interspeech 2008 incorporating SST 08

Agglomerative Hierarchical Speaker Clustering using Incremental Gaussian Mixture Cluster Modeling

Kyu J. Han, 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
kyuhan@usc.edu, shri@sipi.usc.edu

Abstract

This paper proposes a novel cluster modeling method for inter-cluster distance measurement within the framework of agglomerative hierarchical speaker clustering, namely, *incremental Gaussian mixture cluster modeling*. This method uses a single Gaussian distribution to model each initial cluster, but represents any newly merged cluster using a distribution whose pdf is the weighted sum of the pdf's of the respective model distributions for the clusters involved in the particular merging process. As a result, clusters are smoothly transitioned to be modeled by Gaussian mixtures whose components are incremented as merging recursions continue during clustering. The proposed method can overcome the limited cluster representation capability of conventional single Gaussian cluster modeling. Through experiments on various sets of initial clusters, it is demonstrated that our approach consequently improves the reliability of speaker clustering performance.

Index Terms: agglomerative hierarchical speaker clustering, inter-cluster distance measure, incremental Gaussian mixture cluster modeling

1. Introduction

Speaker clustering refers to the process of automatically classifying speech segments by speaker identity, especially in an unsupervised manner. This process is essential for speaker diarization and unsupervised speaker adaptation, and for this reason it has been attracting much attention in the field of speaker and even speech recognition.

Agglomerative hierarchical clustering (AHC) has been the most popular strategy for speaker clustering in the recent decade because it has a simple processing structure, but provides an acceptable level of performance. The details of this strategy are as follows: it considers given speech segments as individual initial clusters, and recursively merges the closest pair of clusters in terms of speaker characteristics. Its recursive process continues until it is decided that extra cluster merging would not improve speaker clustering performance any further.

In order for AHC to achieve reliable speaker clustering performance, two critical questions need to be answered properly: 1) how to select the closest pair of clusters for merging at every recursion step of AHC and 2) how to decide the optimal (recursion) stopping point. For given speech segments, the way of answering the first question determines the error rate that would be achieved by AHC at the optimal stopping point, while the way of answering the second question determines the error rate actually obtained by AHC. Of these two questions we focus on the first one in this paper, because getting error rate at the optimal stopping point lowered as much as possible would be a desirable prior condition for better estimation of the optimal stopping point; as a consequence, AHC could provide a high clustering accuracy overall.

To address this question of selecting the closest pair of clusters for merging at every recursion step of AHC in a state-of-the-art way, Bayesian information criterion (BIC) has been widely utilized since [1]. The BIC-based inter-cluster distance measure requires clusters to be modeled by statistical distributions for its own purpose, which thus causes the choice of cluster modeling methods to be critical to the accuracy of the measure. An ideal cluster modeling method should keep clusters well distinguished in terms of speaker characteristics during AHC.

There currently exist two typical cluster modeling methods for the BIC-based inter-cluster distance measure. One is to model each cluster using a single Gaussian distribution [1, 2], which is simple but might not represent speaker-specific information in clusters adequately enough as merging recursions in AHC continue and the average size of the clusters remaining thus increases, because speaker characteristics are better represented by a complex distribution with multiple modes than by a simple distribution with only one mode [3]. For this reason, a Gaussian mixture model (GMM) is widely regarded as a more suitable (individual) cluster model [4, 5], but this method has a higher computational complexity because of the expectation-maximization (EM) procedure to train GMMs. Besides, this method is not appropriate to be used during the early recursion steps of AHC because some initial clusters might not contain sufficient data to train such complex distributions as GMMs.

In order to tackle this dilemma in the conventional cluster modeling methods for the BIC-based inter-cluster distance measure within the framework of AHC, we propose a novel approach in this paper: *incremental Gaussian mixture cluster modeling*. This method gradually increments the complexity of cluster models from single Gaussian distributions to GMMs with multiple mixture components as more data (in clusters) become available, for better and more dynamic representation of clusters during AHC. In addition, it does not let any GMM trained by the EM procedure, which makes it feasible in terms of computational complexity to use GMMs as cluster models.

The rest of the paper is organized as follows. In Section 2, the further details of the conventional cluster modeling methods are described. Then, we introduce the incremental Gaussian mixture cluster modeling method in Section 3, and compare it

This research was supported in part with funds from NSF, Army and DHS.

with the single Gaussian cluster modeling method in terms of clustering accuracy in Section 4. In Section 5, concluding remarks and future research directions are given.

