SPEAKER DIARIZATION IN MEETING AUDIO

Tin Lay NWE, Hanwu SUN, Haizhou LI and Susanto RAHARDJA
Institute for Infocomm Research (I²R), A*STAR, 1 Fusionopolis Way
Singapore 138632
{tlnma, hwsun, hli, rsusanto}@i2r.a-star.edu.sg

ABSTRACT

This paper describes speaker diarization system on NIST Rich Transcription 2007 (RT-07) Meeting Recognition evaluation data set for the task of Multiple Distant Microphone (MDM). Our implementation includes three components: initial clustering, non-speech removal and cluster purification. Initial clusters are generated using Direction of Arrival (DOA) information and bootstrap clustering. Multiple GMM modeling for speech/non-speech classification is employed for non-speech removal component. In addition, a novel system fusion strategy using information from Receiver Operating Curve (ROC) is proposed for non-speech removal component. Finally, consensus clustering approach together with iterative GMM clustering method is employed for speaker cluster purification. The system achieves the overall DER of 10.81%.

Index Terms— Meetings, pattern classification, clustering methods, speech processing, modeling

1. INTRODUCTION

Rapid progress in computer and network technology makes possible an increasing number of spoken documents. These include broadcast radio, television programs, meetings, voice mails and several others. With these immerse and growing body of spoken documents, efficient and effective searching, indexing and accessing on the collection of the documents become important. Speaker diarization in meeting audio is one of the tasks of great interest. Generally speaking, Speaker diarization system has three fundamental steps. The first step is to segment audio into speech and non-speech segments. The second step is to determine number of speakers in an audio and group speech segments of the same speaker together [1].

For speech/non-speech detection, high and low energy frames are first separated by a threshold. Then, speech and silence models are trained on the high and low energy frames respectively in [2]. In [3], speech/non-speech detection is performed using cepstral coefficients and time derivatives of log-energy. To be more robust to SNR, energy normalization is carried out on voiced frames.

For determining number of speakers and grouping speaker segments, many systems create an enhanced signal from the multiple microphone recordings [2, 4]. Enhanced signal is obtained by beamforming. In [2], agglomerative clustering method is used on the enhanced signal for speaker clustering. This method deduces the number of speakers in a recording, along with the information about where each speaker is speaking. In [4], enhanced audio signal is first segmented using Bayesian Information Criteria (BIC) method. Then, GMM and viterbi-decoding method is used for iterative clustering. Finally, the speaker clusters are purified by MAP adaptation.

We have developed a speaker diarization system [5] which was submitted to NIST Rich Transcription 2007 (RT-07) Meeting Recognition Evaluation. This system uses Direction of Arrival (DOA) [6] information to perform speaker turn detection and clustering. Cluster purification is then carried out by performing GMM modeling on acoustic features. Finally, non-speech and silence removal is carried out to remove unwanted segments.

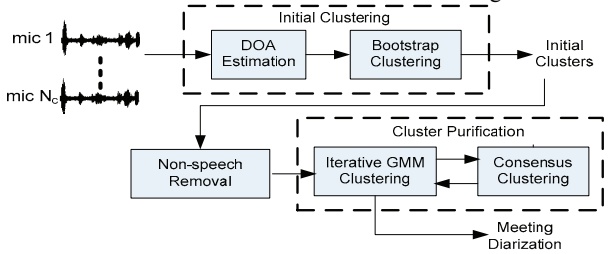

Figure 1. Block diagram of proposed speaker diarization system

In this paper, we attempt to improve the performance of the previous system [5]. Initial clustering is the same as in the previous system [5]. We modify non-speech removal and cluster purification components of previous system. For non-speech removal component, we built two systems. The first system uses energy thresholding on 'Double Layer Windowing' approach. The second system employs training multiple speech and non-speech GMM models. Then, the two systems are fused. For cluster purification, consensus clustering method [7] is integrated to the previous iterative GMM clustering process [5]. We carry out the iterative GMM clustering. Then, consensus clustering process which provides a method to represent the 'consensus' across multiple runs of iterative GMM clustering is carried out. Finally, iterative GMM clustering process is performed to produce speaker clusters. The consensus clustering approach can access the stability of the discovered clusters and is able to discard the unstable clusters. The block diagram of proposed speaker diarization system is illustrated in Figure 1.

The organization of the paper is as follows. Section 2 describes initial clustering. Section 3 presents non-speech removal system. Section 4 explains 'cluster purification' component. Section 5 presents our experimental results and Section 6 concludes the paper.

2. INITIAL CLUSTERING

2.1. Direction of Arrival (DOA) Estimation

MDM task has two or more distant microphone recordings for each meeting. The time delay between the arrivals of a sound

source to a pair of microphone is used to determine speaker turning points [5]. Figure 2 shows DOA estimation for a pair of microphone r[n] (reference microphone) and s[n] (source microphone).

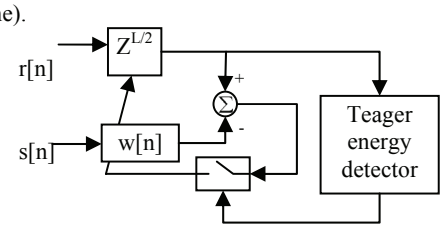

Figure 2. DOA estimation module of a microphone pair [5]

Time delay is estimated for each frame of 512 samples. For each frame, voice activity detection (VAD) is first performed by checking if the Teager energy [8] of the current frame is greater than an adaptive threshold. If the frame is speech signal, adaptive filter is allowed to adapt. Adaptive filter's weight is w[n] and it has length *L*. Normalized Least-Mean Square (NLMS) [9] algorithm is used to adapt the filter. We use $L = 250$ in our experiments.

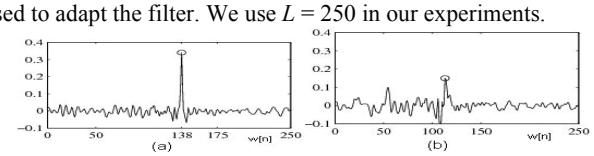

Figure 3. Plot of values for filter coefficients w[n] at frame instance *n*, (a) showing a clear peak at index 138 (b) showing multiple peaks due to reverberation effects [5].

The plots of filter coefficients w[n] for signals with good SNR and with reverberation are shown in Figure 3(a) and (b) respectively. In Figure 3(a), the filter's weights reflect an impulse with peak at index 138. This explains that source signal delays 13 samples (138-125=13) from reference signal. However, for Figure 3(b), the peak present is less pronounced and secondary peaks are present due to reverberation effects. Hence, we use the ratio of maximum peak **w** to its next highest peak as a measure of microphone pair quality.

For each RT-07 task, *K* pairs of microphones are used to generate the direction of arrival. We choose microphone pairs that have 1) High-peak to next-highest-peak ratio on **w**[n], 2) High SNR and 3) Large DOA dynamic range. The matrix *DOA*[n,k] stores the DOA values, specifically,

$$DOA[n,k] = \arg\max_{j=1\ldots L} \{w_j[n,k]\} \quad (1)$$

where $n=1\ldots N$ and $k=1\ldots K$, N being number of frames. $\mathbf{w}_j[n,k]$ is the j^{th} filter coefficient for the k^{th} microphone pair at time n.

2.2. Bootstrap Clustering

Bootstrap clustering uses location information from Eq.(1) to form initial clusters. This process has two steps. The first step is to conduct quantization for each microphone pair. The second step is to carry out quantization among microphone pairs.

As for the first step, a histogram is constructed using the k^{th} column of *DOA*[n,k] to find the frequently occurring positions. Every peak in the histogram indicates that there is a significant amount of speech originating from that particular location. Hence, we assume that each peak belongs to individual speaker. Number of peaks is estimate of the speakers present in each meeting. The peaks in the histogram are taken as centroids and *DOA*[n,k] values are quantized to these centroids using a nearest neighbor approach, as illustrated in Figures 4(a) ~ (c).

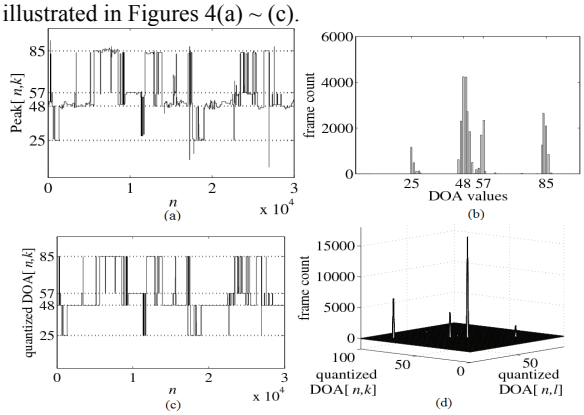

Figure 4. (a) Plot of one column of *DOA*[n,k]. Horizontal dotted lines correspond to histogram centriods. (b) Histogram of *DOA*[n,k] values for selected k^{th} microphone pair. (c) *DOA*[n,k] values after within pair quantization. (d) 2-D histogram of quantized *DOA* values for 2 microphone pairs. Four peaks can be seen [5].

