A Robust Stopping Criterion for Agglomerative Hierarchical Clustering in a Speaker Diarization System

Kyu J. Han, Shrikanth S. Narayanan

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

Abstract

Agglomerative hierarchical clustering (AHC) is an unsupervised classification strategy of merging the closest pair of clusters recursively, and has been widely used in speaker diarization systems to classify speech segments by speaker identity. The most critical part in AHC is how to automatically stop the recursive process at the point when clustering error rate reaches its lowest possible value, for which a BIC-based stopping criterion has been widely used. However, this criterion is not robust to data source variation. In this paper, we examine the criterion to establish the cause for the robustness issue and, based on this, propose an improved stopping criterion. Experimental results based on meeting conversation excerpts randomly chosen from various meeting speech corpora indicate that the proposed criterion is superior to the BIC-based one, showing that clustering error rate is improved on average by 7.28% (absolute) and 34.16% (relative).

Index Terms: Speaker diarization, agglomerative hierarchical clustering, BIC-based stopping criterion

1. Introduction

Speaker diarization refers to the process that automatically produces the transcription of “who spoke when” for a given audio clip [1]. This transcription can provide speaker-perspective statistics for a given audio clip such as frequency of speaking turn change, average speaking time per turn, number of speakers, speaking time distribution over speakers, and so on. These can be used as fundamental features in automatic content analysis of multimedia data. The transcription can also enable one to pick out all the speech uttered by a specific speaker to generate/train a speaker model, which in turn can be utilized for unsupervised speaker adaptation. Because of its broad significance, speaker diarization has drawn much attention since its introduction and is currently one of main evaluation categories in the Rich Transcription Evaluation led by NIST.

A basic speaker diarization system consists of three main steps, which are processed successively right after a feature extraction step. First, ‘speech/non-speech detection’ separates target speech regions from a given audio clip. Then, ‘speaker change detection’ identifies potential speaker changing points in each speech region and consequently divides the speech regions into further speaker-specific segments. Lastly, ‘speaker classification’ classifies the resultant segments by speaker identity. Among these, the last step is the most crucial in the system because its result affects final system performance the most. Typically, agglomerative hierarchical clustering (AHC) has been used for this purpose due to its simple structure and acceptable level of performance.

AHC is an unsupervised classification strategy of merging the closest pair of clusters recursively. Because of the recursive structure, the most critical part in AHC is how to automatically stop the process at the point[1] when clustering error rate reaches its lowest possible value; a BIC-based stopping criterion [2] has been widely used for this purpose. Presently, this criterion is built in to most of the state-of-the-art speaker diarization systems developed by various research sites [3]-[7]. The criterion checks whether or not the closest pair of clusters are homogeneous in terms of speaker identity before every merging, by quantifying statistical distance between the clusters and regarding the clusters as homogenous if the distance is less than a threshold and as not, otherwise. Then, it stops the process right away when the clusters are decided to be not homogeneous, based on the rationale that if the closest pair were not homogeneous then the rest of the pairs would not be as well, and thus there would be no more merging needed. In short, the criterion stops the recursive process in AHC at the point[2] when the distance measured by itself exceeds the threshold for the first time. The distance is measured by computing the difference between two hypotheses for the clusters (H_0: Unmerging and H_A: Merging) in terms of the sum of the log-likelihoods of the whole feature samples within the clusters.

The BIC-based stopping criterion however has the intrinsic drawback that it is not robust to data source variation. In other words, the stopping point estimated by the criterion is not always identical to the optimal one for every possible data source. A main reason for this drawback is that distance measurement in the criterion depends upon the log-likelihood sum affected by the total number of feature samples within clusters under consideration. Therefore, a pair of homogeneous clusters where there are a large number of feature samples could be regarded as more distant to each other (less homogeneous) than a pair of heterogeneous clusters where there are a small number of feature samples. Such dependency makes it difficult to set a global threshold for the distance, because the cardinalities of every closest pair of clusters encountered during AHC are completely data-dependent.

