Speaker diarization system for RT07 and RT09 meeting room audio

Conference Paper *in* Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · April 2010

DOI: 10.1109/ICASSP.2010.5495077 · Source: IEEE Xplore

4 authors, including:

Bin Ma
Institute for Infocomm Research
218 PUBLICATIONS **2,214** CITATIONS

SEE PROFILE

Haizhou Li
National University of Singapore
658 PUBLICATIONS **8,170** CITATIONS

SEE PROFILE

SPEAKER DIARIZATION SYSTEM FOR RT07 AND RT09 MEETING ROOM AUDIO

Hanwu Sun [1], Bin Ma [1], Swe Zin Kalayar Khine [1] and Haizhou Li [1,2]

[1] Institute for Infocomm Research (I²R), A*STAR, 1 Fusionopolis Way, Singapore 138632
[2] Department of Computer Science and Statistics, University of Eastern Finland, FI-80101 Joensuu, Finland
{hwsun, mabin, zkkswe, hli}@i2r.a-star.edu.sg

ABSTRACT

This paper describes an improved speaker diarization system for the Single Distant Microphone (SDM) task in the 2007 and 2009 NIST Rich Transcription Meeting Recognition Evaluations. The system includes three main modules: front-end processing, initial speaker clustering and cluster purification/merging. The front-end processing involves the Wiener filtering for the targeted audio channels and a self-adaptation speech activity detection algorithm. A simple but effective energy based segmentation is applied to chunk the meeting data into small segments to construct the initial clusters. An enhanced purification algorithm is proposed to further improve the performance after the preliminary purification, and the BIC criterion is adopted for the cluster merging. The system achieves competitive overall DERs of 15.67% for RT07 SDM speaker diarization task and 17.34% for RT09 SDM speaker diarization task.

Index Terms: Single Distant Microphone, speaker diarization, speech activity detection, speaker clustering

1. INTRODUCTION

Speaker diarization is one of the tasks in the NIST Rich Transcription (RT) Meeting Recognition Evaluation. It detects "who spoke when" in a meeting by automatically finding the time stamps as to when each meeting participant is talking [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 both Single Distant Microphone (SDM) condition in which a single, centrally located microphone is involved for each meeting, and Multiple Distant Microphones (MDM) condition in which at least 3 microphones are placed on a table in a meeting room [2-9].

In state-of-the-art speaker diarization systems, three components have much impact on the performance, namely, 1) Speech activity detection (SAD), that removes the silence\non-speech from the audio. 2) Initial speaker clustering, which is of great importance to the speaker purification and merging in a later stage. In speaker diarization under MDM condition, Time Difference of the Arrival (TDOA) [12] information based on multiple microphones has been widely used to bootstrap a speaker clustering. As a result, the performance of the MDM tasks is generally much better than that of the SDM tasks [5-9] in which TDOA information is not available. 3) Cluster purification/merging that optimizes the overall performance of a speaker diarization system. Most studies in speaker diarization have been focused on this component [2-9].

In our previous works [6, 9], we developed a GMM based speaker diarization system for the MDM tasks of the NIST 2007 and 2009 Rich Transcription Meeting Recognition Evaluations (hereinafter referred as to RT07 and RT09). The system applied a self-adaptation speech activity detection for silence/non-speech removal, a two-stage histogram quantization strategy based on TDOA information for speaker initial clustering, and a statistical cluster purification method for the final speaker diarization. It produced a good performance on the MDM tasks of both RT07 and RT09 with Diarization Error Rates (DERs) of 8.31% and 9.51%, respectively [6].

In this paper, we extend the previous MDM speaker diarization system to the SDM task by applying the same self-adaptation speech activity detection algorithm as in [6]. We introduce a simple, yet effective initial clustering, and a 2-step purification and merging process. Since there is no TDOA information available in the SDM task, an energy based segmentation method is used to break long segment into relatively short segments during the initial clustering. In the 2-step purification and merging, an enhanced purification and merging is applied after the preliminary purification, where the BIC criterion [6-9] is adopted to guide the cluster-pair merging. We will report the experimental results on both RT07 and RT09 SDM tasks.

The rest of this paper is organized as follows. In Section 2, the front-end of the speaker diarization system is described. In Section 3, the initial clustering and two-step purification and merging strategy is presented. In Section 4 we present the experimental results and finally we conclude in Section 5.

