, l = 1, ..., \hat{n}, k \neq l$
4: merge $\hat{C}_i$ and $\hat{C}_j$
5: $\hat{n} \leftarrow \hat{n} - 1$
6: **until** CER reaches the lowest level
7: **return** $C_i, i = 1, ..., n$

## 4. MODIFIED VERSIONS OF AHC

In this section, we propose three modified versions of AHC to constrain short speech segments so as to minimize their effect on the GLR-based merging-cluster selection scheme. For this, three different methods to prevent erroneous merging processes between short speech segments are introduced in the following sections (4.1–4.3). Experimental results are given in Section 4.4.

### 4.1. Modification of GLR-based scheme

The first method to avoid erroneous merging tries to prevent merging processes between short speech segments from the very beginning. By doing this, merging processes are made to occur only between a short and a long speech segment or between long speech segments. This idea is based on the results in Table 5, showing that the reliability of the GLR-based merging-cluster selection scheme is quite acceptable for both $M_{sl}$ and $M_{ll}$ while relatively poor for $M_{ss}$.

This method, as shown in Algorithm 2, can be implemented by modifying the GLR-based merging-cluster selection scheme so that it can select a pair of clusters (or two speech segments) having the smallest GLR among all the pairs

**Algorithm 3** Modified Version 2 of AHC

**Require:** $\{\mathbf{x}_i\}, i = 1, ..., \hat{n}$: speech segments
$\hat{C}_i, i = 1, ..., \hat{n}', \hat{n}' \leq \hat{n}$: initial clusters
**Ensure:** $C_i, i = 1, ..., n$: finally remaining clusters

sort $\{\mathbf{x}_i\}$ in the descending order of length
$\hat{C}_j \leftarrow \{\mathbf{x}_i\}$ such that $\{\mathbf{x}_i\}$ is a long speech segment $\geq 3$ sec., $i = 1, ..., \hat{n}$ and $j = 1, ..., \hat{n}'$
$m = \hat{n}' + 1$
**do**
  $\hat{C} \leftarrow \{\mathbf{x}_m\}$
  $i \leftarrow \arg\min \text{GLR}(\hat{C}, \hat{C}_k), k = 1, ..., \hat{n}'$
  merge $\hat{C}$ to $\hat{C}_i$
  $m \leftarrow m + 1$
**until** $m > \hat{n}$
**do**
  $i, j \leftarrow \arg\min \text{GLR}(\hat{C}_k, \hat{C}_l), k, l = 1, ..., \hat{n}', k \neq l$
  merge $\hat{C}_i$ and $\hat{C}_j$
  $\hat{n}' \leftarrow \hat{n}' - 1$
**until** CER reaches the lowest level
**return** $C_i, i = 1, ..., n$

either of which is a large size cluster (or a long speech segment), and not among all pairs of remaining clusters.

### 4.2. Pre-classification of short speech segments

The second method is to merge every short speech segment with a long speech segment prior to AHC. It has the same basic idea as the previous method in the sense of preventing merging processes between short speech segments from occurring in AHC, but is a different approach to implementing the idea.

Algorithm 3 shows how this method can be implemented. The method first finds the closest long speech segment for every short speech segment in terms of GLR, and then merges them prior to AHC. After this pre-classification step for short speech segments is done, AHC is performed for the remaining long speech segments.

### 4.3. Sequential classification prior to AHC

The last method is to run leader-follower clustering[4] (LFC) [3] prior to AHC. Instead of pre-screening merging processes between short speech segments like the two methods previously proposed, this method just reduces the proportion of merging processes between short speech segments by letting long speech segments be preferentially considered for merging through LFC.

For this, as shown in Algorithm 4, the first step in the method is to sort speech segments in the descending order of length before running LFC. Then, LFC and AHC are run in a serial manner for the sorted speech segments. The threshold $\eta$ used in LFC is empirically set to be 250.0 in this paper through preliminary experiments for minimizing the average of the lowest levels of CER.

[4] In this sequential clustering strategy, input data are classified in the order of incoming without any pre-trained class model. Thus, the first incoming data automatically becomes the first class and every data thereafter either is merged to one of existing class(es) or becomes another new class.

**Algorithm 4** Modified Version 3 of AHC

**Require:** $\{\mathbf{x}_i\}, i = 1, ..., \hat{n}$: speech segments, $\eta$: threshold
$\hat{C}_i, i = 0, ..., \hat{n}', \hat{n}' \leq \hat{n}$: intermediate clusters
**Ensure:** $C_i, i = 1, ..., n$: finally remaining clusters

sort $\{\mathbf{x}_i\}$ in the descending order of length
$\hat{C}_1 \leftarrow \{\mathbf{x}_1\}, \hat{n}' = 1, m = 2$
**do**
  $\hat{C} \leftarrow \{\mathbf{x}_m\}$
  $i \leftarrow \arg\min \text{GLR}(\hat{C}, \hat{C}_k), k = 1, ..., \hat{n}'$
  **if** $\min \text{GLR}(\hat{C}, \hat{C}_i) > \eta$
    $\hat{n}' = \hat{n}' + 1$
    $\hat{C}_{\hat{n}'} = \hat{C}$
  **else**
    merge $\hat{C}$ to $\hat{C}_i$
  $m \leftarrow m + 1$
**until** $m > \hat{n}$
**do**
  $i, j \leftarrow \arg\min \text{GLR}(\hat{C}_k, \hat{C}_l), k, l = 1, ..., \hat{n}', k \neq l$
  merge $\hat{C}_i$ and $\hat{C}_j$
  $\hat{n}' \leftarrow \hat{n}' - 1$
**until** CER reaches the lowest level
**return** $C_i, i = 1, ..., n$