As for the second step, quantization is performed along the rows of *DOA*[n,k], i.e., to identify the centroids across *K* microphone pairs. Figure 4(d) shows how 4 centroids can be observed by quantizing across 2 microphone pairs. All other histogram bins with low counts will be quantized to these centroids.

This initial clustering uses spatial location information to form clusters. Clusters can have impure segments from other speakers if speakers move or change places. Hence, we need to further purify the clusters obtained in this stage. Please refer to [5] for more details about 'Initial Clustering'. Before purifying the clusters obtained by initial clustering, we remove non-speech segments from audio.

3. NON-SPEECH REMOVAL

The task here is to identify non-speech frames (example, coughs, laughter, breathing, silence) and exclude them. We generate the enhanced recording using the BeamformIt 2.0 toolkit [10]. We build two systems. The first system uses 'Double-Layer-Windowing' and energy thresholding method. The second system employs multi-model GMM training approach. We then fuse these two systems together.

In the first system, audio is divided into frames of 20ms with 10ms overlapping. The energy for each frame is computed. In order to catch only silences that are longer than 300ms tolerance specified for the evaluation [11], second layer window is applied. The length of the second layer window is 300ms long and 10ms overlapping. Average energy is obtained for each 300ms window in second layer. When this energy is found to cross a threshold, the region covered by this window is deemed as non-speech and dropped. The threshold used to make this decision was determined on the RT-05 and RT-06 evaluation data.

As for the second system, we generate the 12 MFCC coefficients for each of 20ms frame from beamformed audio. The frame overlapping is 10ms. Then, we train a total of six GMMs for speech and non-speech categories. These include speech overlaps, clean speech and noisy speech GMMs for speech category and laughter, background noise, silence GMMs for non-speech

category. We use evaluation data of RT-05 and RT-06 to train GMMs. Each GMM has 242 mixtures. We use maximum likelihood approach to determine if the segment contains speech or non-speech.

Finally, we fuse the above two systems. Generally, high and low energy frames can be anticipated as speech and non-speech respectively. Hence, decision by energy thresholding system can be more reliable for high and low energy regions. However, energy thresholding method could confuse to make decisions, especially for frames in middle energy range. Hence, we draw ROC curves (Figure 5) using RT-05 and RT-06 data, to determine the high and low energy regions with lower false alarms by energy thresholding method. We observe that the regions outside two thick dotted lines which are at 0.6% and 33% of mean energy have False Alarm Rate (FAR) of 2% and 4% respectively for speech and non-speech for RT-05 and RT-06 data. Hence, we apply the same settings for our experiments and the classification decision is made by energy thresholding system for these regions. And, the decisions for region within two thick dotted lines are made by multimodal GMM classification system.

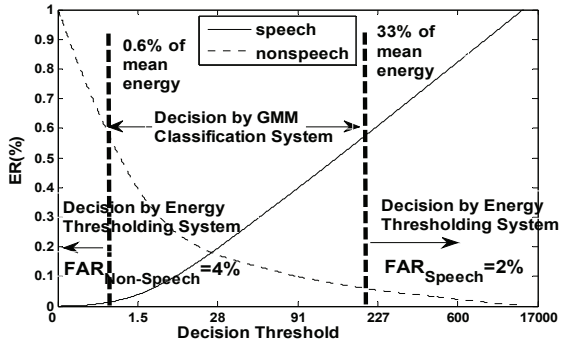

Figure 5. Receiver Operating Curve (ROC) to divide the regions of making decisions by energy thresholding system and GMM classification system

4. CLUSTER PURIFICATION

We access the reliability of the clusters obtained in initial clustering stage. Unreliable clusters are dropped. In addition, impure segments of each cluster are determined and re-assigned. We employ the consensus clustering method together with the iterative GMM modeling to do the task. Multiple runs of the clustering process are obtained by using iterative GMM clustering approach to observe the consensus.

4.1. Iterative GMM Clustering

The segments identified as speech in the non-speech removal process are used to train a root Gaussian Mixture Model (GMM), λ_{Root} using the Expectation Maximization (EM) algorithm [12]. The mixtures of λ_{Root} has full covariance matrices. Here, we use the same MFCC features extracted on enhanced signal as in non-speech removal system. Using the labeled segments of initial clustering process, individual GMMs are adapted from λ_{Root}. Adaptation is performed on the weights, means and variances using the maximum a Posteriori MAP approach [12]. Thus, if there are Q speaker clusters resulting from initial clustering, there will be Q GMMs $\lambda_{i,q}$ where i indicates the iteration number, and $q = 1...Q$ indicates the q^{th} speaker cluster. After each iteration of model adaptation, segmentation and cluster assignment is carried out. A total of 15 iterations are carried out. To observe the 'consensus' across multiple runs of GMM clustering, we repeat the iterative GMM clustering process for 55 times with mixture components, $M = 40,44,48,52,...,256$. Consensus clustering is carried out in the following section.

4.2. Consensus Clustering

Consensus matrix C [7] with the size $T \times T$, stores the proportion of clustering runs in which two speech segments are clustered together. T is number of speech segments in each meeting. The consensus matrix is obtained by taking the average over the connectivity matrices, c, of all clustering runs. There are a total of $h = 1...H$, $H = 825$ clustering runs (15 iterations X 55 times of GMM clustering process = 825) in each meeting. The entries of connectivity matrix for each clustering run h are defined as follows:

$$c^h(x,y) = \begin{cases} 1 & \text{if segments } x \text{ and } y \text{ belong to the same cluster,} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The consensus matrix, C, is average of connectivity matrices over all clustering runs.

$$C(x,y) = \frac{\sum_h c^h(x,y)}{H} \quad (3)$$

The entry (x,y) in the consensus matrix records the number of times the segments x and y are assigned to the same cluster divided by the total number of clustering runs. The entry $C(x,y) = 1$ means that the segments x and y are in the same cluster in all clustering runs and confidence of assigning these segments to the same cluster is high. Using consensus matrix C, we first select the clusters which have the members with entry 1. Then, we retain the clusters which have total duration of members longer than 5 sec and choose these clusters as the speaker clusters in each meeting. We use 5 sec as the threshold since minimum duration of the speakers in RT-05 and RT-06 data set is 5 sec. Figure 6 shows the speaker clusters selected and discarded after consensus clustering for 2 meetings. We conduct the 'iterative GMM clustering' again. Using speaker clusters chosen above, individual GMMs are adapted from λ_{Root}. Adaptation and segmentation is iterated until stabilization is found.

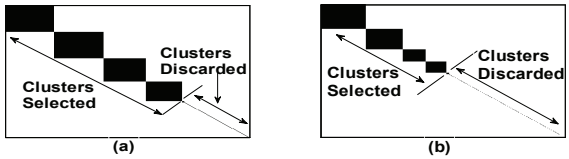

Figure 6. Speaker clusters selected and discarded after consensus clustering (a) CMU_20061115-1030 , (b) EDI_20061114-1500

5. RESULTS AND DISCUSSION

The results presented in Table 1 are from experiments conducted upon the Rich Transcription 2007 (RT-07) Meeting Recognition MDM evaluation. This evaluation consists of 8 tasks as listed in Table 1, for a total length of 3 hours. The number of microphone

recordings range from 3 (for CMU tasks) to 16 (for EDI tasks). Except EDI tasks, recordings were made using distant microphones. The EDI tasks contain recordings made by two sets of microphone arrays.

Table 1. DER (RT-07 eval) on the RT-07

Task No.*	DER (%) After			Previous System [5]
	Initial clustering	Non-speech removal	Cluster Purification	
1[4]	26.16[4]	20.30	19.45[4]	19.36[4]
2[4]	21.33[4]	10.77	10.76[4]	12.46[4]
3[4]	25.58[5]	15.66	14.76[4]	20.69[5]
4[4]	40.41[5]	14.62	9.65[4]	15.00[4]
5[4]	23.85[4]	6.99	5.75[4]	12.66[4]
6[6]	23.90[5]	13.29	9.46[5]	13.36[5]
7[5]	46.43[5]	32.82	6.35[5]	11.32[5]
8[4]	28.96[4]	11.06	10.60[4]	18.45[4]
Overall	**29.71**	**15.91**	**10.81**	**15.32**

*1= CMU_20061115-1030, 2= CMU_20061115-1530,
3= EDI_20061113-1500, 4= EDI_20061114-1500
5='NIST_20051104-1515, 6=NIST_20060216-1347
7=VT_20050408-1500, 8=VT_20050425-1000
Number of speakers for each task is indicated in [ ]

The proposed system achieves overall DER of 10.81%. This system performs better than our previous system [5] with overall DER of 15.32%. Our previous system [5] is 2nd best performer in NIST-2007 evaluation. DER of top performer is 8.51%. The performance improvement of proposed system is contributed by two components. The first is non-speech removal component in which we use 6 GMMs for speech/non-speech modeling. This explains that modeling acoustic classes differently better represents acoustic characteristics and is effective. Our previous system [5] uses only two for speech and non-speech. The second component that contributes to performance improvement is the use of 'consensus clustering' method in cluster purification process. In the previous system [5], speaker GMMs are adapted from λ_{Root}, using all the segments assigned to individual clusters in initial clustering stage. The initial clusters include many impure segments from other speakers. By introducing 'consensus clustering method', many impure segments are removed and we can have better model adaptation. We compute the impurity of each speaker cluster in terms of $FAR_{speaker\text{-}cluster.}$ We notice that $FAR_{speaker\text{-}cluster}$ of initial clustering is 34% and that of consensus clustering is 7%. We use the following equation to calculate the $FAR_{speaker\text{-}cluster}$ for each cluster.

$$FAR_{speaker-cluster} = \frac{total\ duration\ of\ impure\ speaker\ segments}{total\ duration\ of\ speaker\ segments\ in\ a\ cluster} \quad (4)$$

We found that DER of new speech/non-speech removal and old cluster purification [5] is 11.92% and that of old speech/non-speech removal and new cluster purification is 14.4%.

In addition, consensus clustering method can better estimate the number of speakers in each meeting. Number of speaker estimates for meeting numbers 3, 4 and 6 are wrong after 'initial clustering'. Speaker numbers for meetings 3 and 4 are corrected at the consensus clustering stage by removing clusters which have very few members. Figure 6(b) shows correction for meeting 4. Speaker numbers for meeting 6 can not be corrected since the number of speakers estimated by 'initial clustering' stage is less than the actual number of speakers. The proposed system is able to estimate the number of speakers better than our previous system [5].

6. CONCLUSIONS

The proposed speaker diarization system is found to yield better performance on RT-07 evaluation set in comparison with our previous system [5] that was submitted to RT-07 Meeting Recognition MDM evaluation. The use of multiple GMM modeling for speech/non-speech classes and fusion strategy employing information from ROC is effective for non-speech removal component. In addition, the use of 'consensus clustering' method help to remove the impure speaker segments and hence, better adapted initial speaker GMMs are achieved. In addition, 'consensus clustering' helps to better estimate the number of speakers in each meeting.

7. REFERENCES

[1] M. Ben, M. Bester, F. Bimbot, and G. Gravier, "Speaker Diarization Using Bottom-up Clustering Based on A Parameter-derived Distance Between Adapted GMMs," in *Proc. ICSLP*, 2004.

[2] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," In *Rich Transcription 2007 Meeting Recognition Workshop*, Baltimore, USA, May 2007.

[3] X. Zhu, C. Barras, L. lamel, and J-L. Gauvain, "Multi-stage Speaker Diarization for Conference and Lecture Meetings," In *Rich Transcription 2007 Meeting Recognition Workshop*, Baltimore, USA, May 2007.

[4] M. Konečný and D. van Leeuwen, "AMIDA RT07s Speaker Diarization System," In *Rich Transcription 2007 Meeting Recognition Workshop*, Baltimore, USA, May 2007.

[5] C-W.E. Koh, H. Sun, T.L. Nwe, T.H. Nguyen, B. Ma, C.E. Siong, H. Li, and S. Rahardja, "Using Direction of Arrival Estimate and Acoustic Feature Information in Speaker Diarization," *In Proc. Interspeech*, Antwerp, Belgium, August, 2007.

[6] M.S. Brandstein and H.F. Silverman, "A Robust Method for Speech Signal Time-delay Estimation in Reverberant Rooms," In *Proc. ICASSP*, Munich, pp. 375-378, 1997.

[7] S. Monti, P.Tamayo, J.P. Mesirov, and T.R. Golub., "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," *Machine Learning*, 52(1-2): pp. 91-118, 2003.

[8] J.F. Kaiser, "On A Simple Algorithm To Calculate The Energy Of A Signal," in Proc. *ICASSP*, Albuquerque, pp. 381–384, 1990.

[9]S. Haykin, *Adaptive Filter Theory*, (4th Edition). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2002.

[10] X. Anguera, C. Wooters, and J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Trans. On Audio, Speech and Language Processing*, vol. 15, no.7, pp. 2011-2022, 2007.

[11] "Spring 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation Plan," pp. http://nist.gov/speech/tests/rt/2007/docs/rt07-meeting-eval-plan-v2.pdf.

[12] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 1941.