Strategies to Improve the Robustness of Agglomerative Hierarchical Clustering Under Data Source Variation for Speaker Diarization

Kyu J. Han, Student Member, IEEE, Samuel Kim, and Shrikanth S. Narayanan, Senior Member, IEEE

Abstract—Many current state-of-the-art speaker diarization systems exploit agglomerative hierarchical clustering (AHC) as their speaker clustering strategy, due to its simple processing structure and acceptable level of performance. However, AHC is known to suffer from performance robustness under data source variation. In this paper, we address this problem. We specifically focus on the issues associated with the widely used clustering stopping method based on Bayesian information criterion (BIC) and the merging-cluster selection scheme based on generalized likelihood ratio (GLR). First, we propose a novel alternative stopping method for AHC based on information change rate (ICR). Through experiments on several meeting corpora, the proposed method is demonstrated to be more robust to data source variation than the BIC-based one. The average improvement obtained in diarization error rate (DER) by this method is 8.76% (absolute) or 35.77% (relative). We also introduce a selective AHC (SAHC) in the paper, which first runs AHC with the ICR-based stopping method only on speech segments longer than 3 s and then classifies shorter speech segments into one of the clusters given by the initial AHC. This modified version of AHC is motivated by our previous analysis that the proportion of short speech turns (or segments) in a data source is a significant factor contributing to the robustness problem arising in the GLR-based merging-cluster selection scheme. The additional performance improvement obtained by SAHC is 3.45% (absolute) or 14.08% (relative) in terms of averaged DER.

Index Terms—Agglomerative hierarchical clustering (AHC), Bayesian information criterion (BIC), generalized likelihood ratio (GLR), information change rate (ICR), selective agglomerative hierarchical clustering (SAHC), speaker diarization.

I. INTRODUCTION

Speaker diarization refers to the automatic process of dividing a given audio source, predominantly using speech, into speaker-specific segments by transcribing it in terms of “who spoke when” [1]. Such speaker-specific segmentation done by speaker diarization can be beneficial and have many application areas, such as for automatic speech recognition. For instance, speaker diarization enables selecting speaker-specific data that can be utilized for unsupervised speaker adaptation. It also can help provide statistics that rely on speaker-specific information, such as frequency of speaking turn change, average speaking time per turn, number of speakers, speaking time distribution over speakers, and so on. These statistics are useful for multimedia content analysis. Because of its broad significance, speaker diarization is currently regarded as one of the main categories evaluated in the Rich Transcription Evaluation led by the National Institute of Standards and Technology (NIST) [2].

Many state-of-the-art speaker diarization systems have a basic structure in common as shown in Fig. 1, consisting of three main steps following audio feature extraction. One is speech/nonspeech detection, which separates target speech regions from a given audio source. The others are speaker change detection and speaker clustering. Speaker change detection identifies potential speaker changing points in each speech region, and further divides the speech region into smaller speaker-specific segments. Speaker clustering classifies the resultant segments by speaker identity to append a unique label to the segments belonging to the same speaker class. These two steps are in general performed in the order mentioned, i.e., speaker change detection followed by speaker clustering, which the present paper also focuses on. Under this structure for speaker change detection and speaker clustering, we further concentrate on aspects of speaker clustering, specifically, in addressing robustness issues due to data source variation in this paper. It has been shown that data source variation causes significant performance problems in current speaker diarization systems [1], [3].

Agglomerative hierarchical clustering (AHC) [4] has been popularly used as a speaker clustering strategy in many of the speaker diarization systems that have been developed by a number of leading research groups [5]–[10], due to its simple structure and acceptable level of performance. Algorithm 1 shows how it works within the framework of speaker diarization. Using the speech segments given by the speaker change detection step as initial clusters, AHC recursively merges the closest pair of clusters until diarization error rate (DER) reaches the lowest level. In order for AHC to work properly, two critical questions need to be answered.

1) How to estimate when DER reaches the lowest level?
2) How to select homogeneous clusters in terms of speaker identity for merging at every stage of AHC so as to achieve the minimum possible level of DER?

Toward addressing these questions, in the state of the art, a stopping method based on Bayesian information criterion (BIC) [11]...
Despite their popularity, however, both the BIC-based stopping method and the GLR-based intercluster distance measure contribute to the performance degradation of AHC under data source variation. The BIC-based stopping method unreliably estimates the optimal stopping point where DER reaches the lowest level, while the GLR-based intercluster distance measurement unstably selects clusters for merging at every stage of AHC to keep the minimum possible level of DER from being achieved. In this paper, we consider both these issues. In Section II, the data sources and the setup used for experiments in the paper are described. The BIC-based stopping method is investigated in Section III, where we analyze the cause of its sensitivity to data source variation. In Section IV, based on the analysis in Section III, we address the robustness issue in the BIC-based stopping method by proposing a novel alternative based on information change rate (ICR). Through experiments on various meeting conversation excerpts, the ICR-based stopping method is demonstrated to be more robust to data source variation than the BIC-based one. In Section V, we also address the robustness issue in the GLR-based intercluster distance measurement by introducing a simple modified version of AHC, which first runs AHC with the ICR-based stopping method only on the speech segments not shorter than 3 s1 in a data source and then classifies the speech segments shorter than 3 s into one of the clusters given by the initial AHC. This modification that we refer to as selective AHC (SAHC) is motivated by our previous analysis in [14] that the proportion of short speech segments in a data source is one significant source of variability in the minimum achievable DER across data sources. By selective classification of speech segments in terms of length, SAHC mitigates the negative effect of short speech segments on the GLR-based intercluster distance measurement. Finally, we conclude this paper with comments on future work in Section VI.