### 4.4. Experimental Results and Discussion

Table 6 shows the minimum achievable CERs for the three modified versions of AHC with the same data sources used in Section 3. The most noticeable observation found from this table is the huge drop in a CER level for ICSI-I by the third method ($19.29\%$ in Table 2$\rightarrow$ $4.85\%$ in Table 6). This can be explained in terms of the types of merging processes that occur in the earlier stages of AHC. $M_{ll}$ are likely to occur ahead of $M_{ss}$ or $M_{sl}$ in the third method while $M_{ss}$ typically occurs before $M_{sl}$ or $M_{ll}$ in simple AHC. Considering that in Table 5 the reliability of the GLR-based merging-cluster selection scheme for $M_{ll}$ is much higher than for $M_{ss}$, this significant performance improvement by the third method becomes obvious. Based on the observation that most of the results in Table 6 are improved compared to their counterparts in Table 2, we can conclude that our proposed methods achieve their purpose of tackling the negative effect of short speech segments in a data source on CER. The overall performance improvements brought about by these three methods are 17.97%, 20.12%, and 32.49% (relative), respectively.

Comparisons between the proposed methods or with basic AHC would be easier with Fig. 3. One interesting observation is that the performance improvement for ISL-I is not as

**Table 6**. Lowest levels of CER. M1: modified version 1 of AHC, M2: modified version 2 of AHC, and M3: modified version 3 of AHC.

| | ICSI-I | ICSI-II | NIST-I | NIST-II | ISL-I |
|---|---|---|---|---|---|
| M1 | 11.87% | 3.79% | 7.63% | 9.35% | 21.74% |
| M2 | 11.24% | 1.98% | 3.81% | 8.92% | 27.92% |
| M3 | 4.85% | 2.56% | 3.81% | 9.72% | 23.81% |

high as that for ICSI-I. This could mean that the lowest level of CER for ISL-I is not as much affected by short speech segments as that for ICSI-I. Not surprising are the performance improvements for ICSI-II, NIST-I, and NIST-II which are also not high compared to that for ICSI-I, given that short speech segments are not as widespread in these data sources.

## 5. CONCLUSIONS

In this paper, we analyzed the effect of data source variation on clustering error and focused on one factor, namely short length speech segments. We demonstrated that such segments contribute significantly to robustness issues in AHC caused by the GLR-based merging-cluster selection scheme. Following which, we proposed three simple modifications for AHC and experimentally showed performance improvements using the excerpts drawn from a variety of meeting conversations.

There are several directions for future work including further refinements to the proposed solutions. For instance, in AHC based on pre-sequential classification the parameter $\eta$ determines the number of intermediate clusters, which is directly linked to the lowest level of CER. It was chosen empirically here, while finding ways for optimally setting $\eta$ to minimize CER would be beneficial. Other future directions include determining other data factors beyond segment length that contribute to clustering error.

## 6. REFERENCES


[1] D. A. Reynolds and P. A. Torres-Carrasquillo, "Approaches and applications of audio diarization," *Proc. ICASSP 2005*, vol. 5, pp. 953–956, March 2005.

[2] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14(5), pp. 1557–1565, Sept. 2006.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. 2nd edition, John Wiley & Sons, 2001.

[4] D. Moraru, L. Besacier, S. Meignier, C. Fredouille, and J. Bonastre, "Speaker diarization in the ELISA consortium over the last 4 years," *Proc. Fall 2004 Rich Transcription Workshop*, Nov. 2004.


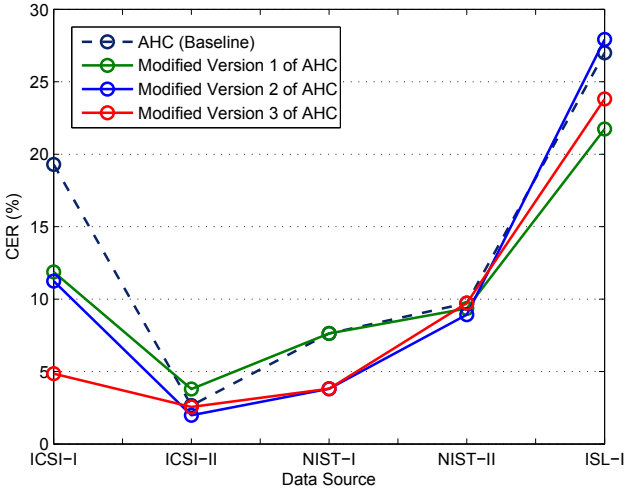


**Fig. 3**. Comparison of all the speaker clustering strategies mentioned in this paper, in terms of the lowest level of CER.


[5] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarisation system," *Proc. INTERSPEECH 2005*, pp. 2437–2440, March 2005.

[6] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," *Proc. MLMI 2005*, pp. 402–414, July 2005.

[7] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Improving speaker diarization," *Proc. Fall 2004 Rich Transcription Workshop*, Nov. 2004.

[8] D. A. Reynolds and P. A. Torres-Carrasquillo, "The MIT Lincoln laboratory RT-04F diarization systems: Applications to broadcast news and telephone conversations," *Proc. Fall 2004 Rich Transcription Workshop*, Nov. 2004.

[9] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6(2), pp. 461–464, March 1978.

[10] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *Proc. DARPA BNTU Workshop*, pp. 127–132, Feb. 1998.

[11] H. Gish, M. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *Proc. ICASSP 1991*, pp. 873–876, May 1991.

[12] K. J. Han and S. S. Narayanan, "A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," *Proc. INTERSPEECH 2007*, pp. 1853–1856, Aug. 2007.