Based on this, we propose a new statistical distance measure in this paper, named information change rate (ICR), which computes difference between H_0 and H_A in terms of average log-likelihood for the whole feature samples within clusters un-

This work was supported by NSF and U.S. Army.

[1] Let us call this *the optimal stopping point*.
[2] Let us call this *the estimated stopping point*.

Table 1: *Development data. 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.).*

	Development Data		
	ICSI-I	ICSI-II	ICSI-III
N_s	7 (5:2)	7 (5:2)	6 (4:2)
T_s	1064.9	931.3	1148.5
N_t	417	278	243
T_a	2.5	3.3	4.7

Table 2: *Evaluation data. The notation is same as above.*

	Evaluation Data				
	ICSI-IV	ICSI-V	NIST-I	NIST-II	ISL-I
N_s	5 (3:2)	7 (6:1)	4 (3:1)	4 (3:1)	4 (2:2)
T_s	674.5	2336.3	835.7	443.4	477.7
N_t	175	610	178	74	118
T_a	3.8	3.8	4.7	5.9	4.0

der consideration. This measure also needs a threshold to decide whether the clusters are homogeneous, but is not affected by the cardinalities of the clusters. As will be discussed later in the paper, the measure tells us how much information (entropy) would be changed (increased) by merging the clusters, and works only if both of the clusters have enough feature samples to represent speaker characteristics. For this reason, our ICR-based stopping criterion for AHC waits until the recursive process is completed so that we can handle clusters of large size. Then, it traces back checking if the latest merging occurred between homogeneous clusters, which is based on the same rationale used in the BIC-based stopping criterion but applies it in a reverse direction. The estimated stopping point in this criterion is thus when ICR between the clusters exceeds the threshold for the last time.

This paper is organized as follows. In Section 2, the setup and data sources for our experiments are described. Then, in Section 3 the BIC-based stopping criterion for AHC is analyzed in terms of why it is not robust to data source variation. In Section 4, we propose the ICR approach, which is followed by comparing the ICR-based stopping criterion with the BIC-based one. In Section 5, we conclude this paper with comments on future work.

2. Experimental Setup

For the purpose of the work in the present paper, we assume that both the speech/non-speech detection and speaker change detection steps are perfectly done for every data source, which allows us to concentrate on the AHC aspects. For this, we manually segmented each data source based on a reference transcription before experiments. In order to avoid distraction in the performance analysis that might result from overlap between the speech segments, we excluded all the segments involved in any overlap with others during data preparation.

Tables 1 and 2 report the chosen development and evaluation data respectively, 8 meeting conversation excerpts (of total length approximately 130 minutes long) randomly chosen from the ICSI, NIST, and ISL meeting speech corpora (LDC2004S02, LDC2004S09, and LDC2004S05). They have distinct speaker-relevant characteristics in terms of number of speakers, gender distribution over speakers, total speaking time, number of speaking turn changes, and average speaking time per turn.

In this paper, mel-frequency cepstral coefficients (MFCCs) are used as general acoustic features. Through 23 mel-scaled filter banks, a 12-dimensional MFCC vector is generated for every 20ms-long frame of a given audio clip. Frame shift rate is fixed at 10ms so that there can be overlap between two adjacent frames. Generalized likelihood ratio (GLR) [8] is used for AHC as an inter-cluster distance measure to select the closet pair among remaining clusters. For measurement of clustering error rate, we used a scoring tool, i.e., md-eval-v21.pl, distributed by NIST [http://www.nist.gov/speech/tests/rt/rt2007].

3. BIC-based Stopping Criterion

In this section, we analyze why the BIC-based stopping criterion is not robust to data source variation. For this, let us first investigate how this criterion works in AHC, as shown below:

1. For the closest pair of clusters C_X and C_Y consisting of features $X = \{x_1, x_2, \cdots, x_M\}$ and $Y = \{y_1, y_2, \cdots, y_N\}$ respectively, compute BIC scores for H_0 and H_A, which are defined as follows:
   • H_0: C_X and C_Y are left unmerged. The clusters are modeled by distinct Gaussian distributions θ_X and θ_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 a Gaussian distribution $\theta_{X \cup Y}$, whose model parameters are estimated by way of maximizing the likelihood of $X \cup Y$.

   $$BIC\left(H_0\right) = \ln P\left(X \cup Y | H_0\right) - \lambda \cdot 2\mathcal{P}$$
   $$BIC\left(H_A\right) = \ln P\left(X \cup Y | H_A\right) - \lambda \cdot \mathcal{P},$$

   where λ is a tuning parameter such that it minimizes average clustering error rate for development data, $\mathcal{P}$ is a penalty term, i.e.,

   $$\mathcal{P} = \tfrac{1}{2}\left\{k + \tfrac{1}{2}k\left(k+1\right)\right\}\ln\left(M+N\right),$$

   and k is feature dimension.
2. Compute $\Delta BIC = BIC\left(H_0\right) - BIC\left(H_A\right)$ and compare it with 0.

   $$\begin{aligned} \Delta BIC &= \ln P\left(X \cup Y | H_0\right) - \lambda \cdot 2\mathcal{P} - \\ &\quad \ln P\left(X \cup Y | H_A\right) + \lambda \cdot \mathcal{P} \\ &= \ln \frac{P\left(X \cup Y | H_0\right)}{P\left(X \cup Y | H_A\right)} - \lambda \cdot \mathcal{P} \\ &= \ln\left(\text{GLR}\right) - \lambda \cdot \mathcal{P} \underset{H_A}{\overset{H_0}{\gtrless}} 0. \end{aligned} \tag{1}$$
3. If $\Delta BIC < 0$ or $BIC\left(H_0\right) < BIC\left(H_A\right)$, decide that C_X and C_Y are homogeneous (and merge them). Otherwise, decide that they are heterogeneous (and stop the process).

In the above, Eq. (1) can be re-written as follows:

$$\begin{aligned} \ln\left(\text{GLR}\right) &\underset{H_A}{\overset{H_0}{\gtrless}} \lambda \cdot \frac{1}{2}\left\{k + \frac{1}{2}k\left(k+1\right)\right\}\ln\left(M+N\right) \\ &\underset{H_A}{\overset{H_0}{\gtrless}} C \cdot \ln\left(M+N\right), \end{aligned} \tag{2}$$

where C is a constant. Considering that GLR is a statistical inter-cluster distance measure, as a consequence, this criterion appears to quantify the statistical distance between clusters under consideration and decide whether or not they are homogeneous comparing the distance with the floating threshold in proportion to the sum of the cardinalities of the clusters. Note that

Table 3: *Clustering error rates (CERs) for evaluation data when the BIC-based stopping criterion with $\lambda = 12.0$ is used. N_c: number of remaining clusters when CER is achieved and CER*: lowest possible CER.*

	ICSI-IV	ICSI-V	NIST-I	NIST-II	ISL-I
CER (N_c)	11.95% (5)	35.79% (8)	9.13% (3)	22.72% (3)	27.00% (2)
CER* (N_c)	11.95% (6)	10.62% (6)	9.13% (3)	7.63% (4)	27.00% (2)

λ is pre-tuned based on development data. In this paper, it is tuned to be 12.0 where all the optimal stopping points for ICSI-I, II, and III were precisely detected by the criterion.

Table 3 shows how well the criterion performed on evaluation data. In the table, the first row contains the clustering error rates obtained when the recursive process in AHC was stopped at the estimated stopping points while the second row consists of the lowest possible clustering error rates that would be achieved at the optimal stopping points. From the comparison of the results in the rows of the table, we can see that the criterion did not correctly pick out the optimal stopping points for ICSI-V and NIST-II although it did for the rest. This confirms that the criterion is not robust to data source variation. A noticeable observation is that the effect of missing the optimal stopping points can be detrimental.

A probable reason for this robustness issue in the criterion could be revealed by examining the $\ln(\text{GLR})$ term, which can be re-written as follows:

$$\begin{aligned} & \ln P\left(X \cup Y | H_0\right) - \ln P\left(X \cup Y | H_A\right) \\ = & \sum_{i=1}^{M} \ln p\left(x_i | \theta_X\right) + \sum_{i=1}^{N} \ln p\left(y_i | \theta_Y\right) - \\ & \left\{ \sum_{i=1}^{M} \ln p\left(x_i | \theta_{X \cup Y}\right) + \sum_{i=1}^{N} \ln p\left(y_i | \theta_{X \cup Y}\right) \right\}. \end{aligned}$$

From this, we can see that the $\ln(\text{GLR})$ term depends upon the log-likelihood sum affected by the sum of the cardinalities of clusters under consideration. This means that a pair of homogeneous clusters where there are a large number of feature samples could be regarded by the criterion as more distant to each other than a pair of heterogeneous clusters where there are a small number of feature samples. Such dependency makes it difficult to set a global threshold for the $\ln(\text{GLR})$ term, because the cardinalities of every closest pair of clusters encountered during AHC are completely data-dependent. The floating threshold in Eq. (2) seems as if to compensate for this undesirable feature of the $\ln(\text{GLR})$ term, but actually does not do so well, which is explicitly shown in Figure 1. In this figure, $\ln(\text{GLR})$ and $\lambda \cdot \mathcal{P}$ for ICSI-I, V, and NIST-II are compared at the final 10 merging processes in AHC. We can easily see from the figure that the threshold curves ($\lambda \cdot \mathcal{P}$) are almost flat and do not vary much across data sources compared to the distance curves ($\ln(\text{GLR})$). This in fact resulted in the disagreement between the estimated and optimal stopping points for the ICSI-V and NIST-II data sets.

4. ICR-based Stopping Criterion

In order to tackle the aforementioned data dependency in $\ln(\text{GLR})$, we propose a new statistical distance measure in this section by normalizing $\ln(\text{GLR})$, which is defined as follows:

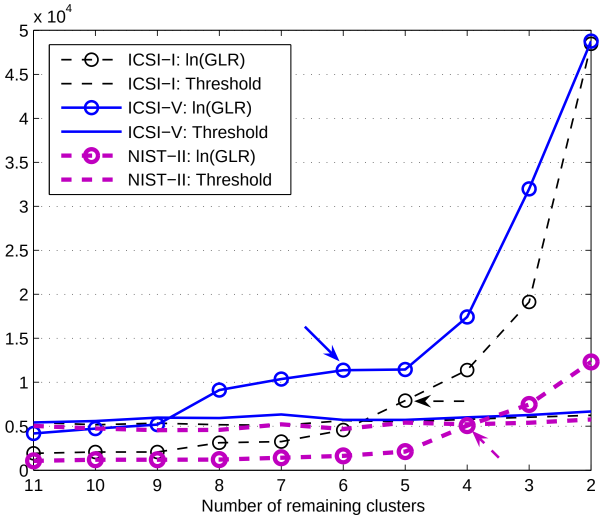

Figure 1: *Comparison of ICSI-I, V, and NIST-II in terms of $\ln(GLR)$ and $\lambda \cdot \mathcal{P}$ where $\lambda = 12.0$. The arrows indicate $\ln(GLR)$ values at the optimal stopping points. The estimated stopping point for each data source is when $\ln(GLR) > \lambda \cdot \mathcal{P}$ for the first time.*

