INTERSPEECH 2009 BRIGHTON

Speaker Diarization for Meeting Room Audio

Hanwu Sun, Tin Lay Nwe, Bin Ma and Haizhou Li

Institute for Infocomm Research (I²R), A*STAR, 1 Fusionopolis Way, Singapore 138632

{hwsun, tlnma, mabin, hli}@i2r.a-star.edu.sg

Abstract

This paper describes a speaker diarization system in 2007 NIST Rich Transcription (RT07) Meeting Recognition Evaluation for the task of Multiple Distant Microphone (MDM) in meeting room scenarios. The system includes three major modules: data preparation, initial speaker clustering and cluster purification/merging. The data preparation consists of the raw data Wiener filtering and beamforming, Time Difference of Arrival estimate and speech activity detection. Based on the initial processed data, two-stage histogram quantization has been used to perform the initial speaker clustering. A modified purification strategy via high-order GMM clustering method is proposed. BIC criterion is applied for cluster merging. The system achieves a competitive overall DER of 8.31% for RT07 MDM speaker diarization task.

Index Terms: Multiple Distant Microphone, speaker diarization, time difference of arrive, speech activity detection, speaker clustering

1. Introduction

Speaker diarization is one of the tasks in the NIST Rich Transcription (RT) Meeting Recognition Evaluation. It is to automatically find the segments of time within a meeting in which each meeting participant is talking, a task to detect Who Spoke When [1]. This requires for marking the start and end times of every speech segment with a speaker identity, from a continuous audio recording of a meeting. In recent years, there has been extensive research on the speaker diarization systems of the Multiple Distant Microphones (MDM) conditions in meeting room scenario [2-6].

A common approach for the MDM task is first to obtain an enhanced signal from the multiple microphone recordings by using beamforming alignments [2, 3] or using weighted channel summation [4]. Then, acoustic features are extracted from the enhanced audio signal for segmentation and clustering purpose.

Although Bayesian Information Criterion (BIC) [2-5] has been widely adopted for the segmentation, but if the recordings are of poor quality, e.g. in low signal-to-noise ratio (SNR) from the distant microphone signals, BIC segmentation often results in a poor speaker change detection. For good performance, a common strategy is to over-segment a recording, and subsequently to regroup the segments by clustering. Over-segmentation however will results in too short segments to have a reliable clustering with the extracted acoustic features. Shorter segments contain less discriminative information to distinguish the speaker identities.

In this paper, we describe an improved speaker diarization system based on our previous system [6] for the MDM task of the 2007 NIST Rich Transcription (RT07) Meeting Recognition Evaluation MDM. In the system, the data preparation module consists of the raw data Wiener filtering and beamforming for enhanced speech signals, time difference of Arrival (TDOA) [9] estimation and speech activity detection (SAD). In the TDOA process, we employ an automatically way to select useful microphone pairs, and then use the TDOA information of these selected pairs to perform speaker turn detection and segmentation. The initial speaker clustering module via a two-stage TDOA histogram distribution quantization approach is applied for initial speaker clustering. Finally, the cluster purification module consists of the cluster purification and merging which is carried out by using a high-order GMM statistical modeling approach [11, 12] and the BIC Criterion. We report the experimental results of RT07 in this paper. The modules described in this paper have contributed to our submission for the 2009 speaker diarization system of the 2009 NIST Rich Transcription Meeting Recognition Evaluation.

The rest of this paper is organized as follows. In Section 2 the data preparation is described. In Section 3, the two-stage speaker initial clustering is presented. In Section 4, the cluster purification method is introduced, in Section 5 we present the experimental results and finally the conclusion is given in Section 6.

2. Data Preparation

The speaker diarization system is depicted in Figure 1. The data preparation module consists of 3 components. They are as follows in the order of system flow:

a) Wiener Filtering and Beamforming
b) Speech Activity Detection (SAD)
c) Time Difference of Arrival (TDOA) estimation

The details are provided in the following sections.

