Speaker Diarization in Meeting Audio for Single Distant Microphone

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

Human Language Technology Department, Institute for Infocomm Research (I^2R), A*STAR, Singapore 138632

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

Abstract

This paper presents speaker diarization system on NIST Rich Transcription 2009 (RT-09) Meeting Recognition evaluation data set for the task of Single Distant Microphone (SDM). A two-step speaker clustering method is proposed. The first step is speaker cluster initialization using speech segments of meeting audio, where we randomly pick a small subset of speech segments and merge them iteratively into a number of clusters. And, the second step is cluster purification, where we introduce a consensus-based speaker segment selection method for efficient speaker cluster modeling that purifies the clusters. The system achieves a promising diarization error rate (DER) of 16.4%.

Index Terms: speaker diarization, rich transcription, single distant microphone

1. Introduction

Speaker diarization is to segment a speech signal according to speaker homogeneous region. In other words, a speaker diarization system is to provide an answer to a question of "Who spoke when?". The information on when each speaker is speaking can be used as the preprocessing step in speech recognition systems [1]. For example, using the output of speaker diarization, vocal tract length normalization and/or speaker adaptation can be carried out in speech recognition systems. Due to the importance of the task, it has been evaluated in the NIST RT evaluation for several years.

There exist many challenges to build a speaker diarization system. Among them, one is to estimate the number of speakers appearing in the recording; another is to group the speech segments of same speaker identity together. This can be referred to as a 'speaker clustering' problem.

There are two general methods for speaker clustering [2]. These are bottom-up and top-down approaches. The bottom-up approach starts with a number of clusters O (which exceeds the predicted number of speakers) and aims to successively merge and reduce the number of clusters until there remains only one for each speaker. Agglomerative Hierarchical Clustering (AHC) method is an example of bottom-up approach. As for top-down approach, the audio is first modeled as a single speaker model. Then, new speaker models are successively added until the full numbers of speakers are deemed to be accounted for.

Several works have been reported in solving the 'speaker clustering' problem. Agglomerative Hierarchical Clustering (AHC) has been the most widely used method in speaker diarization systems [3, 4]. In [3], an investigation is made to reduce manual tuning of initial parameters in AHC. This study proposes an approach for a novel initial parameter estimation method for AHC with Bayesian Information Criterion (BIC) and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs). In [2], improvement to traditional top-down approach is proposed by using expectation maximization (EM) instead of using maximum a posteriori (MAP) adaptation of a background model for speaker modeling. One of the common approaches is the energy based segmentation method followed by a speaker cluster purification process [5], where an open test cluster purification approach is employed to purify the initial clusters.

For speaker clustering, AHC method has been the most popular in speaker diarization [4]. Its bottom-up approach normally starts with 16 to 20 clusters [3] over the entire meeting before cluster merging process. When a starting cluster contains a speech segment which is long, it can include more than one speaker. It is therefore difficult to remove wrong speaker frames in later stages. However, if AHC method starts with a cluster that has a relatively short speech segment to have better cluster purity, it can be computationally very expensive in merging process. For example, if there are O starting clusters, there will be $O!/(2!(O-2)!)$ cluster pairs to consider for merging. Furthermore, in AHC method, an appropriate stopping criterion is necessary to stop merging process when optimal speaker clusters are obtained.

For cluster purification, re-segmentation process is carried out in AHC. This process will be more effective if relatively pure speaker models are available. However, AHC method is not designed to remove impure speaker segments from clusters before re-segmentation. In general, a speech segment is referred to as an 'impure speaker segment' if its speaker identity is different from the speaker identity of the majority of speech segments in a cluster, otherwise referred to as a 'pure speaker segment'.

In this paper, we propose a speaker clustering methodology in an effort to accomplish three tasks: 1) to start each cluster with a speech segment of unique speaker identity, but at a reduced computational load, 2) to circumvent the need of a stopping criterion in clustering and 3) to adopt a scheme to use relatively pure speaker models in cluster purification process.

The rest of the paper is organized as follows. Section 2 describes the proposed speaker diarization system. Section 3 presents the front-end of the proposed speaker diarization system. Section 4 explains initial speaker clustering and purification method. Section 5 presents experimental results and Section 6 concludes the paper.

2. Speaker diarization system

In general, a speaker diarization system has three main steps [6]. The first one is voice activity detection which divides speech and non-speech regions in a recording. The second one is detecting speaker turning points. And, the final step is speaker clustering that groups the speech segments with same

speaker identity together. In this paper, we are only interested in 'speaker clustering'.

To accomplish the 1st and 2nd tasks as stated in Section 1, we propose a random picking and merging process for initial clustering. Initial clustering process starts with small clusters with possible unique speaker identity. To reduce the computational load for selecting a cluster pair to merge, merging process is done on a small subset of speech segments iteratively. As the number of target cluster to achieve is pre-defined, the stopping criterion for merging process is thus not necessary.

As for the 3rd task, we propose a consensus [7] based cluster purification method to remove impure speaker segments in speaker modeling. The block diagram of our proposed system is shown in Figure 1.

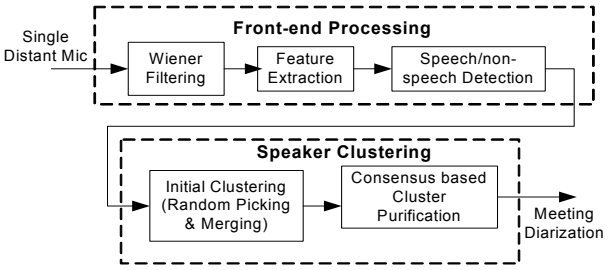

Figure 1: *Diagram of the proposed speaker diarization system*

3. Front-end processing

The front-end preprocessing has three steps as shown in Figure 1. The first step is wiener filtering. The second step is feature extraction and the third step is speech/non-speech detection. The following sections present the details.

3.1. Wiener filtering

The audio quality is relatively poor in single distant microphone recordings in meeting audio. To improve the audio quality, we apply Wiener filtering. The Wiener filtering method we applied is based on tracking a priori SNR using decision-directed method [8].

3.2. Feature extraction

We extract two sets of features. One is for speech/non-speech detection and another is for speaker clustering experiments. For speech/non-speech detection, the recording is divided into frames of 30ms with 15ms overlapping. From each frame, a vector of 36 MFCC features (12 MFCCs together with their first and second order derivatives) and zero-crossing rate are extracted. For speaker clustering experiments, the frame size is 20ms with 10ms overlapping. A total of 12 MFCC features with their first derivatives are extracted from each frame.

3.3. Speech/non-speech detection

We employ the voice activity detection algorithm that is the same as in our previous work [5]. We extract the MFCC features as mentioned in Section 3.2 from each recording to train initial Gaussian models (GMMs). Speech and non-speech models are trained separately using Expectation Maximization (EM) algorithm [9]. The top 10% of the feature frames of the highest energy and relative low zero-crossing rates are selected as training data set for speech GMM. And, top 20% feature frames of the lowest energy and relative high zero-crossing rates are chosen as training data for non-speech GMM [10, 11]. 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 [9], until the relative change of detected speech/non-speech ratio is less than 1%. The re-training process 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 set.

After speech/non-speech detection, we obtain a number of speech segments. Note that a long speech segment may contain multiple speakers; we break a long speech segment into small ones. In practice, segments longer than 5 seconds are broken into 2.5 seconds each. This is part of the preprocessing to prepare the segments for speaker clustering. Mean and Standard Deviation (Std) of durations of speech segments after such breaking are summarized in Table 1.

Table 1. *Mean and Standard Deviation (Std) of durations of speech segments prepared for speaker clustering for each recording in RT09 evaluation set*

Recordings	Mean (seconds)	Std
EDI_20071128-1000	1.7344	1.2117
EDI_20071128-1500	1.4099	1.0913
IDI_20090128-1600	2.1509	1.1679
IDI_20090129-1000	1.5951	1.1424
NIST_20080201-1405	1.4938	1.0219
NIST_20080227-1501	1.8953	1.2008
NIST_20080307-0955	2.0522	1.1397

4. Speaker clustering

We propose a 2-step speaker clustering method. In the first step of initializing speaker clusters, we propose iterative random picking a small subset of speech segments and merging processes. In the second step, we propose a consensus-based pure speaker segment selection method for efficient speaker cluster modeling in cluster purification process. In the following sections, the BIC matrix used for cluster merging process and details of speaker clustering processes are explained.

