An Improved Cluster Model Selection Method for Agglomerative Hierarchical Speaker Clustering using Incremental Gaussian Mixture Models

Kyu J. Han, Shrikanth S. Narayanan

Signal 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

In this paper, we improve our previous cluster model selection method for agglomerative hierarchical speaker clustering (AHSC) based on incremental Gaussian mixture models (iGMMs). In the previous work, we measured the likelihood of all the data points in a given cluster for each mixture component of the GMM modeling the cluster. Then, we selected the N-best component Gaussians with the highest likelihoods to make the GMM refined for the purpose of better cluster representation. N was chosen empirically then, but it is difficult to set an optimal N universally in general. In this work, we propose an improved method to adaptively select component Gaussians from the GMM considered, by measuring the *degree of representativeness* of each Gaussian component, which we define in this paper. Experiments on two data sets including 17 meeting speech excerpts verify that the proposed approach improves the overall clustering performance by approximately 20% and 10% (relative), respectively, compared to the previous method.

Index Terms: agglomerative hierarchical speaker clustering, incremental Gaussian mixture model, cluster model selection, degree of representativeness

1. Introduction

Accurate distance measurement between clusters is critical for speaker clustering. In the framework of agglomerative hierarchical speaker clustering[1] (AHSC), which has been broadly utilized in the field of speaker clustering and diarization [2], generalized likelihood ratio (GLR) or Bayesian information criterion (BIC) based cluster distance metrics are used [3, 4]. Proper statistical cluster modeling is important for these metrics to measure cluster distance reliably. One of the desirable features required for cluster modeling in AHSC is to adaptively increase the complexity of cluster model distributions depending upon the size of the corresponding clusters, since clusters get bigger due to merging throughout AHSC. For this, we recently proposed incremental Gaussian mixture models (iGMMs) [5], which were successful in terms of AHSC performance by incrementing the mixture components of any GMM for a newly merged cluster.

More recently, we introduced mixture selection methods for iGMM-AHSC to refine cluster model distributions for more accurate distance measurement between clusters and obtained improvement in both clustering and diarization performance [6, 7]. Among the methods proposed in [6], it was the best to measure the likelihood of all the data points[2] in a given cluster for each mixture component of the GMM modeling the cluster and select the 16-best components with the highest likelihoods. However, 16 was chosen merely out of the limited number of candidates (4, 8, and 16), and it is hard to verify that the 16-mixture GMM is optimal over every cluster considered during AHSC from a cluster model refinement perspective. The optimal number of mixture components might be different for each cluster depending upon the data statistics of the cluster. It would thus be more reasonable to adaptively select mixture components from a given GMM, considering the statistical variation of the corresponding cluster throughout AHSC.

In this paper, we propose a novel method to select mixture components, which are cluster-dependent, when we refine GMMs in iGMM-AHSC. For this, we define the *degree of representativeness*. This measure estimates how representative each mixture component in a given GMM is in terms of modeling all the data points in the corresponding cluster. By choosing the mixture components showing a *positive* degree of representativeness, we can adaptively select mixture Gaussians depending upon the data statistics of the cluster modeled by the underlying GMM. Through this method we can achieve performance improvement in clustering accuracy (compared to our previous work in [6, 7]) of approximately 20% and 10% (relative) over two different data source sets, respectively.

This paper is organized as follows. In Section 2, we review our previous work on iGMM-AHSC and mixture selection for iGMM cluster modeling. In Section 3, we propose how to measure the degree of representativeness of each mixture component in a given GMM and select the component suitable to represent the GMM better. In Section 4, we explain our data and simulation setup followed by discussion on experimental results. Finally, in Section 5, we conclude this paper with the final remarks and future research directions.

2. Our previous work

2.1. iGMM-based AHSC

Before describing iGMM-AHSC, we need to first see how to measure inter-cluster distance during clustering. For this, we utilize GLR [3] as a cluster distance metric, which assumes two hypotheses for a given pair of clusters for likelihood compu-

This research was supported with funds from the U.S. Army Research Laboratory.

[1] It considers each input speech segment as an initial cluster and merges the closest cluster pair at every stage of clustering until a stopping criterion is satisfied [1].

[2] A single data point mentioned in this paper corresponds to a mel-frequency cepstral coefficient (MFCC).

