INTERSPEECH 2009 BRIGHTON

Signature Cluster Model Selection for Incremental Gaussian Mixture Cluster Modeling in Agglomerative Hierarchical Speaker Clustering

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

Agglomerative hierarchical speaker clustering (AHSC) has been widely used for classifying speech data by speaker characteristics. Its bottom-up, one-way structure of merging the closest cluster pair at every recursion step, however, makes it difficult to recover from incorrect merging. Hence, making AHSC robust to incorrect merging is an important issue. In this paper we address this problem in the framework of AHSC based on incremental Gaussian mixture models, which we previously introduced for better representing variable cluster size. Specifically, to minimize contamination in cluster models by heterogeneous data, we select and keep updating a representative (or signature) model for each cluster during AHSC. Experiments on meeting speech excerpts (4 hours total) verify that the proposed approach improves average speaker clustering performance by approximately 20% (relative).

Index Terms: agglomerative hierarchical speaker clustering, incremental Gaussian mixture model, signature cluster model selection

1. Introduction

In speaker clustering, which refers to the automatic process of classifying speech data by speaker characteristics (or speaker identity) generally in an unsupervised manner, a bottom-up or agglomerative hierarchical strategy has been widely used due to its simple processing structure and acceptable level of performance. We call this approach *agglomerative hierarchical speaker clustering* (AHSC), which works as follows: it initially considers input speech segments as individual clusters and recursively merges the closest pair of clusters in terms of speaker characteristics. Its recursive merging process continues until it is decided that additional cluster merging would not improve speaker clustering performance any further.

Despite its aforementioned merits, AHSC has a major, inherent drawback: its one-way, recursive merging structure. During AHSC, clusters could not have any chance to be purified (in terms of speaker characteristics) once they were mixed with heterogeneous clusters due to incorrect merging. This drawback causes AHSC to be vulnerable to incorrect merging. Furthermore, incorrectly merged clusters are highly likely to cause other incorrect merging subsequently because they are already contaminated by heterogeneous data and have more chances to provide incorrect speaker-specific statistics for inter-cluster distance measurement[1], which depends upon the statistical information of data in clusters. As a consequence, the negative effect of incorrect merging is propagated through AHSC. Therefore, it becomes necessary to tackle this problem in AHSC.

For this, it is natural to consider two directions. One is to design reliable cluster distance measure while the other is to make AHSC robust to incorrect merging. There has been significant work on developing inter-cluster distance measurement, and a few good distance measures and their variations are now widely adopted in the field of speaker clustering, such as generalized likelihood ratio (GLR) [1], cross likelihood ratio (CLR) [2],[3], symmetric Kullback-Leibler distance (KL2) [4], and Bayesian information criterion (BIC) [5]-[7]. However, in terms of improving AHSC's robustness to incorrect merging, there has been comparatively little effort except for some cluster purifying algorithms recently introduced by Anguera, *et al.* [8]. This paper focuses on the latter perspective.

In this paper, we consider AHSC based on *incremental Gaussian mixture models* (*i*GMMs), which we previously introduced in [9],[10] for better statistical cluster modeling (for inter-cluster distance measurement) compared to conventional approaches utilizing normal distributions [1],[4]-[7],[11] or GMMs with fixed numbers of Gaussian mixtures [2],[3],[8],[12]-[16]. In this framework of speaker clustering, we propose a novel idea for making AHSC robust to incorrect merging, which is to select and keep updating a representative (or *signature*) GMM for each cluster through updating mixtures in the respective *i*GMM. Details will be presented in Section 3. We will show that this approach can reduce the negative effect of incorrect merging during *i*GMM-based AHSC by preventing cluster models from being contaminated by heterogeneous data (in terms of speaker characteristics) in clusters.

This paper is organized as follows. In Section 2, *i*GMM-based AHSC is briefly described based on [9] for providing the necessary background for the rest of the work reported here. Then, in Section 3, we propose the idea of selecting signature cluster models and introduce two specific methods for signature cluster model selection/update. In Section 4, we explain our data and simulation setup, and we discuss experimental results. Finally, in Section 5, concluding remarks and future research directions are presented.