II. DATA SOURCES AND EXPERIMENTAL SETUP

Tables I and II present the development and evaluation data sets used for the experiments reported in this paper, obtained from 15 different meeting conversation excerpts (of total length approximately 3 h and 45 min). The data sources are chosen from ICSI, NIST, and ISL meeting speech corpora.2 They 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. The development set will be used during parameter tuning for the stopping methods in AHC while the evaluation set will be used for performance calculation.

For the experiments presented in this paper, we assume that both the speech/nonspeech detection step and the speaker change detection step have been perfectly carried out during speaker diarization, allowing us to concentrate on the clustering issues. To enable this, we manually segmented each data source according to the reference transcription provided by the Linguistic Data Consortium (LDC) prior to the experiments.

1Let us call them long speech segments in this paper. Accordingly, we call the speech segments shorter than 3 s short speech segments.

2LDC2004S02, LDC2004S09, and LDC2004S05, respectively.
analysis that might result from overlaps between segments, we excluded all the segments involved in any overlap during data preparation.

In order to measure DER, we used an official scoring tool, i.e., md-eval-v21.pl, distributed by NIST. This tool calculates DER as the sum of missed speaker time rate, false alarm speaker time rate, and speaker error time rate. The first two error rates indicate missed detection and false alarm caused by speech/non-speech detection and speaker change detection, while the last one comes from speaker clustering. Due to the assumption of perfect speech/non-speech detection and speaker change detection, DER in this paper is determined only by speaker error time rate.

Mel-frequency cepstral coefficients (MFCCs) are used as acoustic features in this paper. Through 23 Mel-scaled filter banks, a 12-dimensional MFCC vector is generated for every 20-ms-long frame of speech. Every frame is shifted with the fixed rate of 10 ms so that there can be an overlap between two adjacent frames.

### III. BIC-BASED STOPPING METHOD FOR AHC

We begin this section by providing relevant background details on GLR and BIC. The former is, as mentioned in Section I, a widely used intercluster distance measure for selecting merging clusters at every stage of AHC, and the latter is a well-known model selection criterion and is utilized for the stopping method considered in this section.

#### A. Generalized Likelihood Ratio (GLR)

Suppose that a pair of clusters $C_x$ and $C_y$ are given and they consist of $n_i$-dimensional acoustic feature vectors $x = \{x_1, x_2, \ldots, x_M\}$ and $y = \{y_1, y_2, \ldots, y_N\}$, respectively. Then, GLR for the pair given is computed as follows:

$$\text{GLR}(C_x, C_y) = \frac{P(x \cup y | H_1)}{P(x \cup y | H_2)}$$

(1)

where

- $H_1$ (Unmerging Hypothesis): $C_x$ and $C_y$ are hypothesized to be left unmerged.
- $H_2$ (Merging Hypothesis): $C_x$ and $C_y$ are hypothesized to be merged so as to be a new cluster $C_z$, where $z = x \cup y$.

In order to mathematically calculate the two likelihoods on the right side of (1), the two hypotheses need to be modeled by probability mass or distribution functions (PMFs or PDFs) respectively. For this, single Gaussian modeling for each cluster considered ($C_x$ and $C_y$ for $H_1$, and $C_z$ for $H_2$) has been popularly utilized since [13]. In this paper, we also follow this approach because single Gaussian modeling for the clusters is not only still popular in GLR computation but also much easier to analyze theoretically than other current modeling approaches such as Gaussian mixture modeling (GMM). Based on [13], $C_x$, $C_y$, and $C_z$ are modeled by (multivariate) single Gaussian distributions $f_X$, $f_Y$, and $f_Z$ with full covariance matrices, respectively. Assuming that the PDFs represent random variables $X$, $Y$, and $Z$, respectively, we can regard $x$, $y$, and $z$ (in the modeling framework of [13]) as the sequences of independently and identically distributed (i.i.d.) random variables drawn according to the PDFs $f_X$, $f_Y$, and $f_Z$ of random variables $X$, $Y$, and $Z$, respectively. The mean vectors and the covariance matrices of $f_X$, $f_Y$, and $f_Z$ are determined by way of maximizing the likelihoods of $x$, $y$, and $z$ for $f_X$, $f_Y$, and $f_Z$, respectively. In other words

$$\tilde{x} = (\mu_x, \Sigma_x) = (\mu_{fx}, \Sigma_{fx}) \Rightarrow \theta_x$$

(2)

$$\tilde{y} = (\mu_y, \Sigma_y) = (\mu_{fy}, \Sigma_{fy}) \Rightarrow \theta_y$$

(3)

and

$$\tilde{z} = (\mu_z, \Sigma_z) = (\mu_{fz}, \Sigma_{fz}) \Rightarrow \theta_z$$

(4)

where $\mu_x$, $\mu_y$, and $\mu_z$ are the sample mean vectors, and $\Sigma_x$, $\Sigma_y$, and $\Sigma_z$ are the sample covariance matrices obtained from $x$, $y$, and $z$, respectively. $\mu_{fX}$, $\mu_{fY}$, and $\mu_{fZ}$ are the mean vectors, and $\Sigma_{fX}$, $\Sigma_{fY}$, and $\Sigma_{fZ}$ are the covariance matrices of $f_X$, $f_Y$, and $f_Z$, respectively. Under this framework, (1) can be rewritten as

$$\text{GLR}(C_x, C_y) = \frac{p(x|f_X; \theta_{fx}) \cdot p(y|f_Y; \theta_{fy})}{p(z|f_Z; \theta_{fz})}$$

