INTERSPEECH 2009 BRIGHTON

Improved Speaker Diarization of Meeting Speech with Recurrent Selection of Representative Speech Segments and Participant Interaction Pattern 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

In this work we describe two distinct novel improvements to our speaker diarization system, previously proposed for analysis of meeting speech. The first approach focuses on recurrent selection of representative speech segments for speaker clustering while the other is based on participant interaction pattern modeling. The former selects speech segments with high relevance to speaker clustering, especially from a robust cluster modeling perspective, and keeps updating them throughout clustering procedures. The latter statistically models conversation patterns between meeting participants and applies it as *a priori* information when refining diarization results. Experimental results reveal that the two proposed approaches provide performance enhancement by 29.82% (relative) in terms of diarization error rate in tests on 13 meeting excerpts from various meeting speech corpora.

Index Terms: speaker diarization, representative speech segments, participant interaction pattern modeling

1. Introduction

Speaker diarization [1] of meeting speech aims to analyze spontaneous conversations in a meeting room without any prior speaker-specific knowledge and annotate them in terms of "who spoke when". The results of this analysis can be utilized in various ways. For example, using diarization results (i.e., meeting metadata) we could summarize and archive large amounts of meeting speech recordings efficiently with minimal or no human help. This would facilitate prompt and punctual retrieval when browsing specific meeting conversations in the archive. Furthermore, based on diarization results, we could utilize speech segments classified by speaker characteristics for (unsupervised) speaker adaptation. This would improve speech recognition accuracy for meeting conversations and thus could provide better understanding of the entire meeting. The Rich Transcription (RT) Evaluation [2], which has been annually offered by the National Institute of Science and Technology (NIST) since 2002, grew out of identifying such potential uses and the relevant challenges in meeting domain speech. As a result, various state-of-the-art speaker diarization systems including [3]-[6] have been introduced and developed.

We recently have developed a speaker diarization system for analysis of meeting speech [7]. The three main unique aspects of our system include: 1) speech activity detection and speaker change detection based on a *leader-follower sequential clustering* structure [8], 2) *incremental Gaussian mixture cluster modeling* [9] for inter-cluster distance measurement in agglomerative hierarchical speaker clustering (AHSC), and 3) stopping point estimation using *information change rate* (ICR) [10],[11] for AHSC, indicated in Figure 1(a). We experimentally verified in [7] that, with these unique features on a basic diarization framework [1], our system could enhance robustness to characteristic variability in meeting data and thus provide reliable diarization performance overall.

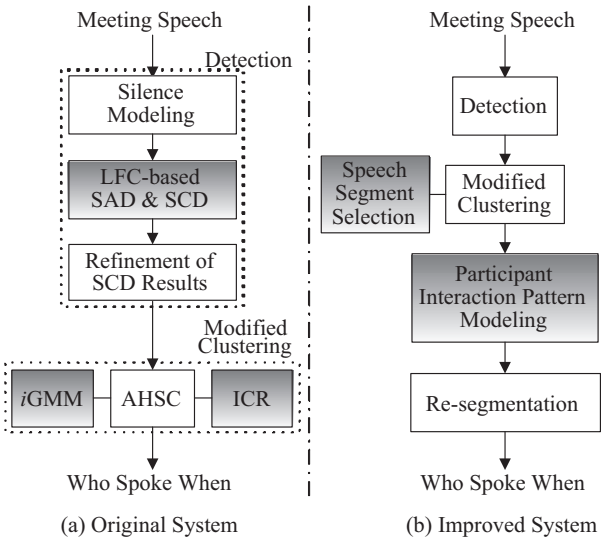

Figure 1: *Comparison of two speaker diarization systems. (a) Original system introduced in [7]. (b) Proposed, improved system in this paper. Notable features in each system are highlighted. LFC: leader-follower clustering, SAD: speech activity detection, SCD: speaker change detection, AHSC: agglomerative hierarchical speaker clustering, iGMM: incremental Gaussian mixture model, and ICR: information change rate.*

In this paper, we further develop this diarization system through two novel approaches, one of which is *recurrent selection of representative speech segments* for speaker clustering and the other is *participant interaction pattern modeling*. The former approach chooses speech segments with high relevance to speaker-specific cluster modeling for inter-cluster distance measurement in speaker clustering. Such representative speech segments are updated throughout speaker clustering procedures to minimize potential contamination in cluster models due to statistically heterogeneous data (or speech segments). This can provide more resilient cluster models for statistical distance comparison between clusters. The latter approach utilizes an m-state (1st-order) Markov chain model to represent con-

versational patterns between meeting participants, where m is determined by the number of clusters that remain after speaker clustering. The transition probabilities in this model are used as *a priori* information when we refine speaker diarization results. This can enhance the overall speaker diarization performance by considering temporal dynamics in dialogues between speakers [12]. The proposed system with these two new features is illustrated in Figure 1(b).

This paper is organized as follows. In Section 2, recurrent selection of representative speech segments for speaker clustering is described along with a brief review of iGMM-based AHSC [9]. Then, in Section 3, we propose participant interaction pattern modeling based on a Markov chain model. In Section 4, we explain data sources and simulation setup, and we discuss experimental results. Finally, concluding remarks and future research directions are presented in Section 5.

2. Representative Speech Segment Selection

In speaker diarization, AHSC performance is known to be decisive in diarization error rate (DER) [1]. Therefore, reliable AHSC performance is necessary for low DER. In general AHSC works as follows: speech segments resulting from speaker change detection (SCD) form initial clusters, and the closest cluster pair is merged recursively until a certain stopping criterion is met. To achieve reliable AHSC performance, it is crucial that inter-cluster distance is calculated accurately. This affects selection of the closest cluster pair at every recursion of AHSC. Since inter-cluster distance is measured by comparing cluster data statistics, modeling clusters reliably during AHSC is important. In the next subsection, we briefly describe our novel cluster modeling approach based on incremental Gaussian mixture models (iGMMs). We previously introduced iGMMs in [9] and showed that they better handle variability in cluster size throughout AHSC when compared to conventional cluster modeling approaches utilizing normal distributions or Gaussian mixture models (GMMs).

2.1. iGMM-based Cluster Modeling

In this cluster modeling framework, clusters are modeled 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. 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.

Figure 2 presents how the pdfs of iGMMs grow through merging in AHSC. In this figure, $i\text{GMM}_1$ and $i\text{GMM}_2$ represent two clusters $\{C_1, C_2, C_3\}$ and $\{C_4, C_5\}$, respectively. Each C_i is an initial cluster (i.e., individual input speech segment to AHSC). In the top row of the figure, the two clusters that have gone through merging between the initial clusters twice and once, respectively, are illustrated. Now suppose that these two clusters are merged and a newly merged cluster $\{C_1, C_2, C_3, C_4, C_5\}$ is represented by $i\text{GMM}_0$, depicted in the bottom part of the figure. The pdf of $i\text{GMM}_0$ is formed by the weighted sum of the pdfs of $i\text{GMM}_1$ and $i\text{GMM}_2$.

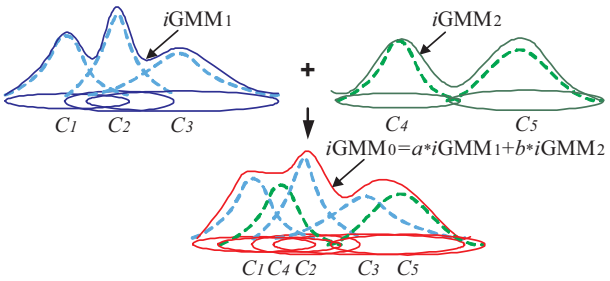

Figure 2: *iGMM-based cluster modeling. $\{C_i\}_{i=1}^5$ are initial clusters for AHSC, and a and b ($a+b=1$) are weights for the respective constituent GMMs. The weights are determined by the cardinalities of $\{C_1, C_2, C_3\}$ and $\{C_4, C_5\}$, respectively. This figure illustrates how iGMMs grow through merging.*

In summary, in this iGMM-based cluster modeling framework, every initial cluster is modeled by the pdf of a single Gaussian distribution, and once any initial cluster is merged into a larger cluster during AHSC then the pdf of its cluster model contributes to the respective iGMM by providing an individual Gaussian component.

2.2. Selection of Representative Speech Segments

In this subsection, we propose a novel idea that can improve modeling capability in this cluster modeling framework, enhancing AHSC performance (and thus DER): *representative speech segment selection*. The basic idea is that, when modeling a certain, large cluster during AHSC, selecting representative initial sub-clusters from the cluster would help because they can represent the cluster statistically better. This method would boost discernibility between clusters by avoiding potential contamination in cluster models due to incorrect merging in the past recursions of AHSC or outlier cluster data in terms of speaker characteristics.

Our way of choosing representative speech segments from a cluster is as follows. Let us consider a cluster $\mathbf{C}$. Suppose that the cluster has gone through merging and contains n initial clusters, i.e., $\mathbf{C} = \{C_1, C_2, \cdots, C_n\}$, where $\{C_i\}_{i=1}^n$ are initial clusters. Then, $i\text{GMM}\{\mathbf{C}\} = i\text{GMM}\{C_1, C_2, \cdots, C_n\} = \lambda(\underline{m}^i, \underline{\underline{\Sigma}}^i, w^i)_{i=1}^n$, where $\lambda(\cdot)$ is a GMM, $\underline{m}^i$ and $\underline{\underline{\Sigma}}^i$ are the sample mean vector and (full) covariance matrix estimated from C_i, respectively, and w^i is a weight for the Gaussian component representing C_i in this GMM.

1) Compute the likelihood of the entire data in the cluster $\mathbf{C}$ for the pdf of every single Gaussian component, i.e.,

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

   Note that we exclude weights $\{w^i\}_{i=1}^n$ in likelihood computation. Otherwise, Gaussian components with large weights in iGMMs would tend to have high likelihood values, which is not desirable for a fair comparison in the next step.

2) Select N-best components in terms of likelihood, where N is less than the total number of Gaussian mixtures in

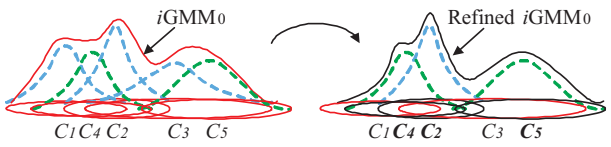

Figure 3: *Selection of representative speech segments for improved cluster modeling. In this case, C_2, C_4, and C_5 are selected as representative speech segments to model $\{C_i\}_{i=1}^{5}$.*

the respective iGMM. The initial clusters (or speech segments) corresponding to the chosen N Gaussian components are considered *representative*. The N components form a new GMM (with N mixtures), which we call a refined iGMM for the cluster **C**.

3) During AHSC, repeat 1) and 2) for every newly merged cluster whose iGMM has the number of Gaussian components greater than N. This can keep updating representative speech segments for clusters throughout AHSC, which is why we call this entire approach *recurrent* selection of representative speech segments for AHSC.

This is simply illustrated in Figure 3 where we reconsider $\{C_1, C_2, C_3, C_4, C_5\}$ and its iGMM$_0$ in Figure 2. Assuming that $N = 3$, $\{C_2, C_4, C_5\}$ are selected as representative speech segments in this case and form a new, refined GMM with 3 Gaussian mixtures.

Note that our interest in this method is to see how universally individual Gaussian components in the iGMM considered represent the entire cluster data. This is because it is reasonable to regard speech segments that correspond to the Gaussian components selected in terms of such universality as representative. This selective approach for cluster modeling using a portion of the entire cluster data can refine representation capability in cluster models in terms of not only keeping statistically representative speech segments but also excluding potentially unnecessary or even degenerate speech segments.

3. Participant Interaction Pattern Modeling

In this section, we propose another idea to draw improvement in the overall diarization performance of our system; *interaction pattern modeling between meeting participants*. This idea was motivated by the expectation that temporal dynamics between participants in meeting conversations are informative from a diarization perspective [12]. Modeling such dynamics would help in understanding the whole meeting speech and would reduce DER.

We estimate participant interaction patterns, which are meeting-dependent, based on diarization results. For this purpose, we use an m-state 1st-order Markov chain model, illustrated in Figure 4 as an example when the number of states is 4. The number of states in this interaction pattern model is set to the number of clusters that remain after AHSC. This number means the estimated number of speakers in the given meeting speech. Each transition probability is decided as follows:

1) "Who spoke when" resulting from speaker diarization is used to count the number of speaking turn transitions (N_{ij}) from the speaker S_i to the speaker S_j, where $1 \leq i, j \leq m$. Every 2s-long segment, which is the smallest unit handled in our original speaker diarization system [7], is considered for transition number counting.

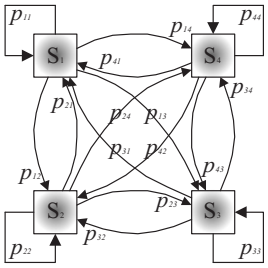

Figure 4: *1st-order Markov chain model for participant interaction patterns when the estimated number of speakers is 4, where p_{ij} is the transition probability from the speaker S_i to the speaker S_j for $1 \leq i, j \leq m$ ($m = 4$ in this case).*

2) Average N_{ij} with N_i, where $N_i = \sum_{j=1}^{m} N_{ij}$. Thus, each transition probability p_{ij} ($1 \leq i, j \leq m$) is determined by

$$p_{ij} = \frac{N_{ij}}{N_i} = \frac{N_{ij}}{\sum_{j=1}^{m} N_{ij}}.$$

The estimated transition probabilities in this model are used as *a priori* information for refinement of diarization results.

The refinement step performs a simple speaker identification task with considering m GMMs[1] for remaining clusters from AHSC as pre-trained speaker models. Specifically, it refines diarization results by classifying every 2s-long segment into one of clusters that remain after AHSC based on maximum *a posteriori*. Suppose that GMMs for clusters that remain after AHSC are λ_i ($1 \leq i \leq m$) and the entire input meeting speech **x** can be split into L 2s-long segments, i.e., $\mathbf{x} = \{x_1, x_2, \cdot, \cdot, \cdot, x_L\}$. The refinement step computes the likelihood of x_l ($1 \leq l \leq L$) for each λ_i and assigns the argument i providing the highest *a posteriori* to x_l as a speaker label, i.e.,

$$\arg\max_i p\left(x_l|\lambda_i\right) p_{ji},$$

where p_{ji} is the transition probability from the speaker S_j to the speaker S_i in the estimated interaction pattern model and it is assumed that the speaker label j is assigned to x_{l-1}.

4. Experiments and Discussion

4.1. Data Sources and Experimental Setup

In this paper, 21 meeting excerpts with an average length of 10 minutes from ICSI, NIST, ISL, and USC meeting speech corpora were used as data sources for our experiments. Eight of them were utilized for training sessions while the rest were used for testing sessions. (These data sets are the extended version of the ones used in [7].) Each data source is distinct from the others in various aspects such as the number of meeting participants, gender and speaking time distribution over participants, and so on.

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.

[1]These GMMs are trained by the EM procedures over representative speech segments in the respective clusters. The number of Gaussian mixtures is empirically set to 32.

Table 1: *Improved speaker diarization performance with the two approaches proposed in this paper, i.e., representative speech segment selection and participant interaction pattern modeling. For the former approach we empirically set $N = 32$. Performance comparison is given in terms of average DER (%) across 13 testing data sources.*

	DER
Original System [7]	32.63
+ Representative Speech Segment Selection	26.76
+ Participant Interaction Pattern Modeling	22.90

BIC [13] was used as an inter-cluster distance measure for AHSC. Diarization performance was evaluated by the scoring tool that NIST officially uses for the RT evaluations, i.e., md-eval-v21.pl [http://www.nist.gov/speech/tests/rt/2006-spring]. Overlapped speeches were excluded in performance evaluation.

4.2. Experimental Results

Table 1 presents speaker diarization performance by our original system in [7] and the modified system with the two approaches proposed in this paper, in terms of average DER across the testing data sources.

The main reason for the diarization performance increase in the modified system with recurrent selection of representative speech segments (32.63% $\rightarrow$ 26.76%, 17.99% relative improvement) is that the proposed approach helped not only in choosing the closest pair of clusters at every recursion of AHSC properly but also in estimating the optimal stopping point[2] for AHSC accurately[3]. This indicates that selecting speech segments with representativeness is better for cluster modeling than using the entire data in clusters. This is reasonable because clusters could contain unnecessary or defective data from a cluster representation perspective due to incorrect merging during AHSC and there is, therefore, a significant need to keep purifying such clusters throughout AHSC.

From the table, we can also see that the second approach contributed to DER reduction as well (14.42% relative improvement), as expected. It is especially advantageous in this high-level modeling approach that interaction patterns between participants, which are hard to be universally modeled due to their data-dependency, can be mathematically represented in an unsupervised fashion based on diarization results. Note that a very accurate stopping point estimation for AHSC is required in the proposed approach because the number of states, m, in a Markov chain model for interaction patterns is determined by the number of clusters that remain after AHSC. This is already bolstered in the modified speaker diarization system by the first approach proposed in this paper.

5. Conclusions

In this paper, we improved our original speaker diarization system introduced in [7] using two novel ideas: *recurrent selection of representative speech segments* for iGMM-based AHSC and *participant interaction pattern modeling* for refinement of diarization results. With these approaches, the modified speaker diarization system was able to obtain significant performance improvement overall by 29.82% (relative) in terms of average DER across 13 meeting excerpts from various meeting speech corpora. Based on the experimental results, we demonstrated that the proposed methods provide an appreciable step toward more reliable speaker diarization over data sources and domains.

Important future work includes finding a way of robustly dealing with overlapped speech in this framework of speaker diarization. The first approach proposed in this paper might provide a relevant hint for this potential research direction, i.e., selective clustering of data can maintain or even boost representativeness in cluster models. With this selective clustering concept, we might be able to exclude defective speech segments like the ones including overlapped speech or having more than one speaker source. This could result in more DER reduction as [14]. These are topics that are currently being investigated.

[2] This is the recursion step of AHSC when additional merging would not improve DER any further.

[3] As emphasized in [10],[11], an incorrect stopping point estimation for AHSC could result in huge DER degradation.

6. References

[1] 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.

[2] Rich Transcription Evaluation Project. National Institute of Science and Technology (NIST). [Online]. http://www.itl.nist.gov/iad/mig/tests/rt.

[3] Huang, J., Marcheret, E., Visweswariah, K., and Potamianos, G., "The IBM RT07 evaluation systems for speaker diarization on lecture meetings," *Proc. NIST CLEAR/RT Workshop (LNCS)*, vol. 4625, pp. 497-508, June 2007.

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

[5] Fredouille, C. and Evans, N., "The LIA RT'07 speaker diarization system," *Proc. NIST CLEAR/RT Workshop (LNCS)*, vol. 4625, pp. 520-532, June 2007.

[6] Zhu, X., Barras, C., Lamel, L., and Gauvain, J., "Multi-stage speaker diarization for conference and lecture meetings," *Proc. NIST CLEAR/RT Workshop (LNCS)*, vol. 4625, pp. 533-542, June 2007.

[7] 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.

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

[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. and Narayanan, S. S., "A Robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system," *Proc. Interspeech 2007*, pp. 1853-1856, Aug. 2007.

[11] Han, K. J., Kim, S., and Narayanan, S. S., "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE Trans. Speech Audio Lang. Process.*, vol. 16(8), pp. 1590-1601, Nov. 2008.

[12] Busso, C., Georgiou, P. G., and Narayanan, S. S., "Real-time monitoring of participants interaction in a meeting using audio-visual sensors," *Proc. ICASSP 2007*, pp. 685-688, April 2007.

[13] 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.

[14] Boakye, K., Vinyals, O., and Friedland, G., "Two's a crowd: Improving speaker diarization by automatically identifying and excluding overlapped speech," *Proc. Interspeech 2008*, pp. 32-35, Aug. 2008.