2. *i*GMM-based AHSC

Accurate inter-cluster distance measurement is important in AHSC to choose the clusters being merged properly at every recursion step. Since most of state-of-the-art distance measures, including the aforementioned ones like GLR, CLR, KL2, and BIC, are statistical methods that rely on cluster properties, ro-

[1]This is performed at every recursion step of AHSC in order to choose the closest pair of clusters.

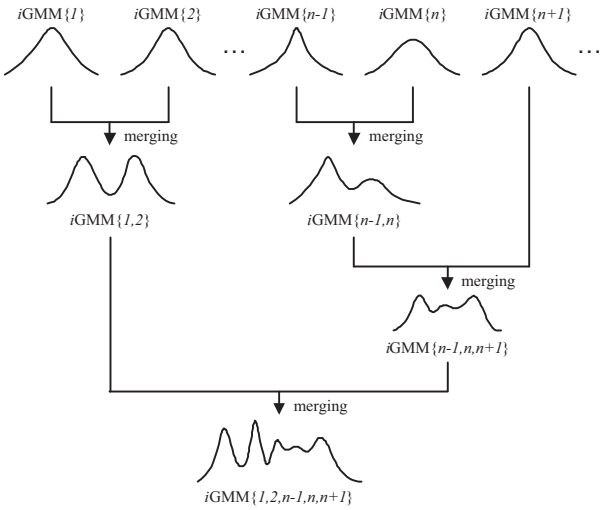


Figure 1: *Pictorial dendrogram for iGMM-based AHSC.*

bust cluster modeling is critical for enabling more accurate cluster comparisons. In [9] we proposed the iGMM-based cluster modeling approach for AHSC with the reasoning that ideal cluster models should consider variability in cluster size as well. In this regard, conventional cluster modeling approaches, both using normal distributions and GMMs with fixed numbers of Gaussian mixtures, lack the flexibility to consider clusters of variable size as the average cluster size increases due to merging during AHSC. The former does not represent large-sized clusters well while the latter has issues in representing smaller clusters. The iGMM-based cluster modeling method attempts to find a simple middle ground between these two approaches, and was empirically verified to provide better speaker clustering performance [9].

In iGMM-based AHSC, cluster modeling is performed as follows:

- Every (initial) cluster in the beginning of AHSC is represented by a normal probability distribution function (pdf) with a sample mean vector and (full) covariance matrix.
- After merging during AHSC, a newly merged cluster is represented by the weighted sum of the pdfs for the clusters being merged.
- The weights are determined by the normalized cardinalities of the merged clusters.

In this way, the pdfs of cluster models not only have smooth transitions from normal pdfs to the pdfs of GMMs but also obtain a gradual increase in the number of Gaussian mixtures in the pdfs of GMMs. This is illustrated in Figure 1 where we can see how the pdfs of iGMMs for initial clusters (top level in the figure) grow through merging in AHSC. Computational complexity for this cluster modeling approach is quite low because there are no training sessions in iGMMs like the expectation-maximization (EM) procedures used for conventional GMMs.

3. Signature Cluster Model Selection

Within this framework of iGMM-based AHSC, iGMMs for initial clusters act as individual Gaussian components in merged clusters at later recursion steps. Incorrect merging during AHSC, therefore, would allow heterogeneous Gaussian mixtures (in terms of speaker characteristics in clusters being represented) to be mixed together, which would cause the pdfs of iGMMs for incorrectly merged clusters to have deficiency in representing the main data characteristics of such clusters. In turn, this would result in inter-cluster distances that would lead to incorrect merging during subsequent AHSC steps. The sequential propagation of errors could severely degrade the overall speaker clustering performance.

To make iGMM-based AHSC robust to incorrect merging, we propose a novel approach, called signature cluster model selection. The basic idea here is that if we were able to preserve representative statistical models for merged clusters, then inter-cluster distance measurement based on such models would be protected from incorrect merging potentially occurring during AHSC. In order to implement this idea in the framework of iGMM-based AHSC, we consider generating signature GMMs for merged clusters by choosing representative Gaussian components from the respective iGMMs. This can exclude unnecessary, outlier Gaussian components being positioned in iGMMs due to incorrect merging and, thus, can mitigate their negative effect on the ensuing inter-cluster distance measurement and the overall speaker clustering performance.