Copyright © 2010 ISCA

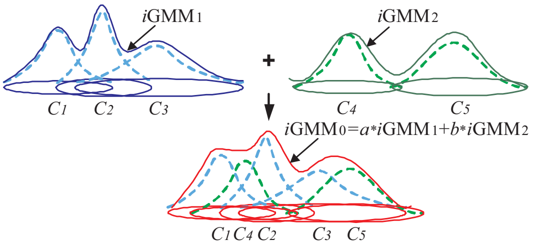


Figure 1: *iGMM cluster modeling.* $\{C_i\}_{i=1}^{5}$ *are initial clusters for AHSC, and a and b* ($a+b=1$) *are weights for the respective constituent GMMs (iGMM*$_1$ *and iGMM*$_2$*).*

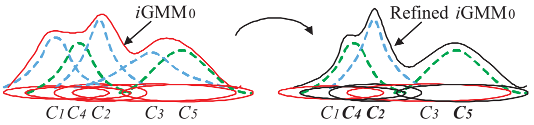


Figure 2: *Selection of representative mixture components for a given GMM in iGMM-AHSC. In this case, mixture Gaussians for* $C_2, C_4,$ *and* C_5 *are presumably selected to model* $\{C_i\}_{i=1}^{5}$.

tation: they are hypothesized to be left unmerged ($\mathcal{H}_1$) versus they are hypothesized to be merged ($\mathcal{H}_2$). $\mathcal{H}_1$ postulates different probability density functions (PDFs) for the pair of clusters, respectively, while $\mathcal{H}_2$ demands only one PDF. Then, the likelihoods of all the data points in the cluster pair are computed and subtracted on a logarithmic scale, resulting in a log-likelihood ratio based distance between clusters. The method to model these hypotheses statistically in this GLR distance metric is unique in the iGMM framework, which is described as follows:

- For $\mathcal{H}_1$, every (initial) cluster in the beginning of AHSC is modeled by a Gaussian PDF with a sample mean vector and a full covariance matrix. For a newly merged cluster, it is represented by the weighted sum of PDFs for the clusters being merged. (The weights are determined by the ratio of the number of data points in the individual cluster to the total number of data points in the merged cluster.)
- For $\mathcal{H}_2$, the pair of clusters is represented in a similar fashion to model the newly merged cluster during AHSC, i.e., by the weighted sum of PDFs for the respective clusters in the pair.

In this way, cluster models not only have smooth transitions from single Gaussian distributions to GMMs but also obtain a gradual increase in the number of Gaussian mixtures in GMMs. This is illustrated in Figure 1, where we can see how GMMs grow through merging in iGMM-AHSC.

2.2. Mixture selection for cluster modeling in iGMM-AHSC

Since AHSC has a hierarchical structure, correct merging (or correct selection of the closest cluster pair) at every stage of clustering is important. Otherwise, error propagation would occur throughout subsequent clustering stages. To reliably choose the closest pair of clusters at a given stage of AHSC, accurate GLR measurement between every pair of clusters is necessary. For this, we recently proposed a mixture selection method for cluster modeling in iGMM-AHSC where GMMs are refined in terms of representing the corresponding clusters [6]. This concept of mixture selection is illustrated in Figure 2.

Let us consider a newly merged cluster $\mathbf{X}$ at a certain stage of iGMM-AHSC. Suppose that it has gone through a few merging steps and contains n initial clusters, $\{\mathbf{x}_i\}_{i=1}^{n}$, i.e., $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$. Then, we can describe a cluster model for $\mathbf{X}$ as

$$i\text{GMM}\{\mathbf{X}\}=i\text{GMM}\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}=\lambda(m_{\mathbf{x}}^{i},\Sigma_{\mathbf{x}}^{i},w_{\mathbf{x}}^{i})_{i=1}^{n},$$

where $\lambda(\cdot)$ is a GMM, $m_{\mathbf{x}}^{i}$ and $\Sigma_{\mathbf{x}}^{i}$ are the sample mean vector and the full covariance matrix estimated from $\mathbf{x}_i$, respectively, and $w_{\mathbf{x}}^{i}$ is a weight for the Gaussian mixture component representing $\mathbf{x}_i$ in this GMM. To refine this cluster model, the proposed method computes the likelihood of all the data points in the cluster for each mixture component in the underlying GMM, i.e.,

$$\left\{p\left(\mathbf{X};m_{\mathbf{x}}^{i},\Sigma_{\mathbf{x}}^{i}\right)\right\}_{i=1}^{n},$$

and selects the N-best component Gaussians having the highest likelihoods. It is reported in [6] that this method of selecting representative mixture components improved clustering performance in iGMM-AHSC, and 16 was the best choice for N among the three candidates of 4, 8, and 16 considered, in terms of clustering accuracy on average.