(5)
can rewrite the equation as below without loss of generality by applying logarithm to both sides

$$\ln \text{GLR}(C_{x}, C_{y}) = \ln \frac{p(x|f_{x}; \theta_1) \cdot p(y|f_{y}; \theta_2)}{p(x|f_{z}; \theta_2) \cdot p(y|f_{z}; \theta_2)},$$

(6)

We can see from (6) that GLR is always greater than or equal to 1 because both of the numerators in the equation are maximal out of the likelihoods of $x$ and $y$, respectively. In other words, $p(x|f_{x}; \theta_1) \geq p(x|f_{z}; \theta_2)$ and $p(y|f_{y}; \theta_2) \geq p(y|f_{z}; \theta_2)$, where the equalities hold only if $C_x = C_y$ or $x = y$. This means that $H_1$ is always more likely than $H_2$, and thus GLR is not adequate to indicate that one hypothesis is more likely than the other. GLR is a measure that provides information on how much more likely $H_1$ is than $H_2$. Therefore, the more likely $H_1$ is for a pair of clusters, the more distant the clusters are regarded in GLR-based distance measurement.

The drawback of GLR as a distance measure is, as mentioned in [14]–[17], that GLR tends to get larger as the total number of feature vectors within a pair of clusters under consideration increases. This can be clearly illustrated in Fig. 2, which shows GLRs between two clusters $C_1$ and $C_2$ along with the numbers of feature vectors $N_1$ and $N_2$, respectively. In order to observe the effect of the number of feature vectors, we fixed the second-order statistics of $\hat{\theta}_1$ and $\hat{\theta}_2$ arbitrarily. (In this case, $\mu_1 = 0, \mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$.) This figure explicitly shows the exponential rising of GLR as the numbers of feature vectors increase. Consequently, in GLR-based intercluster distance measurement, a pair of homogeneous clusters consisting of a small number of feature vectors are likely to have a smaller GLR value and be regarded as mutually closer than those consisting of a large number of feature vectors. Besides, a pair of heterogeneous clusters consisting of a small number of feature vectors might have a smaller GLR value and be regarded as mutually closer than a pair of homogeneous clusters consisting of a large number of feature vectors, which is undesirable.

This undesirable tendency of GLR can be confirmed by analyzing GLR computation with a few basic concepts in the field of information theory. Let us begin this analysis with (5). We

Considering that GLR computation intrinsically assumes the weak law of large numbers$^4$ to be satisfied during its procedure, we can apply the asymptotic equipartition property$^5$ (AEP) widely known as the consequence of the weak law of large numbers in the field of information theory to the right-side term of (7). Then, the equation can be simplified to

$$\ln \text{GLR}(C_{x}, C_{y}) = -M \cdot h(X) - N \cdot h(Y) + (M + N) \cdot h(Z),$$

(9)

where $h$ is entropy. Since entropy for an $n$-dimensional multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ can be obtained (according to [18]) as a closed form of $(1/2) \ln (2\pi e)^n |\Sigma|$ where $| \cdot |$ is determinant, we can further simplify (9) to

$$\ln \text{GLR}(C_{x}, C_{y}) = \frac{-M \cdot 1}{2} \ln (2\pi e)^n |\Sigma_x| - \frac{1}{2} \ln (2\pi e)^n |\Sigma_y| + (M + N) \frac{1}{2} \ln (2\pi e)^n |\Sigma_z|$$

$$= \frac{M + N}{2} \ln |\Sigma_z| - \frac{M}{2} \ln |\Sigma_x| - \frac{N}{2} \ln |\Sigma_y|$$

(10)

where $\Sigma_z$ has the following relation with $\Sigma_x$ and $\Sigma_y$:

$$\Sigma_z = \frac{M \cdot \Sigma_x + N \cdot \Sigma_y}{M + N} + \frac{M \cdot \mu_x \mu_T + N \cdot \mu_y \mu_T}{M + N} - \frac{M \cdot \mu_x + N \cdot \mu_y}{M + N} \cdot \left( \frac{M \cdot \mu_x + N \cdot \mu_y}{M + N} \right)^T$$

(11)

because $z = x \cup y$.

Based on this, suppose that we compute GLR between two clusters $C_{x'}$ and $C_{y'}$, where $x'$ and $y'$ are the sequences of i.i.d. random variables drawn according to the PDFs $f_{x'}$ and $f_{y'}$, and their cardinalities are $2M$ and $2N$, respectively. In other words, $x'$ (or $y'$) has the same second-order statistics with $x$’s (or $y$’s)

$^4$The weak law of large numbers states that a sample mean and a sample variance converge in probability towards the expected value and the second central moment of a corresponding random variable, respectively. In GLR computation, this law is inherent to (2)-(4).

$^5$Let $x_1, x_2, \ldots, x_M$ be the sequence of i.i.d. random variables drawn according to the PDF $f_X$ of a random variable $X$. Then, according to [18], the AEP states that

$$\frac{1}{M} \ln f_X(x_1, x_2, \ldots, x_M) = h(X)$$

(8)

where $h$ is entropy.
but twice the number of feature vectors within $x$ (or $y$). Then, we obtain the same $\Sigma_{z'}$ (where $z' = x' \cup y'$) with $\Sigma_{z}$ using (11), and hence

$$\ln \text{GLR}(C_{x'}, C_{y'}) = (M + N) \ln |\Sigma_{z'}| - M \cdot \ln |\Sigma_{f_x}| - N \cdot \ln |\Sigma_{f_y}| = (M + N) \ln |\Sigma_{z}| - M \cdot \ln |\Sigma_{f_x}| - N \cdot \ln |\Sigma_{y}| = 2 \cdot \ln \text{GLR}(C_{x}, C_{y}).$$

(12)

The above example indicates that $\ln \text{GLR}$ linearly increases (or GLR exponentially increases) with the fixed second-order statistics as the numbers of feature vectors within a pair of clusters under consideration get larger, which is consistent with what is shown in Fig. 2.

B. Bayesian Information Criterion (BIC)

BIC [11] was primarily intended for model (or PDF) selection, specifically for the problem of how to select the best model for given observations from candidate models. A basic model selection strategy based on BIC is as follows.

1) Compute BIC scores for all candidate models

$$\text{BIC}(f) = \ln p(x|f; \theta_f) - P_f = \ln p(x|f; \theta_f) - \frac{1}{2} \#(f) \ln M$$

(13)

where $x = \{x_1, x_2, \ldots, x_M\}$ represents given $M$ observations, $f$ is a model (or PDF), $\theta_f$ is a set of model parameters for $f$, and $\#(f)$ is the total number of model parameters for $f$.

2) Select the model whose BIC score is the highest as the best one to represent the observations.

The core of BIC is that the log-likelihood of given observations for a model is penalized by $P_f$, which is determined by the total number of model parameters and the logarithm of the cardinality of the observations. This prevents the model having the most number of parameters from being chosen all the time as the best one, which is a well-known issue in model selection based on maximum likelihood without penalization.

C. BIC-Based Stopping Method for AHC

Keeping both GLR and BIC in mind, we now investigate the BIC-based stopping method for AHC. This conventional method to search for the optimal stopping point for AHC (when DER reaches the lowest level) was originally introduced in [12] by Chen and Gopalakrishnan. It basically stops AHC at the point when the closest pair among all pairs of remaining clusters are decided to be not homogeneous for the first time, based on the reasoning that if the closest pair of clusters were heterogeneous then so would be any other pair of clusters, and thus there would be no more need for merging in AHC. Decision of homogeneity for the closest pair of clusters at every stage of AHC is done by comparing the BIC scores of the clusters for two hypotheses of “Unmerging” and “Merging.” These two hypotheses are the same as those ($H_1$ and $H_2$) used in GLR computation in Section III-A, and in this case $H_2$ supports homogeneity while $H_1$ supports heterogeneity. As in GLR computation, the two clusters considered are modeled by (multivariate) single Gaussian distributions with maximum-likelihood parameter estimation. The details of how the BIC-based stopping method works for AHC are as follows.\(^6\)

1) For the closest pair of clusters $C_{x}$ and $C_{y}$ consisting of feature vectors $x = \{x_1, x_2, \ldots, x_M\}$ and $y = \{y_1, y_2, \ldots, y_N\}$ respectively, compute the BIC scores of $x \cup y$ for $H_1$ and $H_2$.

$$\text{BIC}(H_1) = \ln P(x \cup y | H_1) - \lambda \cdot P_{H_1} = \ln P(x \cup y | H_1) - \lambda \cdot \frac{1}{2} \#(H_1) \ln N_{\text{total}}$$

$$= \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\} - \lambda \cdot \frac{1}{2} \#(f) \ln N_{\text{total}}$$

$$= \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\} - \lambda \cdot \frac{1}{2} \left\{n + \frac{1}{2}n(n + 1)\right\} \ln N_{\text{total}}.$$  

(14)

$$\text{BIC}(H_2) = \ln P(x \cup y | H_2) - \lambda \cdot P_{H_2} = \ln P(x \cup y | H_2) - \lambda \cdot \frac{1}{2} \#(H_2) \ln N_{\text{total}}$$

$$= \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\} - \lambda \cdot \frac{1}{2} \#(f) \ln N_{\text{total}}$$

$$= \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\} - \lambda \cdot \frac{1}{2} \left\{n + \frac{1}{2}n(n + 1)\right\} \ln N_{\text{total}}.$$  

(15)

In (14) and (15), $\lambda$ is the parameter that should be tuned a priori for minimizing averaged DER with a development set of data sources (which will be explained more in detail later), $N_{\text{total}}$ is the total size (in terms of the number of feature vectors) of the entire clusters given as an input for AHC, and $n$ is the dimension of feature vectors.