2. FRONT-END PROCESSING

The speaker diarization system for the SDM task is depicted in Figure 1. The front-end processing of the system consists of 3 components as follows:

a) Wiener Filtering
b) Feature Extraction
c) Speech Activity Detection (SAD)

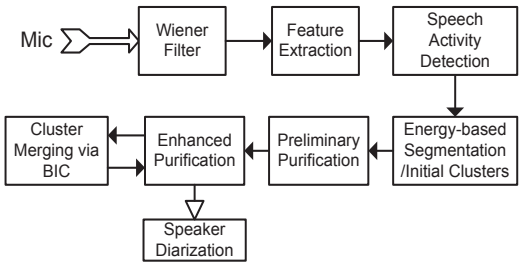

Figure 1: Block Diagram of the Speaker Diarization System

2.1. Wiener Filtering

In the single distant microphone recording scenarios, the quality of the recording data is relative poor. In order to improve the performance of the speech activity detection and speaker clustering, the Winner filtering has been applied on the targeted distant microphone channel data. We adopt the Qualcomm-ICSI-OGI front end tool [10, 11] and apply its Wiener filtering to the targeted audio channels for speech enhancement.

2.2. Speech Activity Detection

We apply the same speech activity detection (SAD) algorithm as in the previous work [6]. A vector of 36 MFCC features (12 MFCCs plus their first and second order derivatives) and the zero-crossing rates of each frame in a 30 ms window with 15 ms shift are generated. We use all the feature vectors extracted from each meeting to train the initial Gaussian mixture models (GMMs). Speech and non-speech models are trained separately using Expectation Maximization (EM) algorithm [13]. The top 10% feature frames of the highest energy and relative low zero-crossing rates are selected as training data set for the speech GMM, while the top 20% feature frames of the lowest energy and relative high zero-crossing rates as the training data for non-speech GMM [4, 6]. Based on such two initial models, we classify all the feature frames into speech and non-speech. The classified frames are then used to iteratively re-train the speech and non-speech GMMs based on the Maximum a Posteriori (MAP) approach [13], until the relative change of detected speech/non-speech ratio is less than 1%. The re-training processing is usually completed in less than 10 iterations. The two GMMs for speech and non-speech have 16 and 4 mixture components, respectively. The data in the NIST RT05 and RT06 have been used as the development test.

3. CLUSTER PURIFICATION AND MERGING

Unlike the MDM task, SDM task doesn't have location-based TDOA information for speaker initial clustering. Therefore, we first introduce a simple energy based frame segmentation approach to break the long segments into segments of a few seconds long as the initial clustering. A 2-stage purification and merging process is introduced to enhance the performance. The BIC based criterion is used for clustering merging [6]. 36 MFCC features (without the zero-crossing rate) are adopted in these tasks.

3.1. Energy-based Segmentation and Initial Clustering

We note that a short speech segment contains insufficient discriminative information to establish the identity of a speaker; on the other hand, a long speech segment may involve multiple speakers, thus making it difficult to separate them in a later stage. Viterbi decoding [7, 8] was usually adopted for the segment alignment. In this paper, we introduce a simple energy-based segmentation approach to break a long speech segment into short speech segments.

Suppose that m is the pre-defined length of the segments in seconds while m is usually set to few seconds and decided by experience [4]. With the SAD, the meeting recording is broken into multiple segments, providing a good initial segmentation. If a segment is still greater than m seconds, we will search the frame with the minimum energy in the range from $m-\Delta m$ seconds and $m+\Delta m$ seconds and break such a segment further into smaller segments, where Δm is less than m. At last, the whole meeting recording is broken into short segments of about m seconds in length each. These short segments constitute the N initial speaker clusters with an equal number of segments. Typically N is much larger than the actual number of speakers in the meeting.

3.2. Preliminary Purification

Based on the above-mentioned initial segmentation clustering, we adopt GMM clustering purification method [6] for the preliminary purification. Multiple iterations of the clustering are conducted in this preliminary purification to purify the speaker clusters. The detailed procedures are as follows:

a) Use the EM algorithm [13] to train a GMM with M *(M=4 initially)* Gaussian mixture components with a diagonal covariance matrix, by using all the data in the meeting recording, named as GMM-Root.
b) For each speaker cluster obtained from the initial clustering in Section 3.1 or updated in the preliminary purification, a GMM is adapted from GMM-Root

(mean adaptation only) by using all the feature vectors in this speaker cluster via the MAP adaptation process [13], resulting in *N* GMMs, denoted as GMM-*1* ... GMM-*N,* for *N* initial speaker clusters.

c) Evaluate all the segments against the resulting *N* GMMs and assign a segment with the cluster whose GMM yields the highest likelihood score.
d) A new set of *N* GMMs are re-trained via the MAP algorithm using the segments assigned in Step c).
e) Repeat the Step c) to d), for several times until no segments change are detected in the clusters.
f) Increase the number of Gaussian mixture in the GMMs by 2 times ($M = M \times 2$ *)*. Repeat the Step a) to e) until $M = 64$.

In the preliminary purification, each of the segments is involved in the GMM training, and assigned to one of the GMMs. We regard this purification process as closed-clustering.

3.3. Enhanced Purification and Cluster Merging

With the results from preliminary purification, we further introduce an enhanced purification process to improve the purification of the speaker clusters. In this process, we adopt an open-clustering strategy as opposed to closed-clustering. The speaker cluster merging is made based on the BIC criterion [6]. This purification process aims to refine the clustering results.

The enhanced purification is similar with the preliminary purification, except that a segment is excluded from the GMM training if it will be evaluated against a GMM. This is to ensure that the evaluation of segments is an open test. After this enhanced purification process, we will re-compute the cluster-pairs BIC matrix. The cluster-pair with the largest merged score is then merged. The enhanced purification and cluster merging will then repeat until the largest cluster-pair merged score is less than zero. The procedures are described as follows:

a) Use the EM algorithm [13] to train a GMM with *M=64* Gaussian mixture components with a diagonal covariance matrix, by using all the data in the meeting, named as GMM-Root.
b) For each speaker cluster obtained from the cluster purification in Section 3.2 or updated in the enhanced purification, a GMM is adapted over all the feature frames in this speaker cluster with the MAP algorithm [13] from GMM-Root (mean adaptation only). For *N* initial speaker clusters, their GMMs are denoted as GMM-*1* ... GMM-*N*.
c) *Do i=1*:*n*, where *n* is number of the speech segments.
d) Re-train the *i*-th segment related GMM by excluding the *i*-th segment features using MAP algorithm.
e) Score the *i*-th segment against the *N* GMMs and assign the *i*-th segment to the speaker cluster whose GMM yields the highest likelihood score.
f) *End i*.
g) Repeat the Step b) to f), for twice.
h) Compute the cluster-pair BIC matrix [6] and find the largest merged score.
i) If the largest BIC Merged score is less than *0*, output the current results and STOP.
j) Merge the cluster-pair (reducing cluster number *N=N-1)*, Go to Step b).

4. EXPERIMENTS ON RT07 AND RT09

We evaluate the above proposed speaker diarization system, on the RT07 and RT09 SDM tasks. The NIST RT05 and RT06 SDM evaluation corpus are used as the development data to fine tune the system. The performance is evaluated by computing the Diarization Error Rate (DER) against the Rich Transcription Time Mark (RTTM) released by NIST [1].

Table 1 shows the performance of SAD on both RT07 and RT09. Both of the SDM and MDM tasks have been included for the comparison. Table 2 presents the performance comparison between the uniform chunking and energy-based segmentation for the initial clustering on the RT07 and RT09 corpus. Table 3 and Table 4 show the final speaker diarization rates on the RT07 and RT09 SDM tasks, respectively. The evaluations consist of 8 meeting recordings for RT07 and 7 meeting recordings for RT09, as listed in two tables. All the meeting recordings were made using one targeted distant microphone or a given channel from microphone arrays [1]. We used 2 seconds with 0.25 second offset ($m = 2$, $\Delta m = 0.25$) for the initial segmentation and set the initial number of cluster *N* to be 20 for the results in Tables 2, 3 and 4.

Table 1. RT07 and RT-09 Speech Activity Detection Rates

	SAD DER (%)	SAD Missed Speaker		SAD False Alarm Speaker	
		Seconds	%	Seconds	%
RT-07-MDM	**3.1**	85.2	1.2	170.2	1.9
RT-07-SDM	**3.4**	93.5	1.3	187.1	2.1
RT-09-MDM	**2.7**	76.7	1.0	135.1	1.8
RT-09-SDM	**2.6**	62.8	0.8	135.3	1.8

From Table 1, we can see that the SAD module produced consistent results between RT07 and RT09 SDM tasks. At the same time, it is also observed that the SAD performance in SDM tasks is comparable with that in MDM tasks, indicating that the SAD proposed for MDM tasks is also robust and efficient for the SDM tasks.

Table 2 shows that the energy-based segmentation is an effective method for the initial clustering, providing a better performance than the uniformly chunking. We have

conducted the experiments with different segmentation lengths and found that m=2 seconds is a good setting for our GMM based speaker diarization system.

Table 2. RT07 and RT-09 Speaker Diarization Error Rates Using Uniform and Energy-based Segmentation

RT07 SDM DER		RT09 SDM DER	
Uniform	**Energy-based**	**Uniform**	**Energy-based**
17.23%	15.67%	18.12%	17.34%

Table 3. RT07 Speaker Diarization Error Rates

Task ID	With overlap	Without overlap
CMU_20061115-1030	30.12%	23.76%
CMU_20061115-1530	11.02%	8.06%
EDI_20061113-1500	22.94%	15.91%
EDI_20061114-1500	23.63%	21.50%
NIST_20051104-1515	10.99%	7.88%
NIST_20060216-1347	10.13%	8.32%
VT_20050408-1500	6.27%	5.81%
VT_20050425-1000	16.17%	11.76%
All	**15.67%**	**12.14%**

Table 4. RT-09 Speaker Diarization Error Rates

Task ID	With overlap	Without overlap
EDI_20071128-1000	7.56%	4.61%
EDI_20071128-1500	18.66%	12.50%
IDI_20090128-1600	8.00%	4.38%
IDI_20090129-1000	14.56%	11.24%
NIST_20080201-1405	59.40%	46.28%
NIST_20080227-1501	11.11%	2.75%
NIST_20080307-0955	17.46%	13.69%
All	**17.34%**	**12.10%**

In Table 3 and Table 4, the SDM evaluation tasks in RT07 and RT09 at both the overlapping speaker condition and non-overlapping speaker condition are presented while the best submitted result for the SDM task in RT07 is 21.74% [5], From Table 3, we can see that the proposed speaker diarization system achieves a promising result. The performance in RT09, shown in Table 4, is as good as that in RT07. After we subtract the without-overlapped DER rate from with-overlapped DER rates in Tables 3 and 4, we obtain the overlapped section only DER about 3.5% for RT07 and 5.2% for RT09. This suggests that RT09 dataset has higher amount of overlapped segments than RT07.

5. CONCLUSIONS

The improved speaker diarization system has been successfully extended to the SDM tasks of RT07 and RT09, and found to yield much better performance on RT07 SDM evaluation set in comparison to other published RT07 results. The results show that the simple energy-based segmentation is effective for the initial clustering, providing a better solution for segmenting the meeting data than the uniform chunking method. In addition, the proposed 2-step purification strategies and merging scheme helps to achieve good speaker diarization performance with an overall DER of 15.67% for RT07 SDM task and DER of 17.34% for RT09 SDM task.

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

[5] J.G. Fiscus, J. Ajot and J.S. Garofolo, "The Rich Transcription 2007 Meeting Recognition Evaluation", , http:www.nist.gov/speech/tests/rt/2007/workshop/RT07-SPKR-v7.pdf". *Lecture Notes in Computer Science, vol. 4625*, pp. 373-389, 2008.

[6] H.W. Sun, T.L. New, B. Ma and H.Z. Li, "Speaker Diarization for Meeting Room Audio", *Interspeech 2009*, pp. 900-903, Brighton, U.K., 2009.

[7] T.H. Nguyen, H.Z, Li and E.S. Chng, "Cluster Criterion Function in special Subspace and Their Application in Speaker Clustering", *ICASSP 2009*, pp.4085-4085, Taipei, April 2009.

[8] A.G. Friendland, B.O. Vinyals, C.Y. Huang and D.C. Muller, "Fusion Short Term and Long Term Features for Improved Speaker Diarization", *ICASSP 2009*, pp.4077-4080, Taipei, April 2009.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.