Although successful overall, this approach is limited in that the empirical choice of N might not work universally over every cluster encountered during AHSC and may harm clustering performance in cases where more or fewer mixture components in iGMMs could better represent the corresponding clusters. It would thus be desirable to refine cluster models adaptively by choosing mixture components in a cluster-dependent fashion throughout AHSC, which motivated us to further develop our idea of mixture selection in this paper.

3. Degree of representativeness

In this section, we propose a novel method for selecting representative mixture components from GMMs adaptively in the framework of iGMM-AHSC. For this, we first define the *degree of representativeness* of each mixture component in a given GMM, which can be used as a measure of how globally each mixture Gaussian represents the entire cluster data. The concept of the degree of representativeness is inspired by the following facts:

- In the framework of iGMM-AHSC, each Gaussian component in GMMs is originally generated to model the corresponding initial cluster, and its parameters are estimated in a maximum likelihood fashion based on data points in the initial cluster.
- For this reason, considering a cluster $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\}$ at some stage of AHSC having gone through several merging steps, where $\{\mathbf{x}_i\}_{i=1}^{n}$ are constituent initial clusters in $\mathbf{X}$, note that

  $$\frac{1}{|\mathbf{x}_i|}p\left(\mathbf{x}_i;m_{\mathbf{x}}^{i},\Sigma_{\mathbf{x}}^{i}\right)>\frac{1}{|\mathbf{X}|}p\left(\mathbf{X};m_{\mathbf{x}}^{i},\Sigma_{\mathbf{x}}^{i}\right),\forall i,\qquad(1)$$

  where $|\cdot|$ means the number of data points in a cluster, and $m_{\mathbf{x}}^{i}$ and $\Sigma_{\mathbf{x}}^{i}$ are the sample mean vector and the full covariance matrix estimated from $\mathbf{x}_i$ through maximum likelihood estimation, respectively.
- The difference between the two (i.e., left and right) normalized likelihood terms in (1) varies over i depending upon how similar $\mathbf{x}_i$ is with $\mathbf{X}$.

Table 1: *Clustering performance evaluation on DS_1 in terms of average speaker error time rate (SPKR, %).*

	SPKR
Original iGMM-AHSC [5]	12.67
+ Mixture Selection [6]	10.14
+ Improved Mixture Selection	**7.96**

Table 2: *Clustering performance evaluation on DS_2 in terms of average speaker error time rate (SPKR, %). This evaluation is done in a diarization framework.*

	SPKR
Original SAIL Speaker Diarization System [8]	34.70
+ Mixture Selection [6]	27.75
+ Improved Mixture Selection	**24.68**

All these facts indicate that the more representative a mixture component, the lesser the difference would be. Therefore, we can measure the degree of representativeness for each mixture component in a given GMM using this difference.

Now we define the degree of representativeness (DR) for the i^{th} mixture component in a given GMM $\lambda(m_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i, w_{\mathbf{x}}^i)_{i=1}^n$, which models $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, by subtracting the two normalized likelihood terms in (1) on a logarithmic scale, as follows:

$$\begin{aligned}\mathrm{DR}(\mathcal{N}_i) &= \mathrm{DR}\left(m_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i\right)\\ &= \frac{1}{\mathrm{DR}_\sigma}\left\{\log \frac{\frac{1}{|\mathbf{x}_i|} p\left(\mathbf{x}_i; m_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i\right)}{\frac{1}{|\mathbf{X}|} p\left(\mathbf{X}; m_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i\right)} - \overline{\mathrm{DR}}\right\}, (2)\end{aligned}$$

where $\mathcal{N}_i$ is the i^{th} mixture component in λ, and $\overline{\mathrm{DR}}$ and DR_σ are the mean and the standard deviation of the degrees of representativeness over the whole n mixture components in λ.

The reason why we normalize each degree of representativeness in the definition (2) is to make it adaptive to decide how many mixture Gaussians should be selected from λ. Based on this normalization, we merely select the mixture components having the *positive degrees of representativeness*, which makes sense in that the other mixture components having the negative degrees of representativeness could be considered to have a negative effect on cluster representation, and thus be unnecessary in terms of representing $\mathbf{X}$. Since the distribution of the degrees of representativeness is determined by cluster data statistics, the number of selected mixture components for each cluster model varies along with the corresponding cluster. It would be small in a case where a few mixture components represent a cluster dominantly, while it would be large when such dominance spreads over all mixture components quite evenly.

4. Experimental results and discussion

In this section we discuss experimental results for our proposed mixture selection method based on the degree of representativeness. Before we proceed, let us first describe experimental data and simulation setup in detail.