2. Conventional Cluster Modeling Methods

In this section, we further investigate the details of the BIC-based inter-cluster distance measure within the framework of AHC, from a viewpoint of cluster modeling. For this purpose, let us consider a certain recursion step during AHC. In order to select the closest pair of clusters for merging at the recursion step, the ΔBIC values of all possible cluster pairs are compared. Then, the cluster pair having the smallest ΔBIC value is picked for merging. Suppose that a certain pair of clusters C_x and C_y at the recursion step consist of n-dimensional data (i.e., acoustic feature vectors) $x = \{x_1, x_2, \cdots, x_M\}$ and $y = \{y_1, y_2, \cdots, y_N\}$, respectively. The ΔBIC value of the pair is presented as follows:

$$\begin{aligned}
&\Delta\mathrm{BIC}\left(C_x, C_y\right) \\
&\quad = \mathrm{BIC}\left(C_x, C_y|\mathcal{H}_1\right) - \mathrm{BIC}\left(C_x, C_y|\mathcal{H}_2\right) \\
&\quad = \ln p\left(x \cup y|\mathcal{H}_1\right) - \frac{\lambda}{2} \cdot N_{\mathcal{H}_1} \cdot \ln N_{\text{total}} - \\
&\qquad \left\{\ln p\left(x \cup y|\mathcal{H}_2\right) - \frac{\lambda}{2} \cdot N_{\mathcal{H}_2} \cdot \ln N_{\text{total}}\right\} \\
&\quad = \ln \frac{p\left(x \cup y|\mathcal{H}_1\right)}{p\left(x \cup y|\mathcal{H}_2\right)} - \frac{\lambda}{2}\left(N_{\mathcal{H}_1} - N_{\mathcal{H}_2}\right) \ln N_{\text{total}}, \quad (1)
\end{aligned}$$

where

- $\mathcal{H}_1$ (unmerging hypothesis): C_x and C_y are hypothesized to be left unmerged.
- $\mathcal{H}_2$ (merging hypothesis): C_x and C_y are hypothesized to be merged so as to become a new, larger cluster C_z, where $z = x \cup y$.

In the above equation, λ (equal to be 1, ideally) is a tuning parameter, $N_{\mathcal{H}_1}$ and $N_{\mathcal{H}_2}$ are the numbers of parameters in the statistical distributions presenting the two hypotheses, respectively, and N_{total} is the total amount of data (i.e., the total number of acoustic feature vectors) in all clusters at the recursion step considered.

In order to present the two hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ statistically, a (multivariate) single Gaussian distribution with a full covariance matrix has been popularly used as a cluster model for each cluster considered in Eq. (1), i.e., C_x and C_y for $\mathcal{H}_1$, and C_z for $\mathcal{H}_2$, due to its simplicity. In this cluster modeling method, C_x, C_y, and C_z are modeled by normal distributions $\theta_x^{\mathcal{N}}$, $\theta_y^{\mathcal{N}}$, and $\theta_z^{\mathcal{N}}$, respectively, and the mean vectors and the (full) covariance matrices of the distributions are determined by way of maximizing the likelihoods of x, y, and z for $\theta_x^{\mathcal{N}}$, $\theta_y^{\mathcal{N}}$, and $\theta_z^{\mathcal{N}}$, respectively [1, 2, 6]. Therefore, Eq. (1) can be expanded further as follows:

$$\begin{aligned}
\Delta\mathrm{BIC}\left(C_x, C_y\right) \quad = \quad & \ln \frac{p\left(x|\theta_x^{\mathcal{N}}\right) p\left(y|\theta_y^{\mathcal{N}}\right)}{p\left(z|\theta_z^{\mathcal{N}}\right)} - \\
& \frac{\lambda}{2}\left\{n + \frac{1}{2}n\left(n+1\right)\right\} \ln N_{\text{total}}, \quad (2)
\end{aligned}$$

where n is the dimension of data.

Despite its popularity, this method however has a critical issue, i.e., the average size of the clusters handled increases as merging recursions in AHC continue, whereas *a single Gaussian distribution has a limited capability in representing clusters of large data size, especially in terms of speaker characteristics* [3]. This could degenerate discernibility between heterogeneous clusters in terms of speaker characteristics at the late recursion steps of AHC, and hence cause speaker clustering performance to degrade severely.