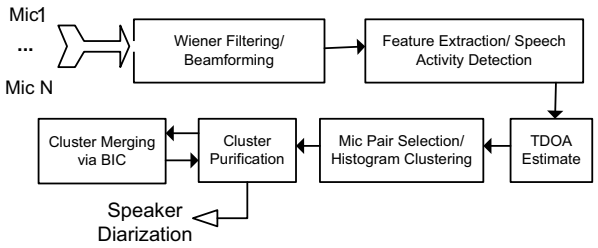

Figure 1: Block diagram of the speaker diarization system

2.1. Wiener Filtering and Beamforming

In the multi-distant microphone recording scenarios, the quality of the recording data are poor. In order to achieve good speech activity detection performance and speaker clustering, we have applied the Winner filtering to all the distant microphone channel data. Based on these enhanced channel speech, a single enhanced channel is obtained by using the beamforming approach. We simply adopt the Qualcomm-ICSI-OGI front end [7] tool and apply its Wiener filtering to all audio channels for speech enhancement. The enhanced audio channels are then filtered and summed to produce a beamformed audio channel using the BeamformIt toolkit [8].

2.2. Speech Activity Detection (SAD)

One of the improvements for SAD is the introduction of a new speech/non-speech detector. In the SAD, we first generate 36 MFCC features (12 MFCCs plus their first and second order derivatives) and the zero-crossing rates of each frame (in 30ms window and 15 ms shift). This feature is also used for the cluster purification describe in Section 4. We use all the features in each meeting recording to train initial speech and non-speech models by Expectation Maximization (EM) method [10]. The 10 percent frame features with the highest energies and relative low zero-crossing rates are selected as training data set for the speech modeling, while the 20 percent frame features with the lowest energies and relative higher zero-crossing rates as the training data for non-speech modeling [5]. Based on these two initial models, we test all the frames in each meeting recording and classify them into speech and non-speech. These classified frames are used to iteratively re-train the speech and non-speech models based on the Maximum a Posteriori (MAP) [10] approach, until the relative change of detected speech to non-speech ratio is less than 1%. The re-training processing is usually less than 10 iterations. The speech and non-speech are modeled in Gaussian mixture models (GMMs) with 16 and 4 mixture component, respectively. The data in the NIST RT05 and RT06 have been used as the development test.

2.3. TDOA Estimation

This component provides the TDOA estimation [9] for all microphone pairs which will be passed to next component for further processing. The TDOA is the result of different placements of the microphones in the meeting room. The difference in placements means that speech originating from a source arrives at the microphones at different times. We make an assumption here that the speech source does not move. Any change in TDOA can be interpreted as a change in the speaker. Instead of applying the Normalized Least-Mean Square (NLMS) method to estimate the time delay of arrival in our previous paper [6], we use a much faster GCC-PHAT [9] method to estimate the TDOA for the MDM task. Given that K pairs of microphones are used to estimate the TDOA, the matrix $TDOA[n,k]$ stores the TDOA values. The TDOA values can be estimated as follows:

For any two microphone signals $S_i^k(n)$ and $S_j^k(n)$ of a given microphone pair k, while *n* is the segment index, GCC-PHAT is given as [9]:

$$R_{PHAT}^k(f,n) = \frac{FFT[S_i^k(n)] \cdot FFT[S_j^k(n)]^*}{\left| FFT[S_i^k(n)] \cdot FFT[S_j^k(n)]^* \right|^\alpha} \quad (1)$$

where FFT is the Fast Fourier transform and the * is the complex conjugate. And α is the normalized factor (between 0 and 1). The TDOA for the microphones segment signals $S_i^k(n)$ and $S_j^k(n)$ is obtained as:

$$TDOA_{i,j}(n,k) = \arg\max_f (IFFT[R_{PHAT}^k(f,n)]) \quad (2)$$

where *IFFT* is the Inverse Fast Fourier Transform, *n* is the segment index, and *k* is the index of microphone pair.

3. Histogram Quantization Clustering

Generally speaking, the overall performance of a speaker diarization system can be largely affected by the initial speaker clustering. After a successful data preparation described in Section 2, this section is to carry out the initial speaker clustering, which will be done using the histogram quantization. Before the histogram quantization initial clustering, we first present how to automatically choose useful microphone pairs from multiple distant microphones.