4.1. Data sources and experimental setup

We group the data sources used for our experiments in this paper into two sets. One is 10 sets of speech segments that are excerpted from meeting speech conversations and manually segmented so that each segment has homogenous data points (or MFCCs) in terms of speaker identity. The other includes 7 unsegmented meeting speech recordings (approximately 70 minutes total, 10 minutes each). Let us call the former data set dataset 1 (DS_1) and the latter dataset 2 (DS_2). DS_1 is for pure clustering performance evaluation and DS_2 is for evaluation of clustering performance in a speaker diarization system. The data sources come from various meeting speech corpora released by ICSI, NIST, and ISL through LDC as well as our own meeting speech recordings. They are distinct from one another in terms of speaker-specific statistics, e.g., the number of speakers, gender distribution over speakers, the total speaking time, the number of speech segments, and so on. In order to avoid any potential confusion in performance analysis that might result from overlaps between segments, we do not consider any segments involved in speech overlap during either data preparation or performance evaluation.

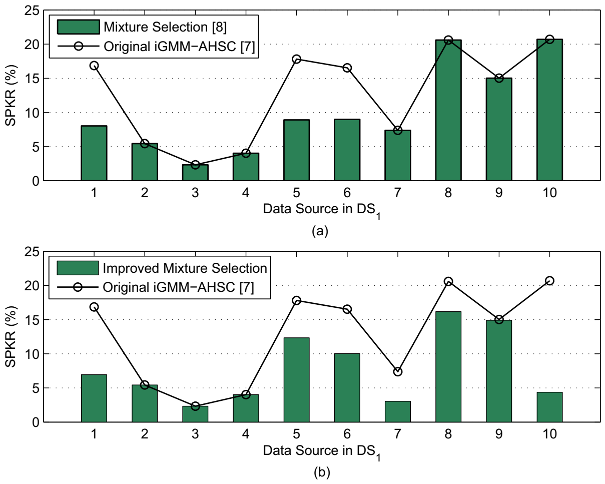


Figure 3: *Data-by-data clustering performance comparison on DS_1. (a) Original iGMM-AHSC vs. iGMM-AHSC with the mixture selection method introduced in [6]. (b) Original iGMM-AHSC vs. iGMM-AHSC with the mixture selection approach proposed in this paper.*

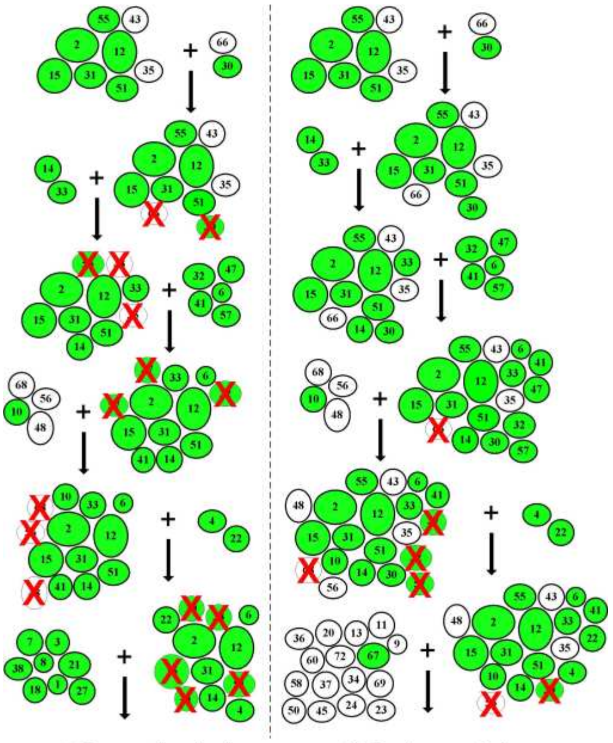
Clustering performance on DS_1 and DS_2 is evaluated in terms of average speaker error time rate, for which we use the scoring tool provided by NIST, i.e., md-eval-v21.pl [http://www.nist.gov/speech/tests/rt/2006-spring]. For the DS_2 evaluation, we utilize our own speaker diarization system introduced in [8].

MFCCs are used as acoustic features. Through 23 mel-scaled filter banks, a 12-dimensional MFCC vector is generated for every 20ms-long frame of speech. Every frame is shifted by 10ms so that there is an overlap between two adjacent frames.