In this context, a GMM is currently regarded as a better cluster model for the BIC-based inter-cluster distance measure within the framework of AHC. In this cluster modeling method, Eq. (1) can be expanded as

$$\Delta\mathrm{BIC}\left(C_x, C_y\right) \quad = \quad \ln \frac{p\left(x|\theta_x^{\mathcal{G}}\right) p\left(y|\theta_y^{\mathcal{G}}\right)}{p\left(z|\theta_z^{\mathcal{G}}\right)}, \quad (3)$$

where $\theta_x^{\mathcal{G}}$, $\theta_y^{\mathcal{G}}$, and $\theta_z^{\mathcal{G}}$ are GMMs with diagonal covariance matrices for modeling C_x, C_y, and C_z, respectively. The parameters of the GMMs (i.e., a mean vector, a variance vector, and a weight for each mixture component) are trained by the EM procedure using x, y, and z, respectively. A notable thing is that there is no penalty term in Eq. (3), i.e., the second term in the right side of Eq. (1), because $N_{\mathcal{H}_2}$ is set to equal $N_{\mathcal{H}_1}$ by equalizing the number of mixture components in $\theta_z^{\mathcal{G}}$ with the total number of mixture components in $\theta_x^{\mathcal{G}}$ and $\theta_y^{\mathcal{G}}$ [4, 5].

However, this method has drawbacks as well. First, *the EM procedure used for GMM training requires a significant amount of processing time* in proportion to the data size of the clusters considered and the number of mixture components in the respective GMMs. Furthermore, *this method can not be used during the early recursion steps of AHC*, because most of the initial clusters handled by speaker clustering usually do not contain sufficient data to train multi-component GMMs; cluster models might be thus overfitted[1].

For all of these reasons, there still exists a strong demand for a more appropriate and more feasible cluster modeling approach to be used for the BIC-based inter-cluster distance measure within the framework of AHC. In the next section, we address this problem by proposing a novel cluster modeling method.

3. Proposed Cluster Modeling Method

In the previous section, we examined several issues in the conventional cluster modeling methods for the BIC-based inter-cluster distance measure within the framework of AHC:

- Single Gaussian cluster modeling
  1. Limited capability in representing speaker characteristics for clusters of large data size
- GMM cluster modeling
  2. EM procedure to require a significant amount of processing time
  3. Inapplicability to the early recursion steps of AHC

First, we consider the problems 1 and 3. One of the possible ways to simultaneously tackle these two problems would be to gradually increase the complexity of cluster models, e.g., from single Gaussian distributions through less complex GMMs with only a few mixture components to more complex GMMs with

[1] This drawback did not need to be considered seriously in [4], as the Viterbi segmentation applied prior to speaker clustering in the work naturally resulted in clusters of large data size enough to train multi-component GMMs.

a lot of mixture components, as merging recursions in AHC continue. Based on this rationale, we propose a novel cluster modeling method, namely, *incremental Gaussian mixture cluster modeling*, which works for the BIC-based inter-cluster distance measure as follows:

a) Models each initial cluster using a normal distribution with a full covariance matrix,

b) Models any merging-hypothesized cluster (e.g., C_z) using the distribution[2] whose pdf is the weighted sum of the pdf's of the respective model distributions for the two clusters considered (e.g., C_x and C_y), and

c) Recursively updates a model for any newly merged cluster using the one presenting the merging hypothesis $\mathcal{H}_2$ when distance between the clusters involved in the particular merging process was measured.

Such a recursive update enables not only cluster models' smooth transition from single Gaussian distributions to GMMs, but also gradual increase in the complexity of GMMs. In this cluster modeling method, Eq. (1) can be thus expanded as follows:

$$\Delta\text{BIC}\left(C_x, C_y\right) \quad = \quad \ln \frac{p\left(x|\theta_x^{\Lambda}\right) p\left(y|\theta_y^{\Lambda}\right)}{p\left(z|\theta_z^{\Lambda}\right)}, \qquad (4)$$

where θ_x^{Λ}, θ_y^{Λ}, and θ_z^{Λ} are (proposed) incremental Gaussian mixture models for the clusters considered, C_x, C_y, and C_z, respectively, ($\theta_x^{\Lambda} = \theta_x^{\mathcal{N}}$ and $\theta_y^{\Lambda} = \theta_y^{\mathcal{N}}$, if C_x and C_y are initial clusters) and

$$f(\theta_z^{\Lambda}) \quad = \quad \frac{M}{M+N} \cdot f(\theta_x^{\Lambda}) + \frac{N}{M+N} \cdot f(\theta_y^{\Lambda}). \qquad (5)$$

In the above equation, M and N are cardinalities for C_x and C_y, respectively, and $f(\theta)$ is the pdf of a distribution θ.

A notable thing is that our proposed method does not further train any GMM using the EM procedure, which enables us not to concern the problem 2 mentioned above any more. This is because we found out from a simple test for the effectiveness of the EM procedure on inter-cluster distance measurement that the EM procedure was actually of no benefit to cluster modeling for the BIC-based inter-cluster distance measure. The test results are given in Figure 1, which presents comparison of distance based on Eq. (3) for a particular pair of homogeneous clusters in terms of speaker characteristics (black) and for three different pairs of heterogeneous clusters (grey), as the number of iterations in the EM procedure increases. Interestingly, we can observe from the leftmost sub-figure that distance for the heterogeneous cluster pair becomes smaller than that for the homogeneous cluster pair as iterations in the EM procedure continue, which is completely undesirable. In addition, even in the other sub-figures, the EM procedure is demonstrated not to significantly improve discernibility[3] between clusters. These observations can be explained as follows: the EM procedure iteratively adapts the parameters of any GMM toward maximum likelihood, and thus increases $p\left(z|\theta_z^{\mathcal{G}}\right)$ in Eq. (3) regardless of homogeneity between the clusters considered, which however does not help increase inter-cluster discernibility and even might make it worse as shown in the leftmost sub-figure in Figure 1.

[2] As a consequence, this distribution has a mixed form of weighted Gaussian distributions, which is a GMM.

[3] We can roughly define this term as difference between distance for homogenous clusters and for heterogeneous clusters. The bigger such difference, the easier to discern (i.e., classify) clusters correctly.

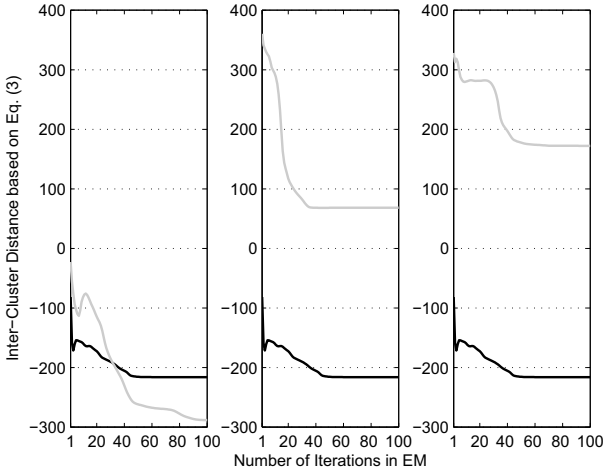

Figure 1: *Comparison of inter-cluster distance based on Eq. (3) for a particular pair of homogeneous clusters (black) and for three different pairs of heterogenous clusters (grey), along with the number of iterations in the EM procedure.*

In summary, the proposed incremental Gaussian mixture cluster modeling method has the advantage that it can keep clusters well represented in terms of speaker characteristics during AHC with feasible computational complexity. In the next section, this advantage is verified through experiments.

4. Experimental Results

4.1. Data Sources and Experimental Setup

Table 1 presents the data sources used for our experiments. These data sources are 15 sets of speech segments with approximately 4hr-long total duration, and are randomly chosen from the ICSI, NIST, and ISL meeting corpora. They are distinct from one another in terms of the number of speakers (N_s), gender distribution over speakers, the total speaking time (T_s), and the number of speech segments (N_{ss}).

For preparing each set of speech segments, we manually segmented each audio clip 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.

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

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

4.2. Experimental Results and Discussions

Figure 2 shows AHC performance comparison of incremental Gaussian mixture cluster modeling and single Gaussian cluster

Table 1: *Data source. N_s: number of speakers (male:female), T_s: total speaking time (sec.), and N_{ss}: number of speech segments.*

	Data Source				
	1	2	3	4	5
N_s	7 (5:2)	7 (5:2)	7 (6:1)	6 (4:2)	5 (1:4)
T_s	1064.9	931.3	2336.3	1148.5	805.1
N_{ss}	418	279	611	244	228
	Data Source				
	6	7	8	9	10
N_s	6 (5:1)	5 (5:0)	4 (4:0)	9 (7:2)	4 (3:1)
T_s	1664.9	1609.1	1475.9	659.7	443.4
N_{ss}	532	591	478	159	75
	Data Source				
	11	12	13	14	15
N_s	4 (3:1)	6 (4:2)	8 (4:4)	4 (2:2)	4 (0:4)
T_s	835.7	624.1	272.4	477.7	429.1
N_{ss}	179	144	93	119	95