3.1. Microphone Pairs Selection

It is a normal case in MDM meeting where a number of microphones are available with a large number of microphone pair permutations. For example, in the RT-07 EDI 20050216-1051 task, it has 8 distant microphones in a circle array. Such array has possible 28 possible sensor pairs in combination. It is not wise to choose microphone pairs according to the physical location of multiple distant microphones because sometimes such information may not be available. Since the TDOA is the result of the different placements of the microphones in a meeting room, therefore, the higher dynamic range of TDOA histogram peak distribution, the better it presents the speaker physical position in the meeting room. Under such circumstances, we can judiciously choose pairs that generate discriminative peak distribution automatically by:

a) Compute the TDOA values of all the possible combination of microphone pairs using (2) with normalized factor 0.995.
b) Generate the histogram distribution of each pair's TDOA vectors and find the peaks, which satisfy the following threshold:

$$T_{Peak\_Thres} = \frac{T}{B} \quad (3)$$

where *T* is the total number of frames and *B* is the histogram bin number.
c) Find dynamic ranges among these peaks (peak to peak distance).
d) Choose the largest 5 peak dynamic microphone pairs if the microphone pairs in each meeting are greater than 5.

3.2. Within-Pair-Quantization

With the selected microphone pairs in Section 3.1, we are able to perform within-pair-quantization on the TDOA values. A histogram is made from the TDOA values. Peaks can be located in the histogram and these peaks match the physical locations where much of the speech signals originate. The peaks will be used as centroids and all other frames are clustered to one of these centroids using a nearest neighbor approach. Then, the quantized TDOA information from multiple microphone pairs is merged by performing a second level of quantization. Every speech frame will have a set of labels obtained by performing within pair quantization on all retained microphone pairs. The unique labels are clustered into initial speaker clusters. Continuous frames with the same cluster labels then are grouped together to form continuous segments. The details are described in the following.

For the k[th] microphone pair, the $TDOA[n,k]$ values are quantized to locate frequently occurring TDOA positions for this pair. The frequently occurring positions are those $TDOA[n,k]$ found by constructing a histogram. Every peak in the histogram indicates that there is a significant amount of speech signals originating from that particular location. We make the assumption that speech signals originating from a single location will very likely belong to a single homogenous speaker. As such, the number of peaks can be used as an estimation of the number of speakers present in the meeting. The peaks in the histogram are taken as centroids and

$TDOA[n,k]$ values are quantized to these centroids using a nearest neighbor approach.

Another important issue is how to decide the peak threshold and the number of the histogram bins used for the TDOA quantization. From the development data from the NIST RT06, we have found that we can achieve the robust results according to the rule for the peak threshold selection as follows.

If a meeting recording has a total of *N* speech frames, each having 512 samples in 32ms based on 16k sample rate, and the number of the bins used for the histogram quantization is *B*. We used the threshold given in (3) for the peak detection threshold. Only the peak of the bins in the histogram which is greater than the threshold T_{Peak_Thres} will be regarded as the within-pair centroids. In addition, we have found that it is a good setting if the number of the histogram bins, B, is chosen between 40 and 50 in RT06 data. We will apply this setting for the RT07 data.

3.3. Inter-Pair Quantization

The previous section discusses quantization along the columns of $TDOA[n,k]$. The next step will perform quantization along the rows of $TDOA[n,k]$, i.e., to identify centroids across K microphones pairs. The same procedures as in Section 3.2 are used to find the row peak in $TDOA[n,k]$ matrix. Based on these two quantization matrix, we choose the *N* highest centroids as our initial clusters. All other histogram bins with low counts will be quantized to one of these centroids. We choose a reasonable number (e.g. *N*=9) as our initial speaker clustering number.

The number of the initial speaker clusters is usually greater than actual number of speakers in the meeting. In the subsequent modules, we will exploit acoustic features for the cluster purification and cluster merging.

4. Cluster Purification and Merging

In Section 3.3, we have pointed out that only location-based information (TDOA) is used for histogram quantization clustering. Obviously, if the speakers had moved significantly or changed positions during the meeting, it is postulated that the results would be significantly poorer. In addition, we selected the *N* highest centroids as our initial speaker clusters. Obviously, we need to have extra efforts to obtain a better speaker clustering. In this section, a speaker cluster purification algorithm based on high-order GMM modeling approach and a cluster merging via BIC Criteria are conducted. A high-order statistical GMM approach is applied for segment purification and BIC based criterion is used for clustering merging. Only 36 MFCC features defined in Section 2.2 are used again in this section.

4.1. Cluster Purification

We access the reliability of the clusters obtained in the initial clustering. We employ the consensus [11], [12] or statistical 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 statistical information for cluster purification. The procedures are as follows:

a) Use the EM algorithm [10] to train a GMM with *M=256* Gaussian mixture components in which the covariance matrix is diagonal, by using all the data, named as GMM-Root.
b) For each speaker cluster obtained from the initial clustering in Section 3.3, the corresponding feature vectors are then to obtain a GMM by adaptive training from GMM-Root. The adaptation is done with the mean vectors only via MAP algorithm [10]. For *N* initial speaker clusters, their GMMs are GMM-1 ... GMM-N.
c) Score every segment against the *N* GMMs and assign each segment to the cluster whose GMM yields the highest likelihood score.
d) A new set of *N* GMMs re-adapted using the feature vectors assigned in Step c.
e) Repeat the Step a) to c), until the segment assignments have been stabilized (usually in less than 20 iterations).
f) Reduce the number of Gaussian components in the GMM by 8 each time and repeat the Step a) to f) (each time. M=M-8) until the number of Gaussian components has been reduced to 32.
g) Based on the above multi-iterative runs, each cluster only chooses the segments with more than 60% total iterative runs counts inside its cluster, and use them to re-train the GMM of this cluster via EM. The GMM size is set to 256. We found that we still have the same cluster number *N* GMM_1 to GMM_N.
h) The remaining segments in each of the clusters are then scored against GMM-1 to GMM-N. These segments will be assigned to the cluster whose GMM yields the highest likelihood score.

4.2. Cluster Merging via BIC

Based on above-mentioned cluster purification, BIC criterion is applied to check if some clusters can be further merged together. Suppose that there are two initial clusters X and Y, and let *Z=X U Y* to be the merged cluster. With the BIC criterion, we can decide whether these initial clusters are from the same speaker or not as follows:

$$T_{thres} = S_1 - S_0 - \frac{\lambda}{2}\log(M_x + M_y) \tag{4}$$

and

$$S_1 = \sum_{i=1}^{M_x}\log p(x_i \mid \theta_z) + \sum_{i=1}^{M_y}\log p(y_i \mid \theta_z) \tag{5}$$

$$S_0 = \sum_{i=1}^{N_x}\log p(x_i \mid \theta_x) + \sum_{i=1}^{N_y}\log p(y_i \mid \theta_y) \tag{6}$$

where the λ is the penalty factor (set to be 1) and *M* is total number of frames in the cluster X or Y.

Based the purified *N* clusters in Section 4.1, we have $N(N-1)/2$ combination of BIC scores. If lowest score in these BIC values is less than zero, we will merge the related two clusters into one. This merging procedure is performed until the lowest score of clusters BIC scores is greater than zero. We use EM to train the model and the model size is set to be 4.

After completing the first round purification and the cluster merging, we repeat the purification process described in Section 4.1 again and obtain the final speaker diarization results.

5. Experiments on RT07

Based on the above proposed system, we evaluated our system on the NIST RT07 MDM evaluation task. The NIST RT05 and RT06 MDM evaluation corpus are used as the development data to fine tune the system for the NIST RT07 MDM

evaluation corpus. The system is also used to evaluate the speaker diarization system of the 2009 NIST RT MDM task. The system performance is evaluated by computing the Diarization Error Rate (DER) against the official Rich Transcription Time Mark (RTTM) released by NIST [1].

Table 1 and Table 2 present the speech activity detection (SAD) DER rates and speaker diarization rates on the NIST RT07 MDM evaluation task. The evaluations consist of 8 tasks, as listed in Table 2. The number of microphone recordings ranged from 3 (for the CMU tasks) to 16 (for the EDI tasks). The recordings for all tasks are made using distant microphones or microphone arrays. Table 3 summarized the results of NIST09 MDM task.

In Table 1, the SAD module produced significant improvements. It reduces an absolute reduction of 11.2% in DER, yields an absolute reduction of 5.6% over the SAD in our previous system [6]. The results are also consistent with ICSI results reported in [5].