4.2. Results and discussion

Tables 1 and 2[3] present performance comparisons of original iGMM-AHSC with and without mixture selection for cluster modeling in terms of average speaker error time rate on DS_1 and DS_2, respectively. As for the mixture selection methods, we compare our newly proposed approach to the previous work. In both of these tables, we can see that the proposed method, which

[3] The results in Table 2 are obtained in a diarization framework. Input speech segments for iGMM-AHSC result from voice activity detection and speaker change detection, which happens prior to AHSC in speaker diarization in general. Thus, they might have some portions either from different speaker sources or from overlap speech from spontaneous meeting conversations. This causes relatively low clustering performance in Table 2, compared to Table 1.



(a) Proposed method (b) Previous method

Figure 4: *Illustration of a part of the merging steps during iGMM-AHSC related to a randomly-selected cluster from the data source 10 in DS$_1$. Each ellipse with the unique identification number is an individual mixture component in the GMMs representing the corresponding clusters. The 'X' mark means that those clusters are discarded by the mixture selection method.*

not only selects representative mixture components based on the degree of representativeness but also decides the number of chosen mixture components adaptively, significantly boosts clustering performance by 37.14% and 28.88% (relative) for DS$_1$ and DS$_2$, respectively, compared to iGMM-AHSC without any mixture selection method applied for cluster modeling. Even compared to the method we introduced in [6], which selects the 16-best mixture components having the highest likelihoods of all the data points in the cluster modeled by the underlying GMM, we can note that our new approach works better on average over both data sets DS$_1$ and DS$_2$ by approximately 20% and 10% (relative), respectively.

The proposed approach is also superior to the previous method in terms of generalizing better across data sources, as shown in Figure 3. It has a comparatively lower error rate on 6 data sources (1, 5, 6, 7, 8, and 10) in DS$_1$ than iGMM-AHSC without mixture selection for cluster modeling, while its counterpart is helpful only for 3 data sources (1, 5, and 6). This suggests that the proposed approach is more adaptive to the variation of data sources, specifically to the variation of data statistics in clusters during iGMM-AHSC on each data source, which the previous method fails to achieve. Figure 4 illustrates how the proposed approach works more adaptively. In Figure 4(a), we observe that the proposed approach keeps discarding undesirable mixture components (white colored, representing a cluster from a different speaker source) after each merging step by selecting only a moderate number (from 7 to 10) of representative mixture components. Thus, cluster models can be further refined at every stage of AHSC and then correct merging continues to occur. On the other hand, we see from Figure 4(b) that the previous method does not discard undesirable mixture components efficiently. This is not only because the metric used in [6] to decide the 16-best mixture components is not as reliable as the degree of representativeness defined in this paper, but also because it inherently needs to keep the pre-set number (16 in this case) of mixture components throughout AHSC regardless of variation in the data statistics of a given cluster, which should also be adaptively adjusted like our new approach does. In sum, we can conclude that the approach proposed in this paper suitably develops upon our previous idea of selecting mixture components in a more adaptive way.

5. Conclusions

In this paper we developed our previous mixture selection method for cluster modeling in iGMM-AHSC into a more adaptive one that can reliably select representative mixture components throughout the clustering procedure. By this method, we were able to further refine each cluster model, and demonstrated the overall clustering performance improvement across various meeting speech data.

As one potential research direction, we could think of other metrics to measure the degrees of representativeness for mixture components in the underlying GMM than what we defined in this paper. For example, we might consider the coverage of each mixture component across all the data points in the corresponding cluster as the degree of representativeness. In this case, we could apply various matrix norms, e.g., Frobenius norm, to the covariance matrix of each mixture component in order to capture how broadly the mixture component stretches out over the entire cluster data. It would be worthwhile to compare this to what we proposed in this paper, which might lead us into deeper understanding of the relation between mixture components in the underlying GMM and the corresponding cluster in the framework of iGMM-AHSC.

6. References

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

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

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

[4] 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.

[5] Han, K. J. and Narayanan, S. S., "Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling," *Proc. Interspeech 2008*, pp. 20-23, Sept. 2008.

[6] Han, K. J. and Narayanan, S. S., "Signature cluster model selection for incremental Gaussian mixture cluster modeling in agglomerative hierarchical speaker clustering," *Proc. Interspeech 2009*, pp. 2547-2550, Sept. 2009

[7] Han, K. J. and Narayanan, S. S., "Improved speaker diarization of meeting speech with recurrent selection of representative speech segments and participant interaction pattern modeling," *Proc. Interspeech 2009*, pp. 1067-1070, Sept. 2009.

[8] Han, K. J., Georgiou, P. G., and Narayanan, S. S., "The SAIL speaker diarization system for analysis of spontaneous meetings," *Proc. MMSP 2008*, pp. 966-971, Oct. 2008.