2) Compute $\Delta \text{BIC}(C_{x}, C_{y}) = \text{BIC}(H_1) - \text{BIC}(H_2)$.

$$\Delta \text{BIC}(C_{x}, C_{y})$$

$$= \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\} - \lambda \cdot \frac{1}{2} \left\{n + \frac{1}{2}n(n + 1)\right\} \ln N_{\text{total}}$$

$$- \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\} + \lambda \cdot \frac{1}{2} \left\{n + \frac{1}{2}n(n + 1)\right\} \ln N_{\text{total}}$$

$$= \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\}$$

$$\Delta \text{BIC}(C_{x}, C_{y})$$

$$= \ln \{p(x|f; \theta_f) \cdot p(y|f ; \theta_f)\} - \lambda \cdot \frac{1}{2} \left\{n + \frac{1}{2}n(n + 1)\right\} \ln N_{\text{total}}$$

$$- \lambda \cdot \frac{1}{2} \left\{n + \frac{1}{2}n(n + 1)\right\} \ln N_{\text{total}}.$$  

(16)

\(^6\)We used the same notation in Section III-A for single Gaussian modeling for clusters.
3) If $\Delta \text{BIC}(C_x, C_y) < 0$ or $\text{BIC}(H_1) < \text{BIC}(H_2)$, decide that $C_x$ and $C_y$ are homogeneous and merge them. Otherwise, do not merge them and stop AHC.

The stopping criterion mentioned above can be rewritten as

$$\ln \text{GLR}(C_x, C_y) \leq \frac{H_1}{H_2} \lambda \cdot c \cdot \ln N_{\text{total}}$$

(17)

where $c = (1/2)(n + (1/2)n(n + 1))$ is a constant. This criterion could be replaced by

$$\ln \text{GLR}(C_x, C_y) \leq \frac{H_1}{H_2} \lambda \cdot c \cdot \ln (M + N).$$

(18)

This modified criterion was introduced in [10] based on its better performance for estimating the optimal stopping point for AHC than (17). In this paper, we will consider (18) as a baseline stopping criterion for the BIC-based stopping method for this reason. From this point on, the stopping criterion that we mention through the rest of paper thus points to (18), not (17).

D. Tuning Parameter $\lambda$

An important aspect to note for this BIC-based stopping method is the use of the tuning parameter $\lambda$ in (14) and (15). This parameter is not included in the original BIC score computation as shown in (13), which means that the parameter was intentionally introduced when applying BIC to devise a stopping method for AHC. Unfortunately, there is no explicit explanation in [12] of why $\lambda$ is necessary and how it can be optimally chosen. In the field of speaker diarization, however, the parameter is widely considered as a weighting factor to lift up the level of the whole right-side term of (18), and is generally tuned so as for the stopping criterion to provide the minimum averaged DER for a development data set. (In this paper, we set $\lambda$ to be 12.0 because $\lambda = 12.0$ minimized averaged DER for our development data set presented in Section II).

A problem is that $\lambda$ does not work universally. In other words, tuning $\lambda$ in this stopping method cannot guarantee the stopping criterion to correctly estimate the optimal stopping points for every data source. This problem is clearly confirmed in Fig. 3, where comparison of the minimum possible levels of DERs for the evaluation data set described in Section II with the respective DERs achieved by AHC with the BIC-based stopping method with $\lambda = 12.0$. We can see from the figure that with $\lambda = 12.0$ the BIC-based stopping method does not reliably estimate when DER reaches the lowest level for the evaluation data set. In our experiments, the impact of incorrect estimation of the optimal stopping point is detrimental specifically for C-5, C-6, N-2, and I-2, while it is not the case for C-4, C-8, and I-3. Average DER degradation due to such incorrect estimation is about 9.65% (absolute) per data source.

In order to handle this problem, one interesting approach was proposed in [19] based on the idea of [20], which is to automatically erase $\lambda$ by equalizing $(H_1)$ to $(H_2)$ in the computation of BIC scores for $H_1$ and $H_2$. For this, a GMM with $m$ model parameters for each cluster considered ($C_x$ and $C_y$) for $H_1$ and another GMM with $2m$ model parameters for a hypothetically merged cluster ($C_z$) for $H_2$ were utilized, respectively. By doing so, this approach can avoid parameter tuning. However, it has some side effects such as increased computing time for training GMMs at every stage of AHC. Moreover, the approach does not directly take care of a fundamental cause for the robustness issue of the BIC-based stopping method, which is the stopping criterion being not robust to data source variation.

E. Sensitivity of the Stopping Criterion To Data Source Variation

The stopping criterion of the BIC-based method (18) has an intrinsic flaw in terms of robustness to data source variation because it utilizes GLR. As aforementioned in Section III-A, GLR is sensitive to the numbers of feature vectors within the clusters considered. As a result, the left-side term of (18), $\ln \text{GLR}$, is affected by several aspects in the entire speech segments given as an input data source for AHC beyond just the statistical difference between the clusters considered. This is because the size of the clusters considered by the BIC-based stopping method at a certain stage of AHC is determined jointly by the total length of the segments given as an input for AHC, the distributions of the segments in length and speaker identity, and merging procedures at the previous stages of AHC. One might claim that the right-side term of (18) is also affected by the numbers of feature vectors within the clusters considered due to $\ln (M + N)$, so the stopping criterion looks robust to data source variation. However, $\ln \text{GLR}$ grows in a linear fashion in proportion to $M$ and $N$, while $\ln (M + N)$ increases in a logarithmic fashion, which is well shown in Fig. 4. $\ln \text{GLR}$ is fast increasing along $M$ and $N$, but $\ln (M + N)$ looks relatively flat in the figure. This indicates that the right-side term of (18) cannot compensate for the data dependency of the left-side term fully enough, and the stopping criterion is thus highly likely to vary across data sources. For this reason, it is too difficult to set a global $\lambda$.

In this experiment, GLR was used as an intercluster distance measure for AHC to select the closest pair of clusters at every stage.

We confirmed in Section III-A that GLR exponentially increased in proportion to the numbers of feature vectors within the clusters considered.
On the other hand, NLLR is very similar (22) Merging.

GLR represents how much amount of information cannot compensate the and GLR.

Considering that entropy can a pair of clusters under consideration was inspired from ana-

This means that ICR is a normalized version of GLR and fixed second order statistics, \( f_i \).

Distance Measures

In fact, there have been several ICR-like intercluster distance measures to normalize GLR in the field of speaker diarization, specifically for speaker change detection. Table III compares two of such measures, i.e., penalized likelihood ratio (PLR) [16] and normalized log-likelihood ratio (NLLR) [17], with ICR. PLR normalizes GLR with the \( f_i \) over feature vectors by merging the clusters considered, could avoid being affected by the size of the clusters. ICR satisfies such an expectation, which is the reason why we named our proposed distance measure information change rate. From (19) and (20), we can obtain a different version of ICR as follows:

\[
\text{ICR}(C_x, C_y) = h(Z) - \frac{M \cdot h(X) + N \cdot h(Y)}{M + N}, \quad (21)
\]

Let us consider how ICR is expressed for two extreme examples.

- Ex 1: \( C_x = C_y \) or \( x = y \)

\[
\text{ICR}(C_x, C_y) = \text{ICR}(C_x, C_x) = h(X) - \frac{M \cdot h(X) + M \cdot h(X)}{M + M} = h(X) - h(X) = 0,
\]

- Ex 2: \( C_x \) and \( C_y \) are mutually independent

\[
\text{ICR}(C_x, C_y) = h(X) + h(Y) - \frac{M \cdot h(X) + N \cdot h(Y)}{M + N} = \frac{(M + N) \cdot h(X) + (M + N) \cdot h(Y)}{M + N} - \frac{M \cdot h(X) + N \cdot h(Y)}{M + N} = \frac{N \cdot h(X) + M \cdot h(Y)}{M + N}.
\]

B. Comparison of ICR With Other ICR-Like Intercluster Distance Measures

In fact, there have been several ICR-like intercluster distance measures to normalize GLR in the field of speaker diarization, specifically for speaker change detection. Table III compares two of such measures, i.e., penalized likelihood ratio (PLR) [16] and normalized log-likelihood ratio (NLLR) [17], with ICR. PLR normalizes GLR with the \( \alpha \)th power of the sum of feature vectors within the clusters considered. However, it does not appear promising in terms of mitigating the effect of cluster size on distance measurement, because

\[
\text{PLR}(C_x, C_y) = \ln \text{GLR}(C_x, C_y) - \alpha \cdot \ln (M + N), \quad (22)
\]

As shown in Section III-E, \( \ln (M + N) \) cannot compensate the dependency of \( \ln \text{GLR} \) on cluster size entirely. Thus, it is difficult to set a global \( \alpha \). On the other hand, NLLR is very similar to ICR and its relation to ICR is shown as follows:

\[
\text{NLLR}(C_x, C_y) = \frac{1}{n} \text{ICR}(C_x, C_y), \quad (23)
\]
However, it has a different physical meaning from that of ICR because it further normalizes $\ln \text{GLR}$ with the dimension of feature vectors.

C. ICR as a Measure To Decide Homogeneity for Clusters

Since ICR represents what amount of information would be changed on average over feature vectors by merging the clusters considered, it is natural to expect ICR to be very small when the clusters considered are homogeneous in terms of speaker identity and each cluster is large enough to fully cover the intraspeaker variance of 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 for representing full speaker characteristics. In contrast, ICR would be relatively large when the clusters considered are heterogeneous, or when they are homogeneous but contain small feature vectors to cover only a part of the speaker characteristics. Thus, 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. In this paper, we assume that a cluster containing feature vectors which correspond to more than 30 s is such a large enough cluster. This assumption is based on the fact that it requires long speech utterances (at least longer than 20 s) to derive reliable speaker characteristics [21]–[23].

Fig. 5 displays distributions for correct and incorrect merging in terms of ICR. The threshold $\eta$ is set so as to minimize classification error between the two distributions. All the merging processes used for obtaining the distributions were picked up from our development data set, and they corresponded to more than 30 s.

1) Wait until AHC reaches the end of its merging processes, i.e., wait until all the clusters given for AHC are merged to one cluster.

2) For the pair of clusters merged at the last stage of AHC, $C_x$ and $C_y$, consisting of feature vectors $x = \{x_1, x_2, \ldots, x_M\}$ and $y = \{y_1, y_2, \ldots, y_N\}$ respectively, compute ICR.

3) Compare ICR with $\eta$

$$\text{ICR}(C_x, C_y) \begin{cases} H_1 & \text{if } H_2 \geq \eta \\ H_2 & \text{if } H_1 < \eta \end{cases}$$

If ICR($C_x, C_y$) > $\eta$, decide that $C_x$ and $C_y$ are heterogeneous in terms of speaker identity and consider the pair of clusters merged at the next latest stage of AHC. Otherwise, stop considering more merging processes and select the stage previously considered as the stopping point.

The ICR-based stopping method depends upon the reasoning that all merging processes during AHC after the optimal stopping point would occur between heterogeneous clusters. The reason why this stopping method starts its consideration from the pair of clusters merged at the last stage of AHC is because such a strategy can make the stopping criterion (24) consider large clusters only. As mentioned in the previous subsection, ICR can properly work as a homogeneity decision measure only for large enough clusters to represent full speaker characteristics, respectively.

Equation (24) can be rewritten as follows:

$$\ln \text{GLR}(C_x, C_y) \begin{cases} H_1 & \text{if } H_2 \geq \eta \cdot (M + N) \\ H_2 & \text{if } H_1 \geq \eta \cdot (M + N) \end{cases}$$

The BIC-based stopping method for AHC also relies on the same reasoning.

<table>
<thead>
<tr>
<th>ICR ($C_x, C_y$)</th>
<th>PLR in [16]</th>
<th>NLLR in [17]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{M+N} \ln \text{GLR} (C_x, C_y)$</td>
<td>$\frac{1}{(M+N)^2} \text{GLR} (C_x, C_y)$</td>
<td>$\frac{1}{(M+N)^2} \ln \text{GLR} (C_x, C_y)$</td>
</tr>
</tbody>
</table>
TABLE IV
ICR-BASED STOPPING METHOD VERSUS BIC-BASED STOPPING METHOD, $c = \frac{1}{2} \{ n + (1/2)n(n + 1) \}$, WHERE $n$ IS THE DIMENSION OF FEATURE VECTORS. $n = 12$, $\eta = 0.18603$, AND $\lambda = 12.0$ IN THIS PAPER

<table>
<thead>
<tr>
<th>Criterion</th>
<th>ICR-based Stopping Method</th>
<th>BIC-based Stopping Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right side term in criterion</td>
<td>ICR $(C_x, C_y) \frac{H_2}{H_1} \geq \eta$</td>
<td>$\ln \text{GLR} (C_x, C_y) \frac{H_2}{H_1} \geq \lambda \cdot c \cdot \ln (M + N)$</td>
</tr>
<tr>
<td>Computational complexity for criterion</td>
<td>Complexity for computing $\ln \text{GLR} (C_x, C_y)$ and $\eta \cdot (M + N)$</td>
<td>Complexity for computing $\ln \text{GLR} (C_x, C_y)$ and $\lambda \cdot c \cdot \ln (M + N)$</td>
</tr>
<tr>
<td>Order of clusters considered</td>
<td>From the pair of clusters merged at the last stage of AHC</td>
<td>From the pair of clusters merged at the first stage of AHC</td>
</tr>
</tbody>
</table>

Comparing this criterion with (18) for the BIC-based stopping method, we can see that the difference of computational complexity between the two stopping methods is thus negligible. For easier understanding of the ICR-based stopping method for AHC, Table IV is presented.

Fig. 6 shows $\ln \text{GLR}, \theta_{\text{bic}} = \lambda \cdot c \cdot \ln (M + N)$, and $\theta_{\text{icr}} = \eta \cdot (M + N)$ for the data source C-6 in our evaluation data set, where $\lambda = 12.0$ and $\eta = 0.18603$. The stopping point estimated by the ICR-based stopping method is identical to the optimal one in this case.

sources except C-4, C-8, and C-9. Even for the three data sources, gaps between DERs at the estimated stopping points and those at the optimal ones are shown to be insignificant. Compared to the results obtained using AHC with the BIC-based stopping method for the same data set (shown in Fig. 3), the results in this figure are much improved overall, and indicate that the ICR-based stopping method is superior to the BIC-based one in terms of robustness to data source variation. Consequently, the ICR-based stopping method for AHC led to average DER improvement by 8.76% (absolute) and 35.77% (relative) per data source, compared to the conventional BIC-based one.

V. SELECTIVE AGGLOMERATIVE HIERARCHICAL CLUSTERING (SAHC)

In this section, we tackle the robustness issue of intercluster distance measurement for AHC. As mentioned in Section I, GLR is widely used as such a measure to select the closest pair of clusters at every stage of AHC, but its sensitivity to data source variation in terms of accuracy results in the severe variability of the minimum possible level of DER across data sources.
sources. This can be confirmed in Figs. 3 and 7, where the minimum possible levels of DERs severely vary across the data sources considered. A possible key factor contributing to this robustness issue was analyzed in [14], where we found out that the large fraction of the segments shorter than 3 s in the input speech segments to AHC affected the minimum possible level of DER. To avoid such data dependency of the accuracy of the GLR-based intercluster distance measurement, we introduce here a simple modified version of AHC, namely selective AHC (SAHC).

SAHC first runs AHC (with the ICR-based stopping method) only on the segments longer than or equal to 3 s among the speech segments given for AHC, and then classifies the rest of the segments (shorter than 3 s) into one of the final clusters provided by the initial AHC, which is described in Algorithm 2.

By doing this, the modified clustering strategy can enhance the accuracy of the GLR-based intercluster distance measurement during the initial AHC. Fig. 8 shows that AHC for a subset (of a given data source) containing only the segments longer than or equal to 3 s can, in general, achieve better performance than AHC for the entire given segments. Considering $T_0$ in Table I in Section II, we can easily identify from the figure that such performance improvement is remarkable specifically for the data sources with many short segments, i.e., C-1, C-2, and I-1.

Fig. 9 shows SAHC performance for the evaluation data set. From the figure, we can see that SAHC is a reasonable strategy to tackle the robustness issue of the GLR-based intercluster distance measurement. The severe variability of the minimum level of DER across data sources is mitigated to some degree by SAHC. This mitigation was obtained significantly for C-8, C-9, and I-2. The overall DER improvement achieved by SAHC is 21.92% (relative) compared to simple AHC with the ICR-based stopping method.

VI. CONCLUSION

In this paper, we addressed the robustness issues of AHC to data source variation within the framework of speaker diarization, which are faced by the BIC-based stopping method and the GLR-based intercluster distance measurement in AHC. To tackle the problem caused by the BIC-based stopping method we proposed a novel ICR-based alternative. Furthermore, we introduced SAHC as a simple solution to tackle the severe variability of the minimum possible level of DER across data sources due to the sensitivity of the accuracy of the GLR-based intercluster distance measurement to data source variation. Through experimental results on excerpts obtained from meeting corpora, AHC with the ICR-based stopping method and SAHC were shown to outperform and be more robust to data source variation than basic AHC with the BIC-based stopping method. Table V presents performance comparison results of AHC with the BIC-based stopping method, AHC with the

Algorithm 2 Selective Agglomerative Hierarchical Clustering (SAHC)

Require: $\{x_i\}, i = 1, \ldots, \hat{n}$: speech segments $C_i, i = 1, \ldots, \hat{n}, \hat{n}' \leq \hat{n}$: initial clusters

Ensure: $C_i, i = 1, \ldots, n$: finally remaining clusters

1: permute $\{x_i\}$ in the descending order of length
2: $\hat{C}_j \leftarrow \{x_i\}$ such that $\{x_i\}$ is a long speech segment $\geq 3$ sec., $i = 1, \ldots, \hat{n}$ and $j = 1, \ldots, \hat{n}'$
3: $m = \hat{n}'$
4: do
5: $i, j \leftarrow \arg\min_{l, k} \text{GLR}(\hat{C}_k, \hat{C}_l), k, l = 1, \ldots, m, k \neq l$
6: merge $\hat{C}_i$ to $\hat{C}_j$
7: $m \leftarrow m - 1$
8: until DER is estimated to have reached the lowest level
9: return $C_i, i = 1, \ldots, n$
10: $m = \hat{n}' + 1$
11: do
12: $\hat{C} \leftarrow \{x_m\}$
13: $i \leftarrow \arg\min_{k} P(\hat{C} | \hat{C}_k), k = 1, \ldots, n$
14: merge $\hat{C}$ to $\hat{C}_i$
15: $m \leftarrow m + 1$
16: until $m > \hat{n}$
17: return $C_i, i = 1, \ldots, n$
ICR-based stopping method, and SAHC for the evaluation data set. A reason for the improvements achieved by our proposed methods in terms of averaged DER across the data sources in the evaluation data set is because of the undesirable tendency of GLR where it tends to get larger as the total number of feature vectors within a pair of clusters under consideration was removed (in the case of AHC with the ICR-based stopping method), and the negative effect of the segments shorter than 3 s in the speech segments given for AHC on the minimum possible level of DER was mitigated (in the case of SAHC).

One potential future direction is to identify the lower bound for cluster size that guarantees ICR to be reliable as a statistical distance measure, more specifically as a homogeneity decision measure, between the clusters considered. In this paper, we avoided the possibility that ICR would not work properly, by checking ICR-based intercluster homogeneity starting from the pair of clusters merged at the last stage of AHC under the assumption that clusters at the later stages of AHC would be large enough for reliable ICR. This assumption worked for the meeting conversation excerpts used for the experiments presented in the paper because most of the speakers involved in the conversations uttered longer than at least 30 s, which is empirically known to be long enough to represent speaker characteristics adequately. The assumption could be however broken for other data sources which have a preponderance of short speech segments that are inadequate to reveal the speaker characteristics completely.

Another future direction would be to search for the factors in a given data source for AHC that affect the reliability of the GLR-based intercluster distance measurement, other than the portion of short speech segments that we previously discovered. These could include the ratio of male and female speakers, the degree of intrinsic discemibility between speakers in terms of MFCC, and so on.

In this paper, we assumed perfect speech/nonspeech detection and speaker change detection. For real applications, we need to extend our research without this assumption. In this context, it would be a good opportunity for us to plug in the basic ideas of this work into practical applications such as the SMARTROOM project [24].

REFERENCES


Kyu J. Han (S’00) received the B.S. degree in radio communications engineering from Yonsei University, Seoul, Korea, in 2000, and the M.S. degree in electrical engineering from Seoul National University, Seoul, in 2002. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Southern California, Los Angeles. His dissertation focus is on robust speaker clustering under data source variation for reliable speaker diarization. His general research interests include speech and speaker recognition, audio signal processing, and multimedia content analysis.

Samuel Kim received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2003 and 2005, respectively. He is currently pursuing the Ph.D. degree at the University of Southern California, Los Angeles. His research interests include audio, speech, and music signal processing.

Shrikanth S. Narayanan (S’88–M’95–SM’02) received the Ph.D. degree from the University of California, Los Angeles, in 1995. He is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member of its Technical Staff from 1995–2000. At USC, he is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center. He has published over 275 papers and has 14 granted/pending U.S. patents.

Dr. Narayanan is a recipient of an NSF CAREER Award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 Best Paper Award from the IEEE Signal Processing Society. Papers by his students have won best student paper awards at ICSLP’02, ICASSP’05, MMSP’06, and MMSP’07. He is an Editor for the Computer Speech and Language Journal (2007–present) and an Associate Editor for the IEEE Signal Processing Magazine and the IEEE TRANSACTIONS ON MULTIMEDIA. He was also an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004). He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication Committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu.