A NOVEL INTER-CLUSTER DISTANCE MEASURE COMBINING GLR AND ICR FOR IMPROVED AGGLOMERATIVE HIERARCHICAL SPEAKER CLUSTERING

Kyu J. Han and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering
University of Southern California, Los Angeles, CA, USA
E-mails: kyuhan@usc.edu and shri@sipi.usc.edu

ABSTRACT

Agglomerative hierarchical clustering (AHC) has been a popular strategy for speaker clustering, due to its simple structure but acceptable level of performance. One of the main challenges in AHC that affects clustering performance is how to select the closest cluster pair for merging at every recursion. For this, generalized likelihood ratio (GLR) has been widely adopted as an inter-cluster distance measure. However, it tends to be affected by the size of the clusters considered, which could result in erroneous selection of the cluster pair to be merged during AHC. To tackle this problem, we propose a novel alternative to GLR in this paper, which is a combination of GLR and information change rate (ICR) that we recently introduced for addressing the aforementioned tendency of GLR. Experiments on various meeting speech data show that this combined measure improves clustering performance on average by around 30% (relative).

Index Terms— Speaker clustering, agglomerative hierarchical clustering (AHC), generalized likelihood ratio (GLR), information change rate (ICR)

1. INTRODUCTION

Speaker clustering is the process of automatically classifying speaker-specific speech segments by speaker identity, especially in an unsupervised manner. This process is essential for speaker diarization [1] or unsupervised speaker adaptation. The most popular strategy for speaker clustering has been agglomerative hierarchical clustering (AHC) [2] because AHC provides an acceptable level of performance despite its simple processing structure. The details of how this strategy works are shown in Algorithm 1 (inset, next page). In other words, using given speech segments as initial clusters, AHC recursively merges the closest pair of clusters. Its recursive process is stopped when it is decided that extra cluster merging does not improve clustering performance any more.

In order for AHC to achieve reliable performance, two critical questions need to be answered: 1) how to select the closest pair of clusters for merging at every recursion of AHC and 2) how to decide the optimal (recursion) stopping point. For a certain set of speech segments, the first question relates to the minimum possible error rate that would be obtained during AHC, while the second one relates to the specific error rate finally obtained by AHC. Of these two questions we focus on the first one in this paper, because getting the minimum possible error rate lowered as much as possible is a desirable prior condition for properly tackling the second one; consequently, AHC can provide low overall clustering performance eventually.

To address this question of selecting the closest pair of clusters for merging at every recursion of AHC, generalized likelihood ratio (GLR) has been widely adopted as a state-of-the-art distance measure between clusters [3]. However, as mentioned in [4]-[5], GLR has the intrinsic drawback that it tends to be affected by the size of clusters under consideration, which is undesirable in that GLR-based inter-cluster distance measurement is controlled by a factor beyond just the statistical difference between the clusters considered. Thus, this tendency of GLR might result in incorrect selection of the closest pair of clusters for merging during AHC. Such erroneous selection of clusters obviously makes the minimum possible error rate higher, and could also cause incorrect estimation of the optimal stopping point; as a result, overall clustering performance degradation.

Recently we introduced information change rate (ICR) as a measure for deciding whether the clusters considered are homogeneous in terms of speaker identity [4]. Through experiments on various meeting speech data, this measure was demonstrated to have less dependency on the size of clusters under consideration than GLR, and was applied as a stopping criterion to AHC. A notable thing about ICR is that the measure works only if it handles large size clusters to represent speaker characteristics completely. This is because ICR measures how much information would be changed (or increased) by merging a certain pair of clusters; if two homogeneous clusters do not have sufficient information for capturing speaker identity, then their ICR score would be quite large and exceed a pre-determined threshold, meaning that they would be regarded as heterogeneous (in terms of ICR) despite their inherent homogeneity. For this reason, ICR was not further considered as an inter-cluster distance measure beyond as a

Table 1. Data sources (from ICSI Meeting Speech). N_s: total number of speakers (male:female), T_s: total speaking time, N_t: total number of speaking turn changes, and T_a: average speaking time per turn.