Table 1. RT07 Speech Activity Detection Rates

	SAD DER (%)	SAD Missed Speaker		SAD False Alarm Speaker	
		seconds	%	Seconds	%
Without SAD	14.3	0.0	0.0	1005.2	14.3
With SAD	3.1	85.2	1.2	170.2	1.9
Previous System SAD [6]	8.7	326.3	4.7	280.3	4.0

Table 2. RT07 Speaker Diarization Error Rates

Task	DER (%)				Previous System [6]
	Histogram Clustering	Initial Purifying	BIC Merging	Final Purifying	
1	18.07	18.04	15.90	15.80	19.36
2	7.98	8.27	7.34	7.40	12.46
3	13.87	13.91	13.43	11.04	20.69
4	37.13	37.04	5.65	5.86	15.00
5	35.79	33.67	8.23	6.71	12.66
6	17.87	7.53	5.70	5.29	13.36
7	31.29	17.95	11.08	5.26	11.32
8	36.98	34.23	15.08	9.61	18.45
All	**24.87**	**21.13**	**10.27**	**8.31**	**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

Table 3. Summarized Results of RT09 MDM Task

Task	Speaker Diarization Error DER (%)				SAD DER(%)
	Histogram Clustering	Initial Purifying	BIC Merging	Final Purifying	
All	**29.63**	**28.00**	**10.04**	**9.54**	**2.74**

According to the results shown in Table 2, the system yielded an overall DER of 24.87% after histogram quantization clustering which provides a good start point for the MDM speaker diarization.

After applying the first purification (without BIC merging), the system yielded an overall DER of 21.13%. We observed that for some meeting recordings, this component can significantly reduce their DER rates, e.g. for VT 20050408-1500, its DER drops from 31.29% down to 17.95%, and for NIST_20060216-1347, it drops from 17.87% down to 7.53%. More important, this component has purified the initial clusters and make it easy for BIC based merging process to merge the initial speaker clusters into the final speaker clusters.

In the BIC merging component, the overall DER has been reduced from 21.13% to 10.27%. It is observed that the system has identified exact same number of speakers with the actual number of speakers for RT07 tasks. In our primary experiment, we found that the BIC merging module can not accurate identify the correct numbers without the initial purification.

After the cluster merging, the purification is further applied on the merged speaker clusters. Finally, the system yielded an overall DER 8.31%. This system greatly outperformed our previous system [6] with overall DER of 15.32%.

From Table 3, we can see that our proposed system achieved similar results with RT07 for RT-09 evaluation data.

6. Conclusions

The proposed improved speaker diarization system is found to yield much better performance on NIST RT07 evaluation set in a comparison to our previous system [6] that was submitted to NIST RT07 Meeting Recognition MDM evaluation. The results show that the proposed speech activity detector is effective for non-speech removal component. In addition, the proposed statistical purification strategy and merging scheme help to improve overall speaker diarization performance and yield an overall DER of 8.31%.

References

[1] Spring 2007 Rich Transcription meeting recognition evaluation plan - http://www.nist.gov/speech/tests/rt/rt2007/docs/rt07-meeting-eval-plan-v2.pdf.

[2] D. A. v. Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06smeeting data," in *Proc. NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop*, Washington DC, pp. 371-384, 2006.

[3] X. Anguera, C. Wooters, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, San Juan, 2005.

[4] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J. F. Bonastre, "NIST RT05S evaluation : Pre-processing techniques and speaker diarization on multiple microphone meetings," in *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh, 2005.

[5] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," *Lecture Notes in Computer Science, vol. 4625*, pp. 509-519, 2008.

[6] E.C.W. Koh, H.W. Sun, T.L. Nwe and T.H. Nguyen, B. Ma, E.S. Chng, H. Li and Rahardja, S., "Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information: The I^2R-NTU Submission for the NIST RT 2007 Evaluation," *Lecture Notes in Computer Science, vol. 4625*, pp. 484-496, 2008.

[7] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-icsi-ogi features for asr" in *Proc. ICSLP*, vol. 1, pp. 4-7, 2002.

[8] BeamformIt acoustic beamformer. - http://www.xavieranguera.com/beamformit/.

[9] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. ICASSP*, pp. 375-378, 1997.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

[11] T.L. Nwe, H. W. Sun and H. Li*,"*, "Spkeaker Diarrization in Meeting Audio', *ICASSP*, pp.4073-4076, Taipei 2009.

[12] 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.