An important practical question now is how to choose representative Gaussian components from iGMMs. We propose two specific methods for this purpose in the next two subsections.

3.1. Global Likelihood Comparison

Our first approach to selecting representative Gaussian components from iGMMs is based on *global likelihood comparison*, which is to compute the likelihood of the entire data in a cluster for the pdf of every single Gaussian component[2] and pick the N-best components in terms of likelihood, where N is less than the total number of Gaussian mixtures in the respective iGMM. The chosen N Gaussian components form a new GMM (with N mixtures), which is a signature cluster model to newly represent the cluster from this step forward. This process is illustrated in Figure 2(a), where the red and green Gaussian components (the two leftmost ones) are selected from the given iGMM and form the signature GMM with 2 mixtures, assuming that $N = 2$ in this case and the likelihood of the entire cluster data (the grey region under the iGMM) for the blue Gaussian component (the rightmost one) is the lowest.

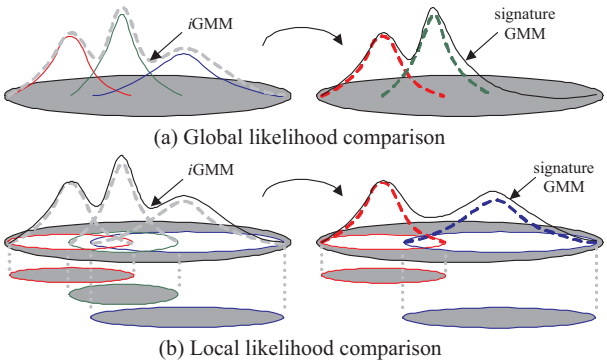
The reason why we call this approach global likelihood comparison is that the likelihoods of the *entire* data, not just a portion of them, in the cluster considered are compared to choose representative Gaussian components. Global likelihood comparison, as a consequence, provides Gaussian components that can represent the entire data in a cluster universally.

3.2. Local Likelihood Comparison

The second approach is based on *local likelihood comparison*, which is to compute the normalized likelihood[3] of data in the sub-cluster corresponding to each Gaussian component in the iGMM considered (i.e., initial cluster in the beginning of iGMM-based AHSC) for the pdf of the iGMM and pick the N-best sub-clusters in terms of likelihood. As in Section 3.1,

[2]When we compute these likelihoods, we exclude weights for the respective Gaussian components in our consideration. Otherwise, Gaussian components with larger weights in iGMMs would tend to have higher likelihood values and would presumably be selected as representative components, which is not a fair comparison.

[3]It is the likelihood divided by the cardinality of the sub-cluster considered.



(a) Global likelihood comparison

(b) Local likelihood comparison

Figure 2: *Illustration of the two proposed methods for selecting representative Gaussian components.*

N should be less than the total number of Gaussian mixtures in the iGMM. The Gaussian components corresponding to the chosen sub-clusters form a signature GMM. This process is illustrated in Figure 2(b), where the red and blue Gaussian components (the leftmost and the rightmost ones) are selected from the given iGMM and form the signature GMM, assuming that $N = 2$ and the likelihood of data in the sub-cluster corresponding to the green Gaussian component (the middle one) is the lowest.

The main difference of this approach from the previous one is that the likelihoods of a *portion* of the entire data in a cluster, i.e., data in the sub-clusters corresponding to Gaussian components in the iGMM considered, are compared to select representative Gaussian components, which is why this approach is named as local likelihood comparison. In this way, local likelihood comparison provides Gaussian components that have high level of membership to the iGMM considered.

3.3. General Summary

To summarize these two proposed methods for signature cluster model selection, let us consider a certain cluster $\mathbf{x}$ at one recursion step in the middle of iGMM-based AHSC. Suppose that it has gone through merging and contains n initial clusters, i.e., $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, where $\{\mathbf{x}_i\}_{i=1}^n$ are initial clusters. Then, $i\text{GMM}\{\mathbf{x}\} = i\text{GMM}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} = \lambda(\underline{m}_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i, w_{\mathbf{x}}^i)_{i=1}^n$, where $\lambda(\cdot)$ is a GMM, $\underline{m}_{\mathbf{x}}^i$ and $\Sigma_{\mathbf{x}}^i$ are the sample mean vector and (full) covariance matrix estimated from $\mathbf{x}_i$, respectively, and $w_{\mathbf{x}}^i$ is a weight for the normal distribution (or Gaussian component) representing $\mathbf{x}_i$ in this GMM.

For signature cluster model selection, global likelihood comparison computes and compares

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

while local likelihood comparison considers

$$\{p\left(\mathbf{x}_i|\lambda\right)\}_{i=1}^n = \left\{ \sum_{j=1}^n w_{\mathbf{x}}^j \cdot p\left(\mathbf{x}_i; \underline{m}_{\mathbf{x}}^j, \Sigma_{\mathbf{x}}^j\right) \right\}_{i=1}^n.$$

4. Experiments and Discussion

In this section we discuss experimental results for our proposed signature cluster model selection methods. Before we proceed, let us describe experimental data and simulation setup in detail.

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	6
N_s	7 (5:2)	7 (5:2)	6 (4:2)	7 (6:1)	6 (4:2)	5 (1:4)
T_s	1065	931	674	2336	1149	805
N_{ss}	418	279	176	611	244	228
	Data Source					
	7	8	9	10	11	12
N_s	6 (5:1)	8 (4:4)	5 (5:0)	9 (7:2)	4 (2:2)	4 (3:1)
T_s	1665	968	1609	659	407	443
N_{ss}	532	305	591	159	114	75
	Data Source					
	13	14	15	16	17	
N_s	6 (4:2)	8 (4:4)	4 (2:2)	3 (1:2)	4 (0:4)	
T_s	624	272	478	365	429	
N_{ss}	144	93	119	73	95	

4.1. Data Sources and Experimental Setup

In Table 1, the data sources used for our experiments are presented. These data represent 17 different sets of speech segments with approximately 4hr total durations and were chosen from ICSI, NIST, and ISL meeting corpora. They are distinct from one another in terms of speaker-specific statistics, such as 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 data source, we manually segmented the respective audio clip at every point of speaking turn changes based on the reference transcription provided. 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.

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.

As an inter-cluster distance measure for AHSC, BIC [5] was used. It was assumed that recursion stopping point estimation detects when to stop AHSC properly. Speaker clustering performance was evaluated by speaker error time rate, which has been officially used as a performance measure for speaker clustering. For this, we used the scoring tool, i.e., md-eval-v21.pl [http://www.nist.gov/speech/tests/rt/2006-spring].

4.2. Experimental Results

Table 2 presents comparison of iGMM-based AHSC with and without signature cluster model selection in terms of average speaker error time rate for our various data sources. In the experiments, we considered 3 different N values (4, 8, and 16) for N-best selection of representative Gaussian components from iGMMs in order to see how many Gaussian components would be empirically appropriate for signature cluster model selection. From this table, we can observe that regardless of N our signature cluster model selection approaches improve iGMM-based AHSC in terms of average performance. This is as expected because selection of representative Gaussian components from iGMMs during AHSC helps in minimizing the negative effect of heterogeneous Gaussian components on inter-cluster distance measurement. Between the two methods for

Table 2: *Comparison of iGMM-based AHSC with and without signature cluster model selection (SCMS) in terms of average speaker error time rate (%). N: number of the representative Gaussian components selected by SCMS, GLC: global likelihood comparison, and LLC: local likelihood comparison.*

N	AHSC with SCMS		AHSC w/o SCMS
	GLC	LLC	
4	12.43	***11.01***	12.58
8	11.01	11.08	
16	***10.27***	11.81	

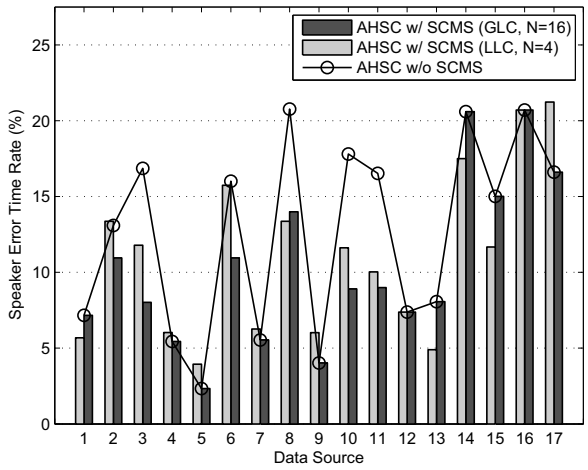


Figure 3: *Performance comparison of global likelihood comparison (GLC) and local likelihood comparison (LLC) with their best N values.*

representative Gaussian component selection, global likelihood comparison is shown to be better overall. One interesting observation is that global likelihood comparison tends to show better performance as N increases while local likelihood comparison provides the best performance at $N = 4$ without much difference. This indicates that the latter is less sensitive to N than the former.

Figure 3 compares the two methods with their best N values in a data-by-data perspective and shows that global likelihood comparison provides more stable performance enhancement for iGMM-based AHSC than local likelihood comparison; although, for some data sources, e.g., Data Sources 13, 14, and 15, the latter outperforms the former. (Note that global likelihood comparison does not degrade AHSC, but local likelihood comparison sometimes does, as shown in the results for Data Sources 9 and 17.)

5. Conclusions

In this paper, we proposed the idea of *signature cluster model selection* for making iGMM-based AHSC robust to incorrect merging. We introduced two specific methods for choosing the Gaussian components deemed to represent the data of the clusters considered. We also experimentally verified that our proposed approaches could boost the robustness of clustering performance to incorrect merging and, as a result, improve the reliability of iGMM-based AHSC across data sources. It should be noted that a number of different factors such as differences in conversation/interaction patterns between speakers or inherent speaker-specific discernibility in an MFCC space could contribute to incorrect merging scenarios during AHSC.

An important future step would be to find the optimal N value in selecting representative Gaussian components because experiments reveal N to be data-dependent. In this paper we only considered three N values (4, 8, and 16), and global likelihood comparison and local likelihood comparison fit to 16 and 4, respectively. We could potentially obtain better reliability in iGMM-based AHSC if we were able to choose a proper N value adaptively depending upon speaker-specific characteristics in data sources. This is a part of our ongoing work.

6. References

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

[2] Reynolds, D. A., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Comm.*, vol. 17(1-2), pp. 91-108, Aug. 1995.

[3] Le, V., Mella, O., and Fohr, D., "Speaker diarization using normalized cross likelihood ratio," *Proc. Interspeech 2007*, pp. 1869-1872, Aug. 2007.

[4] Siegler, M. A., Jain, U., Raj, B., and Stern, R. M., "Automatic segmentation, classification and clustering of broadcast news audio," *Proc. DARPA SR Workshop*, pp. 97-99, Feb. 1997.

[5] 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.

[6] Delacourt, P. and Wellekens, C. J., "DISTBIC: A speaker-based segmentation for audio data indexing," *Speech Comm.*, vol. 32(1-2), pp. 111-126, Sept. 2000.

[7] Zhou, B. and Hansen, J. H. L., "Efficient audio stream segmentation via the combined T^2 statistic and Bayesian information criterion," *IEEE Trans. Speech Audio Process.*, vol. 13(4), pp. 467-474, July 2005.

[8] Anguera, X., Wooters, C., and Hernando, J., "Purity algorithms for speaker diarization of meeting data," *Proc. ICASSP 2006*, pp. 1025-1028, May 2006.

[9] 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.

[10] 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.

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

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

[13] Sinha, R., Tranter, S. E., Gales, M. J. F., and Woodland, P. C., "The Cambridge university March 2005 speaker diarisation system," *Proc. Interspeech 2005*, pp. 2437-2440, Sept. 2005.

[14] Meignier, S., Moraru, D., Fredouille, C., Bonastre, J., and Besacier, L., "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech Lang.*, vol. 20(2-3), pp. 303-330, July 2006.

[15] Barras, C., Zhu, X., Meignier, S., and Gauvain, J., "Multistage speaker diarization of broadcast news," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14(5), pp. 1505-1512, Sept. 2006.

[16] Wooters, C. and Huijbregts, M., "The ICSI RT07s speaker diarization system," *Proc. NIST CLEAR/RT Workshop (LNCS)*, vol. 4625, pp. 509-519, June 2007.