	Data Source (DS)						
	DS-1	DS-2	DS-3	DS-4	DS-5	DS-6	DS-7
N_s	7 (5:2)	7 (5:2)	5 (3:2)	6 (5:1)	5 (5:0)	4 (4:0)	9 (7:2)
T_s	1064.9 sec	931.3 sec	674.5 sec	1664.9 sec	1609.1 sec	1475.9 sec	659.7 sec
N_t	417	278	175	531	590	477	158
T_a	2.5 sec	3.3 sec	3.8 sec	3.1 sec	2.7 sec	3.1 sec	4.1 sec

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** no more extra cluster merging is needed
7: **return** $C_i, i = 1, ..., n$

Algorithm 2 AHC with a proposed method

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: **if** all $\{\hat{C}_i\}_{i=1}^{\hat{n}}$ contain data of more than 10 sec.
4: $i, j \leftarrow \arg\min[w_{\text{GLR}}^{k,l} \cdot R_{\text{GLR}}(\hat{C}_k, \hat{C}_l) + w_{\text{ICR}}^{k,l} \cdot R_{\text{ICR}}(\hat{C}_k, \hat{C}_l)]$,
R_{GLR}, R_{ICR} : inter-cluster distance rankings (in the ascending order),
$w_{\text{GLR}}^{k,l}, w_{\text{ICR}}^{k,l}$: weighting factors,
$k = 1, ..., \hat{n}$, and $l = k + 1, ..., \hat{n}$
5: **else**
6: $i, j \leftarrow \arg\min \text{GLR}(\hat{C}_k, \hat{C}_l)$,
$k = 1, ..., \hat{n}$, and $l = k + 1, ..., \hat{n}$
7: merge $\hat{C}_i$ and $\hat{C}_j$
8: $\hat{n} \leftarrow \hat{n} - 1$
9: **until** no more extra cluster merging is needed
10: **return** $C_i, i = 1, ..., n$

stopping criterion.

However, ICR could be utilized as a complement inter-cluster distance measure to GLR in the sense that it could possibly compensate for the aforementioned undesirable tendency of GLR if we are able to manipulate it to handle large clusters only. With this motivation, we propose a novel method for inter-cluster distance measurement in this paper, which combines GLR and ICR so as to provide complementary performance improvement. For this purpose, the paper is organized as follows. In Section 2, the data sources and the setup used for experiments in the paper are described. In Section 3, we introduce and explain a new inter-cluster distance measurement method using GLR and ICR. Performance comparisons between AHC with GLR only and AHC with the proposed method are illustrated as well. In Section 4, we conclude the paper with remarks on future work.

2. DATA SOURCES AND EXPERIMENTAL SETUP

Table 1 presents the data sources used for the experiments reported in this paper. These data sources represent 7 different meeting conversation excerpts with a total length of approximately 2 hours and 45 minutes, and chosen from ICSI Meeting Speech (LDC2004S02). They are distinct from one another in terms of total number of speakers (N_s), gender distribution over speakers, total speaking time (T_s), total number of speaking turn changes (N_t), and average speaking time per turn (T_a).

For preparing input speech segments to AHC, we manually segmented the data sources at every point of speaking turn changes according to the respective reference transcriptions beforehand. In order to avoid any potential confusion in performance analysis that might result from overlaps between segments, we excluded all the segments involved in any overlap during data preparation.

AHC performance is evaluated by speaker error time rate in this paper, which has been officially used as a measure for speaker clustering within the framework of speaker diarization in the Rich Transcription Evaluation by the National Institute of Standards and Technology (NIST). For this, we use the scoring tool, i.e., md-eval-v21.pl, distributed by NIST [http://www.nist.gov/speech/tests/rt/2006-spring].

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

3. COMBINATION OF GLR AND ICR

Algorithm 2 explains our proposed method for inter-cluster distance measurement within the framework of AHC. This

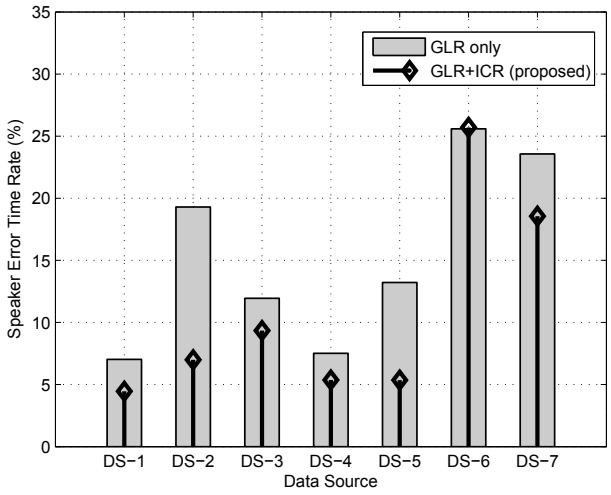

Fig. 1. Comparison of AHC with GLR only and AHC with the proposed method (GLR+ICR) in terms of the minimum possible speaker error time rates for various data sources.

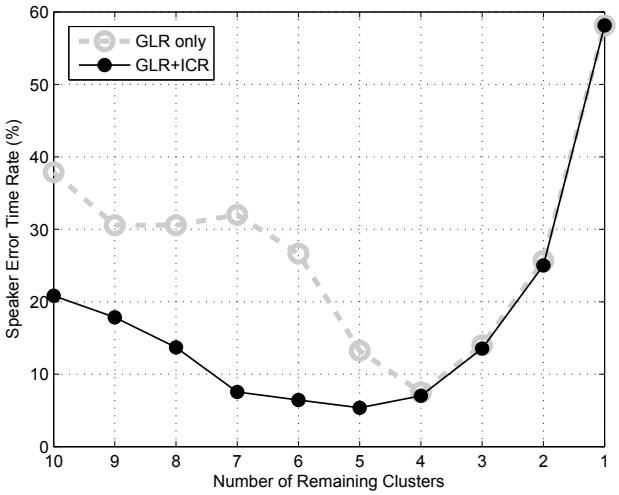

Fig. 2. Final 10 merging recursions in AHC for DS-4.

method for selecting the closest pair of clusters basically depends upon GLR at every recursion of AHC, but additionally considers ICR when all remaining clusters contain data samples of more than 10 seconds[1]. The proposed method is motivated by the followings:

1. As AHC proceeds to the end, erroneous selection of clusters for merging becomes much more detrimental to the minimum possible speaker error time rate than that at earlier recursions. This is because average cluster size increases as merging recursions in AHC continue, and thus incorrect merging of such large size clusters would raise the error rate much more than that of small size clusters. Therefore, inter-cluster distance measurement needs to be more accurate at the later recursions of AHC.

2. ICR is hypothesized to be complementary to GLR at the later recursions of AHC, specifically when all remaining clusters contain data samples of more than 10 seconds. This is based on the assumption that 10-second-long data samples are large enough for reliable ICR, i.e., do not need more information to represent speaker characteristics.

To consider both GLR and ICR at the later recursions of AHC, the proposed method utilizes the weighted sum of rankings in terms of GLR and ICR as a means of information fusion. The specific reason why such a high level fusion strategy is used in this case is because GLR is empirically shown to have wider variance than ICR for given cluster pairs, and thus low level fusion strategies like score normalization could cause GLR to be extremely dominant over ICR in decision of clusters for merging. The weighting factors $w_{\text{GLR}}^{k,l}$ and $w_{\text{ICR}}^{k,l}$ of the proposed method for measuring distance between a pair of clusters $\hat{C}_k$ and $\hat{C}_l$ are dynamically determined as follows:

$$w_{\text{GLR}}^{k,l} = f\left\{\frac{\text{GLR}(\hat{C}_k,\hat{C}_l)-\mu_{\text{GLR}}}{\sigma_{\text{GLR}}}\right\}$$

$$w_{\text{ICR}}^{k,l} = f\left\{\frac{\text{ICR}(\hat{C}_k,\hat{C}_l)-\mu_{\text{ICR}}}{\sigma_{\text{ICR}}}\right\}$$

where μ and σ are mean and standard deviation for the entire pairs of (remaining) clusters in terms of GLR or ICR, and $f(\cdot)$ is a cumulative density function for normal distribution with zero mean and unit variance. With these weighting factors, the proposed method chooses a pair of clusters having the smallest weighted sum of rankings of GLR and ICR as the one for merging.

Fig. 1 illustrates comparison of AHC with GLR only and AHC with the proposed method in terms of the minimum possible speaker error time rate. This figure shows us how much error rate (achievable) could be further lowered by the proposed method. We can clearly observe from the figure that the proposed method helps AHC achieve better clustering performance for every data source, except for DS-6 where no significant gain was obtained. Note that improvement for DS-2, 5, and 7 is relatively large, which leads to overall performance improvement due to the proposed method by 29.94% (relative).