$$\text{ICR} \triangleq \frac{1}{M+N} \ln(\text{GLR}). \tag{3}$$

The reason why we named this measure ICR, which is an acronym for information change rate, is because it measures how much information (entropy) would be changed (increased) by merging clusters under consideration. This information-theoretic perspective is based on the following, according to [2]:

$$\begin{aligned} \text{ICR} &= \frac{1}{M+N} \cdot \frac{1}{2} (M+N) \ln |\Sigma_{\theta_{X \cup Y}}| - \\ & \quad \frac{1}{M+N} \cdot \frac{1}{2} \left(M \ln |\Sigma_{\theta_X}| + N \ln |\Sigma_{\theta_Y}|\right) \\ &= \frac{1}{2} \ln (2\pi e)^k |\Sigma_{\theta_{X \cup Y}}| - \\ & \quad \frac{M \cdot \frac{1}{2} \ln (2\pi e)^k |\Sigma_{\theta_X}| + N \cdot \frac{1}{2} \ln (2\pi e)^k |\Sigma_{\theta_Y}|}{M+N} \\ &= H\left(\theta_{X \cup Y}\right) - \frac{M \cdot H\left(\theta_X\right) + N \cdot H\left(\theta_Y\right)}{M+N}, \end{aligned} \tag{4}$$

where H is entropy. Since it is obvious that merging heterogeneous clusters would increase entropy much more than merging homogeneous ones, this measure with a proper (fixed) threshold can be a reasonable alternative to $\ln(\text{GLR})$ with the floating threshold in the previous section. Given the threshold η, $ICR < \eta$ means that clusters under consideration are homogeneous.

A minor issue is that ICR works only if both clusters under consideration are large enough thus have a number of feature samples to fully represent speaker characteristics. Otherwise, ICR could exceed the threshold even if the clusters were homogeneous. In order to resolve this issue, we suggest an ICR-based stopping criterion by applying the same rationale used in the BIC-based one in a reverse direction, i.e., waiting until the recursive process in AHC is completed and tracing back to check if the latest merging occurred between homogeneous clusters. This helps ICR measurement to handle large size clusters. The

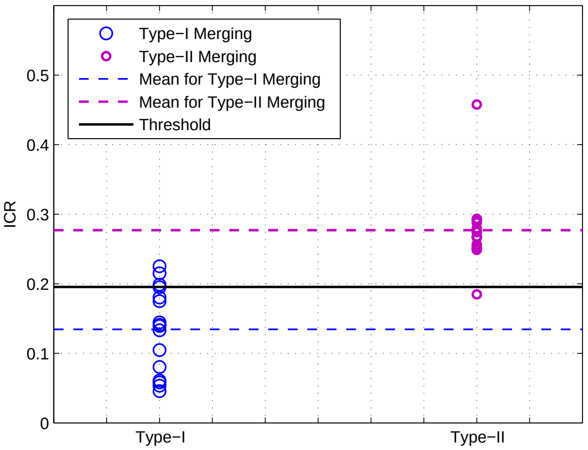

Figure 2: *ICR values at final 10 merging processes in AHC for each of development data, i.e., total 30 ICR values are plotted. The solid line in the middle indicates $\eta = 0.19547$.*

estimated stopping point in this criterion is thus when ICR between the clusters exceeds the threshold for the last time.

Figure 2 displays how well ICR distinguishes merging between homogeneous clusters (Type-I Merging) from merging between heterogeneous clusters (Type-II Merging) at the final 10 merging processes in AHC for each of ICSI-I, II, and III. The particular reason that we picked up the final 10 merging processes for this experiment is that, as mentioned above, they are likely to be the merging processes that occurred between the clusters of large size to fully represent speaker characteristics. Based on these results, we set η to be 0.19547, i.e., mean plus standard deviation for the ICR values for Type-I Merging under the assumption of Gaussian distribution. Given the threshold, we identified three outliers for Type-I Merging and one outlier for Type-II Merging, but most of them were the merging processes that occurred between clusters of relatively small size and thus regarded as negligible.

Table 4 shows experimental results on the evaluation data when the ICR-based stopping criterion with $\eta = 0.19547$ is used. Compared to the same experiments performed using the BIC-based one (shown in Table 3), the results in this table verify that the ICR-based stopping criterion is superior to the BIC-based one. Note that this criterion exactly detected the optimal stopping points for ICSI-V and NIST-II as well as NIST-I and ISL-I. Even for ICSI-IV where the optimal stopping point was not detected, difference between clustering error rates at the estimated stopping point and the optimal one is not huge. Therefore, we can conclude that the ICR-based stopping criterion is more robust to data source variation and consequentially leads to improvement in terms of clustering error rate. In our case, the resultant improvement was 7.28% (absolute) and 34.16% (relative) on average.

Table 4: *Clustering error rates (CERs) for evaluation data when the ICR-based stopping criterion with $\eta = 0.19547$ is used. For every data source except ICSI-IV, CER = CER*.*

	ICSI-IV	ICSI-V	NIST-I	NIST-II	ISL-I
CER (N_c)	15.80% (4)	10.62% (6)	9.13% (3)	7.63% (4)	27.00% (2)
CER* (N_c)	11.95% (6)	10.62% (6)	9.13% (3)	7.63% (4)	27.00% (2)

5. Conclusions

In this paper, we established the cause for the potential data-based robustness issue in the widely-used BIC-based stopping criterion for AHC and, based on it, proposed a novel alternative. Although both methods take into account the total number of feature samples in a pair of clusters under consideration, the proposed one does so more flexibly as the total number of the feature samples increases. The effectiveness of the proposed stopping criterion was empirically verified through experiments on the data drawn from a variety of meeting speech corpora.

One of potential future directions is to identify the optimal cluster size for ICR to work well. This might be dependent on several factors other than the cardinalities of the clusters. Although it is not a serious problem when a speaker diarization system handles long audio clips, this issue could be challenging in either selecting the threshold η or running the ICR-based stopping criterion on data sources from other domains, like TV shows.

Another future direction would be to test the ICR-based stopping criterion without the assumption of perfect speech/non-speech and speaker change detection. In those cases, the clusters can be more severely contaminated and possibly affect the performance of this criterion.

6. Acknowledgement

The authors would like to thank NSF and U.S. Army for supporting this work.

7. References

[1] Tranter, S. E. and Reynolds, D. A., “An overview of automatic speaker diarization systems,” IEEE Trans. Audio, Speech, and Language Proc., 14(5):1557–1565, 2006.

[2] Chen, S. S. and Gopalakrishnan, P. S., “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” Proc. DARPA BNTU Workshop, 127–132, 1998.

[3] Moraru, D., Besacier, L., Meignier, S., Fredouille, C., and Bonastre, J., “Speaker diarization in the ELISA consortium over the last 4 years,” Proc. Fall 2004 RT-04 Workshop, 2004.

[4] Sinha, R., Tranter, S. E., Gales, M. J. F., and Woodland, P. C., “The Cambridge university March 2005 speaker diarisation system,” Proc. INTERSPEECH 2005, 2437–2440, 2005.

[5] Anguera, X., Wooters, C., Peskin, B., and Aguilo, M., “Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 siarization system,” Proc. MLMI 2005, 402–414, 2005.

[6] Barras, C., Zhu, X., Meignier, S., and Gauvain, J., “Improving speaker diarization,” Proc. Fall 2004 RT-04 Workshop, 2004.

[7] Reynolds, D. A. and Torres-Carrasquillo, P. A., “The MIT Lincoln laboratory RT-04F diarization systems: Applications to broadcast news and telephone conversations,” Proc. Fall 2004 RT-04 Workshop, 2004.

[8] Gish, H., Siu, M., and Rohlicek, R., “Segregation of speakers for speech recognition and speaker identification,” Proc. ICASSP 1991, 873–876, 1991.