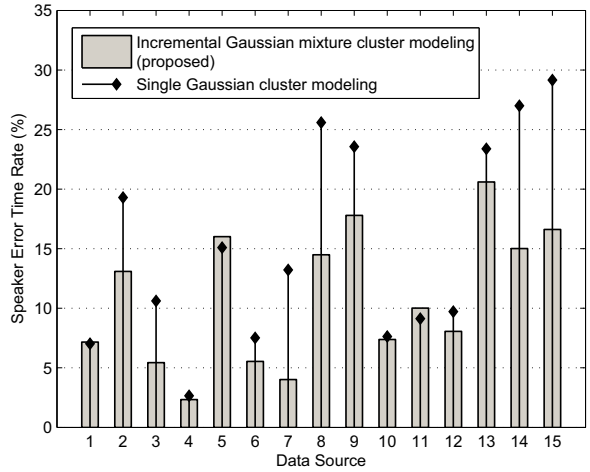

Figure 2: *AHC performance comparison of incremental Gaussian mixture cluster modeling and single Gaussian cluster modeling.*

modeling in terms of speaker error time rate. From this figure, we can easily see that AHC performance for the particular set of data sources 2, 3, 7, 8, 9, 14, and 15 was significantly improved by the proposed cluster modeling method, which is 8.86% (absolute) on average. Such improvement for this particular data set can be explained like this: the cluster representation capability enhanced by the proposed cluster modeling method increased discernibility between clusters as well, and thus prevented incorrect merging processes between heterogeneous clusters during AHC better than the single Gaussian cluster modeling method did. The reason why performance improvement by the proposed cluster modeling method was not that significant for the other data sources seems to be because clustering errors occurring during AHC for the data sources were caused mainly by other (unknown) factors than cluster modeling.

As a consequence, we can tell that the proposed cluster modeling method mitigates the severe fluctuation of AHC performance across data sources in the single Gaussian cluster modeling method case. This means that the reliability of AHC performance can be improved by the proposed cluster modeling method. Conversely, we can conclude that the choice of cluster modeling methods for inter-cluster distance measurement is one of the factors to decide the reliability level of speaker clustering performance.

5. Conclusions

In this paper, we proposed the incremental Gaussian mixture cluster modeling method for the BIC-based inter-cluster distance measure within the framework of AHC. The proposed method tackled the problems of the two conventional cluster modeling methods, i.e., single Gaussian cluster modeling and GMM cluster modeling, by smoothly updating cluster models from normal distributions to GMMs during AHC. The accuracy of the BIC-based inter-cluster distance measure using this method was demonstrated to be more reliable than that using the single Gaussian cluster modeling method. As a result, we obtained 4.47% (absolute) performance improvement on average across 15 sets of speech segments from meeting conversations, in terms of speaker error time rate.

One of future efforts would be to implement an end-to-end speaker diarization system so as to compare our idea with others through the RT evaluation. Such a system also could enable us to see how the proposed cluster modeling method works in the cases that there exist some clusters including data from more than one speaker source. In this work, we just handled speech segments each of which contains homogenous data in terms of speaker characteristics.

Another interesting research direction would be to find out more factors that degrade the reliability of AHC performance across data sources. In this work, we found out that the choice of cluster modeling methods for inter-cluster distance measurement was one of such factors. Previously, we also claimed in [7, 8] that the reliability of AHC performance would be affected by the choice of stopping point estimation methods and the portion of small size initial clusters (i.e., short speech segments). Possible factors might be the statistics of given speech segments in terms of speaker characteristics, such as speaking time distribution over speakers or speaking style differences between speakers.

6. References

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

[2] 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 Rich Transcription Workshop (RT-04)*, Nov. 2004.

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

[4] Ajmera, J. and Wooters, C., "A robust speaker clustering algorithm," *Proc. ASRU 2003*, pp. 411-416, Nov. 2003.

[5] Anguera, X., Wooters, C., Peskin, B., and Aguilo, M., "Robust speaker diarization for meetings: ICSI RT06S meetings evaluation system," *Proc. MLMI 2006*, pp. 346-358, May 2006.

[6] Gish, H., Siu, M., and Rohlicek, R., "Segregation of speakers for speech recognition and speaker identification," *Proc. ICASSP 1991*, vol. 2, pp. 873-876, May 1991.

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

[8] Han, K. J., Kim, S., and Narayanan, S. S., "Robust speaker clustering strategies to data source variation for improved speaker diarization," *Proc. ASRU 2007*, pp. 262-267, Dec. 2007.