The additional advantage offered by the proposed method is explicitly illustrated in Fig. 2, which shows the final 10 merging recursions for DS-4 as a clear example comparing AHC with GLR only and AHC with the proposed method.

[1] In this paper, we conservatively assume that 10-second-long data samples include sufficient information for capturing speaker characteristics, based on [6] where it was reported that almost perfect speaker identification accuracies were obtained with 10-second-long testing speech.

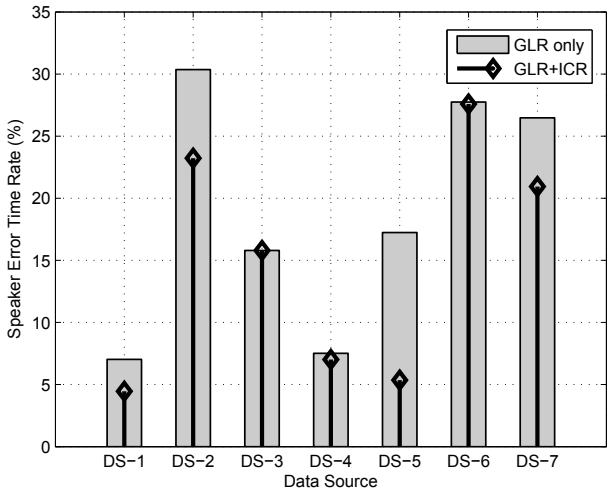

Fig. 3. Comparison of AHC with GLR only and AHC with the proposed method when a recursion stopping method for AHC is applied.

From this figure, we see that AHC with the proposed method makes lower speaker error time rates across merging recursions than AHC with GLR only, which means that there could be more chances to obtain a low error rate even though the optimal stopping point was not estimated exactly. Fig. 3 presents this advantage more clearly. This figure compares AHC with GLR only and AHC with the proposed method when a recursion stopping method for AHC is applied. The stopping method used here is what we proposed in [4], which was verified to be superior to a conventional BIC-based one [7] in terms of robustness to data source variation. From the figure, we can see that even in cases of incorrect estimation of the optimal stopping points (e.g., for DS-2, 3, and 4) AHC with the proposed method provides better performance than its counterpart. In other words, error rate increase due to mismatch between the stopping point estimated and the optimal one in AHC with the proposed method for DS-2, 3, and 4 are smaller than those in AHC with GLR only.

4. CONCLUSIONS

In this paper, we addressed the drawback of GLR as an inter-cluster distance measure within the framework of AHC. The tendency that GLR is affected by the size of clusters under consideration could influence effectiveness in distance measurement, which in turn might lead to degradation in AHC performance. To tackle this problem, we proposed a new distance measure combining GLR and ICR. The latter compensates for the undesirable tendency of the former and thus plays a critical role as a complementary criterion.

One potential future work would be to analytically identify the lower bound for cluster size that guarantees ICR to be reliable as a statistical distance measure between clusters. In this paper, we tried to avoid the possibility that ICR would not work properly, by applying ICR only when all remaining clusters contain data samples of more than 10 seconds under the empirically established assumption that such clusters are large enough for reliable ICR. This assumption worked for the data sources used for the experiments presented in the paper, but might be violated for other data sources where remaining clusters at the later recursions of AHC would be still too small to reveal the respective speaker characteristics completely. Clear identification of the bound mentioned could give us more flexibility to generalize our proposed method across different data domains.

Another future work might be about how to optimally fuse two different statistical information on the same object. In this paper, we used the weighted sum of rankings in terms of GLR and ICR for that purpose, but it is not theoretically proven to be optimal to the task considered in this paper. Establishing more systematic frameworks for selection of information fusion methods could be one of directions.

A final remark is that the method introduced in this paper can be used within clustering frameworks (other than AHC) wherever inter-cluster distance measurement is required, e.g., competitive learning or leader-follower clustering [2].

5. REFERENCES

[1] 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.

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

[3] H. Gish, M. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *Proc. 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 873-876, May 1991.

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

[5] K. J. Han, S. Kim, and S. S. Narayanan, "Robust speaker clustering strategies to data source variation for improved speaker diarization," *2007 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 262-267, Dec. 2007.

[6] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3(1), pp. 72-83, Jan. 1995.

[7] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127-132, Feb. 1998.