4.1. BIC metric for cluster merging

Among multiple clusters, we use the Bayesian Information Criterion (BIC) [11] to select two clusters to merge. Suppose the two clusters of interest are *X* and *Y*. Let *Z=X U Y* to be the merged cluster. We use the following BIC metric to decide whether or not to merge the clusters X and Y.

$$T_{thres} = S_1 - S_0 - \frac{\lambda}{2}\log(F_x + F_y) \quad (1)$$

$$S_1 = \sum_{i=1}^{F_x} \log p(x_i \mid \theta_z) + \sum_{i=1}^{F_y} \log p(y_i \mid \theta_z) \quad (2)$$

$$S_0 = \sum_{i=1}^{N_x} \log p(x_i \mid \theta_x) + \sum_{i=1}^{N_y} \log p(y_i \mid \theta_y) \quad (3)$$

Where λ is penalty factor (set to be 1) and F is total number of frames in clusters *X* or *Y*. We use the likelihoods of Gaussian distributions to compute S_1 and S_2. If there are O clusters in total, a total of $O!/(2!(O-2)!)$ pairs are considered

for merging. Out of these pairs, the two clusters which have the highest BIC scores are merged.

4.2. Step 1: Initializing speaker clusters

We group the speech segments obtained in Section 3.3 to initialize the speaker clusters. We propose a random picking and merging method which includes two steps. In the first step we pick a small subset of speech segments randomly. In the second step, we merge the picked speech segments using the BIC metric discussed in Section 4.1. Before we start these processes, we define a total number of target clusters N. We assume N is larger than the number of speakers which will appear in the recordings. We note that very short speech segments have insufficient information on speaker identity. Hence, we use the long speech segments with 4.5 ~ 5 seconds in length to start the clusters.

We randomly pick a subset including R segments ($R > N$) from the group of long speech segments to start the initial speaker clustering process. Each of the R speech segments is treated as individual cluster. The 2 clusters out of R clusters are merged each time until the target cluster number of N is achieved. Then, the next incremental subset of r speech segments ($N + r = R$) is randomly picked and each speech segment is treated as individual cluster. Merging process is repeated until the target number of N clusters is achieved.

This random picking of speech segments and cluster merging process is repeated until there are no more speech segments which are longer than 2 seconds in each recording. The speech segments which are smaller than 2 seconds are left from the merging process as BIC metric is not reliable for the short speech segments as these segments contain insufficient discriminative information for speaker identity.

At the end of the above processes we obtain N speaker clusters for each recording. Finally, each of the segments which are shorter than 2 seconds are assigned to each of N clusters using GMM scores. We train Universal Background Model (UBM) on all the speech segments of the entire recording using EM algorithm [9]. Model adaptation is carried out for each of N clusters on UBM using MAP algorithm [9]. Each of short speech segments are scored against all N adapted cluster models and assigned to the cluster with the highest GMM score.

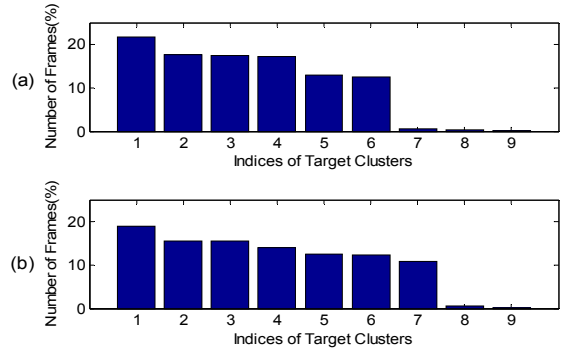

Figure 2: *Percentage of number of frames assigned to each cluster after initial speaker clustering (a) NIST_20080227-1501 and (b) NIST_20080307-0955*

We then check the number of speech frames assigned to each clusters as shown in Figure 2. From the figure, we found that some of the clusters are assigned very little frames and we discard these clusters from further processing. The clusters which have percentage of the total frames less than a threshold, T_i, are discarded. The speech segments from the discarded clusters are assigned to the clusters retained using GMM scores. The proposed approach does not need a stopping criterion for cluster merging process as in AHC method [3] as the target number of clusters to achieve is pre-defined.

4.3. Step 2: Purifying speaker clusters

In this step, we purify the clusters obtained in Section 4.2 by correcting the incorrect assignments. If we can remove impure speaker segments and select only pure speaker segments for speaker modeling, cluster purification will be more effective. To arrive at a reliable speaker clustering, we propose a consensus [7] based strategy that is inspired by the re-sampling methods in pattern recognition. Consensus based clustering method involves performing multiple clustering runs and made decisions globally over all clustering runs for cluster assignment [12]. In this way, we group speech segments that are together consistently across all runs to clusters based on consensus, and remove inconsistently assigned speech segments as they are deemed impure speaker segments. MFCC coefficients described in Section 3.2 are used as features.

Firstly, we perform multiple clustering runs as follows:

1. Train a root Gaussian Mixture Model (GMM), λ_{Root}, using the speech segments of entire meeting. EM algorithm [10] is used to train λ_{Root}. The mixtures of λ_{Root} has diagonal covariance matrices and the number of mixture, M=128.
2. Perform GMM adaptation from λ_{Root} for each cluster produced by Section 4.2 of initial clustering. Adaptation is performed on the weights, means and variances using MAP approach [10]. If there are q initial clusters, there will be $\lambda_1 \ldots\ldots \lambda_q$ adapted GMMs.
3. Score each speech segment of entire meeting against $\lambda_1 \ldots\ldots \lambda_q$ GMMs and assign each segment to the GMM with the highest likelihood score. Then, new set of q clusters are produced.
4. The new set of $\lambda_1 \ldots\ldots \lambda_q$ GMMs is re-adapted from λ_{Root} for q new clusters.
5. Iterate the steps 3 and 4 for 10 times.
6. Steps 1 to 5 is repeated for M=126, 124, 122,….,4.

The process of multiple clustering runs will produce 10 (iterations) x 63 (63 different M) clustering results.

Then, we generate q clusters using the results of 10 x 63 clustering runs. We looks for the speech segments which are clustered together in all 10 x 63 clustering runs and group them as a cluster. Then, the small clusters which have percentage of total frames less than a threshold, T_p, are discarded. The members of discarded clusters are taken as impure speaker segments. Finally, we obtain the clusters which include pure speaker segments. Repeat the steps 1 and 2 once to use clusters retained for speaker modeling. And, the steps 3 and 4 are iterated until the assignment of speech segments to individual cluster has been stabilized. A total of less than 10 iterations are necessary for stabilization. The number of mixture M is 128 in this process.

5. Experiments and results

We conduct experiments using the Rich Transcription 2009 (RT-09) Meeting Recognition SDM evaluation. The

experiments are conducted following the guidelines of RT-09 evaluation [13]. This evaluation consists of 7 meeting recordings as listed in the first column of Table 2.

We conduct experiments for speech/non-speech detection or Voice Activity Detection (VAD) as mentioned in Section 3.3. We set the experimental parameters to conduct speaker cluster initialization experiments as mentioned Section 4.2. We set target cluster number N to be *9*, the number of segment R for random picking to be 15, and the number of incremental segments r to be 6. We select the threshold, T_i, as 10% of frames of the entire meeting to discard the small clusters out of 9 clusters. Finally, we purify the clusters obtained from speaker initialization experiments as mentioned in Section 4.3. The threshold, T_p, for cluster purification is set as 5% of total speech frames of entire recording. We use RT-06 evaluation set as the development data to select the thresholds, T_i and T_p.

We report the DERs for speech/non-speech detection or Voice Activity Detection (VAD), initial clustering and cluster purification in the 2nd, 3rd, and 4th columns respectively in Table 2. DER is decomposed into 3 components [3]: missed speech, false alarmed speech and Speaker Error (SE). And, SE of our system is also included in the 5th column of Table 2. The reported results have taken into account overlapping speech.

Table 2. *Speaker Diarization Error (DER) and Speaker Error (SE) on SDM system of the RT-09 evaluation set (Scoring overlapped speech is accounted in the error rates.)*

Data*	DER(%)			SE(%)
	VAD	Initial clustering	Cluster purification	Cluster purification
1[4]	2.0	36.07[4]	15.02[4]	9.7
2[4]	4.2	48.1[4]	24.84[4]	12.2
3[4]	1.1	16.43[4]	6.65[4]	2.0
4[4]	5.2	30.65[4]	15.38[4]	5.8
5[5]	4.4	56.98[5]	36.45[5]	17.7
6[6]	0.6	42.22[6]	17[6]	8.0
7[11]	1.8	27.95[7]	10.9[7]	5.6
all	**2.7**	**34.4**	**16.41**	**7.8**

*1= EDI_20071128-1000, 2= EDI_20071128-1500,
3= IDI_20090128-1600, 4= IDI_20090129-1000
5= NIST_20080201-1405, 6= NIST_20080227-1501
7= NIST_20080307-0955
The number of speakers for each task is indicated in [ ]

Overall performance of proposed system is 16.4% (DER) or 7.8% (SE). We have evaluated the proposed techniques on NIST RT-09 database, which is a common platform for many RT systems in the literature. It is worth noting that the proposed random picking and merging methodology for initial clustering followed by cluster purification method outperforms other reported RT-09 SDM systems, such as 1) AHC method (19% SE) [3], 2) top-down clustering strategy (21.1% DER) [2] and 3) energy based segmentation followed by cluster purification method (17.34% DER) [5]. Table 2 shows the DER at each step of the system flow to appreciate the contribution of the individual modules. Instead of repeating the prior work, we only cite their results on the same database for comparison.

The advantages of our proposed method over AHC [3] is that the proposed system 1) starts the clustering process with clusters which include speech segments of possible single speaker identity 2) does not need stopping criterion and 3) provides a facility to remove impure speaker segments from speaker modeling. In addition, initial clustering strategy in our approach helps to estimate the number of speakers correctly for all the recordings except recording 7. The numbers of speakers estimated are presented in the third and the fourth columns (in brackets) of Table 2. After correctly estimating the number of speakers, cluster purification process further help to reduce the DER as consensus clustering method help to remove impure speaker segments when speaker modeling is carried out for each speaker cluster.

6. Conclusions

The proposed speaker diarization system is found to yield better performance on RT-09 SDM evaluation set in comparison with other reported RT-09 SDM systems [2, 3, 5]. The random picking and merging methodology for initial clustering is effective for cluster initialization and provides good estimation for number of speakers present in each recording. In addition, consensus based impure speaker segment removal process helps to obtain better speaker models and to achieve effective cluster purification process to improve DER.

7. References

[1] Ajmera, J. and Wooters, C., "A Robust Speaker Clustering Algorithm", IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp 411-416, 2003.

[2] Simon, B., Nicholas, W. D. E. and Corinne, F., "The LIA-EURECOM RT'09 Speaker Diarization System: Enhancements in Speaker Modelling and Cluster Purification", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.4958-4961, 2010.

[3] David, I. and Gerald, F., "Tuning-Robust Initialization Methods for Speaker Diarization", Accepted for publications in IEEE Transactions on Audio, Speech, and Language Processing.

[4] Chen, S. S. and Gopalakrishnan, P. S., "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, USA, February 1998.

[5] Sun, H. W., Ma. B., Khine, S. Z. K., and Li, H. Z., "Speaker Diarization System for RT07 and RT09 Meeting Room Audio", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.4982-4985, 2010.

[6] Han, K. J, Kim, S. and Narayanan, S. S., "Robust Speaker Clustering Strategies to Data Source Variation for Improved Speaker Diariation", Automatic Speech Recognition and Understanding (ASRU), pp. 262-267, Dec. 2007.

[7] Monti, S., Tamayo, P., Mesirov, J. P., and Golub, T. R., "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] Scalart, P. and Vieira Filho, J., "Speech Enhancement based on a Priori Signal to Noise Estimation", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.629-632, 1996.

[9] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[10] Wooters, C., and Huijbregts, M., "The ICSI RT07s Speaker Diarization System," Lecture Notes in Computer Science, vol.4625, pp. 509-519, 2008.

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

[12] Duda , R.O., Hart, P.E., and Stork, D.G., Pattern Classification, New York: John Wiley & Sons, 2001.

[13] "The 2009 (RT09) Rich Transcription Meeting Recognition Evaluation Plan", http://itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf