The NIST SRE Summed Channel Speaker Recognition System

Hanwu Sun and Bin Ma
Human Language Technology Department, Institute for Infocomm Research,
A*STAR, Singapore 138632
{hwsun, mabin}@i2r.a-star.edu.sg

Abstract

This paper presents an improved speaker recognition system for the summed channel evaluation tasks in the 2008 NIST SRE (SRE08) with multiple summed-channel excerpts for speaker training and one summed-channel excerpt for testing. The system includes three main modules in which a hybrid speaker purification and clustering algorithm is adopted to segregate the summed-channel speech, a common speaker identification is proposed by mapping multiple summed-channel excerpts for a common speaker cluster, and the GMM-SVM-NAP algorithm is used for the speaker recognition system. The system achieves an overall EER of 7.82% for all the trials and 4.19% for English trials in the SRE08 3summed-summed task.

Index Terms: speaker recognition, speaker diarization, summed channel

1. Introduction

In the most training conditions and test conditions with the telephone channel in the NIST Speaker Recognition Evaluations (SREs), the training/test segments are excerpts from two-channel telephone conversations while the target speaker is designed from one of the two channels [1]. The two-channel excerpt often refers to the four wires (4-wire) telephone recording. However, in many practical applications, the 4-wire recording is not always available, such as the conversations using a typical analogue telephone set at home or in office. In such case, the conversations in the two channels are usually summed to a single track. It is denoted as 2-wire or summed-channel recording.

From 2005 to 2010, the summed-channel speaker recognition was one of the evaluation tasks in the NIST SRE. In such a task, the speech signals from both sides of the conversation are summed together. The speech segments of the target speaker have to be distinguished from that of another speaker. Assuming the speakers take turns to speak most of the time, we can apply the multi-speaker segmentation or speaker diarization methods to segregate the voices by different speakers [2,3,4,5].

In this paper, we are interested in speaker recognition of the summed-channel tasks with multiple training conversations, so called 3summed-summed test in the NIST SRE 2008 [1]. The training data for each speaker consist of 3 summed channel conversational excerpts, each with approximately 5 minutes of speech, involving one common target speaker in these 3 training summed excerpts. The test data consist of 1 summed-channel conversational excerpt of approximately 5 minutes of speech. The challenge for the speaker model training is to identify the speech segments of the common speaker from these multiple summed excerpts.

We take advantage of the speaker diarization techniques developed for the 2007 and 2009 NIST Rich Transcription (RT) Meeting Recognition Evaluations (RT-07 and RT-09) [2, 3] for the summed-channel speaker segregation. Speaker diarization is a task to detect "who spoke when" in the meeting recordings. We adopt the purification process [3] in combination with the Viterbi decoding algorithm to cluster all the summed channel speech into two separate clusters.

After all the summed-channel excerpts are diarized into speaker dependent clusters, we introduce a two-stage cluster mapping method to extract the desired training speaker clusters by identifying the common speaker for training the common speaker model.

Finally, the speaker recognition system is based on the GMM-SVM-NAP modeling technique [6]. Figure 1 shows the speaker recognition system based on speaker diarization, common speaker modeling, and speaker recognition.

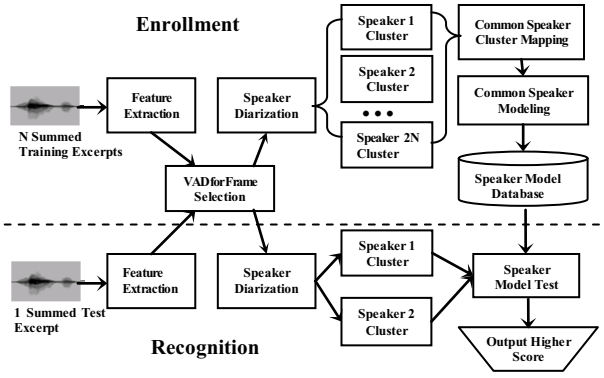

Figure 1: Diagram of speaker recognition system for the NIST SRE summed training and summed testing tasks.

The paper is organized as follows. In Section 2, the GMM-SVM-NAP speaker recognition system is presented. The hybrid speaker diarization process is described in Section 3. The common speaker modeling for the desired speaker is introduced in Section 4. The experimental results are reported in Section 5. Finally, we conclude in Section 6.

2. GMM-SVM-NAP Speaker Recognition

The speaker recognition system using in this study is based on the GMM based support vector machine (GMM-SVM) and the *nuisance attribute projection* (NAP) technique for channel compensation, in short, the GMM-SVM-NAP as reported in 6].

We use the MFCC feature in this study. In particular, a 16-dimension MFCC features are generated for each speech frame with a window of 30ms and a frame shift of 12.5ms. By including the 16-dimension of the first derivatives and the 14-dimension of the second derivatives, a MFCC feature vector consists of 46 dimensional features.

The spectral subtraction technique is used to assist the voice activity detection (VAD) for selecting useful speech frames [7]. The MFCC feature vectors are then processed by RASTA filtering [8] and followed by *mean and variance* normalization (MVN).

In the GMM-SVM-NAP speaker recognition, each of the utterances in variety of durations will be represented by a high-dimensional vectors referred to as the GMM supervector. Channel compensation and speaker detection are then performed in the high-dimensional vector space.

Let $\Lambda = \{ \omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i; i = 1, 2, \ldots M \}$ be the parameters of the universal background model (UBM), where M is the number of mixture components, ω_i are the mixture weights, $\mathbf{m}_i$ are the mean vectors, and $\boldsymbol{\Sigma}_i$ are the covariance matrices assumed to be diagonal. For a given utterance X_s, the Baum-Welch statistics are used to adapt the mean vectors of the UBM using the maximum *a posterior* (MAP) [9]. The adapted mean vectors are concatenated to form a GMM supervector, as follows:

$$\mathbf{m}(s) \equiv \left[\mathbf{m}_1^{\mathrm{T}}(s), \mathbf{m}_2^{\mathrm{T}}(s), \ldots, \mathbf{m}_M^{\mathrm{T}}(s) \right]^{\mathrm{T}}, \quad (1)$$

where T denotes transposition. The mean vectors are then normalized by its standard deviation and weighted by the squared root of the mixture weights:

$$\mathbf{m}'_i(s) = \sqrt{\omega_i}\, \boldsymbol{\Sigma}_i^{-1/2} \mathbf{m}_i(s), \; i = 1, 2, \ldots, M. \quad (2)$$

The normalization in (2) allows the similarity between two GMM supervectors to be computed by taking their inner product.

Based on the NAP matrix derived from the labeled or clustered NAP training dataset, the NAP projection compensated GMM supervector is given as [6]:

$$\hat{m} = \left(I - EE' \right) \cdot m' \quad (3)$$

Here the E is the eigenvectors of the NAP matrix.

With GMM-SVM-NAP, one SVM is trained for each target speaker using the NAP related supervector. Let Ω_k be the target speaker utterance (i.e., the positive example), and $\Re$ the set of supervectors pertaining to background speakers. An SVM solver for the dual formulation [6],

$$f_k \leftarrow \mathrm{SVM}\left(\Omega_k \,\|\, \Re \right), \quad (4)$$

with the Lagrange multipliers α associated to all the supervectors in the training set and a bias parameter β, which essentially forms a linear model f_k for a target speaker, as follows:

$$f_k(\mathbf{m}') = \left[\sum_{\Omega_k} \alpha_i \mathbf{m}'(i) - \sum_{\Re} \alpha_j \mathbf{m}'(j) \right]^{\mathrm{T}} \mathbf{m}' + \beta. \quad (5)$$

It should be noted that, the background set $\Re$ is used for all target speakers enrolled to the system. In this study, the SRE04 corpus was used as the background speaker data set for the SVM training and also used to train the gender-dependent UBMs with 1024 mixture components, as well as NAP training dataset. The rank of NAP is set to be 60 in the experiments. TZnorm was used for score normalization [10], where SRE05 data was selected for training the cohort models for Tnorm and SRE04 data was used as imposture utterances for Znorm.

3. Speaker Diarization

In the speaker diarization, the initial speaker clusters might affect the subsequent speaker merging and clustering. We adopt a hybrid speaker diarization strategy for the summed channel audio segmentation to improve the initial clusters. This is inspired by the findings in RT-07 an RT-09 speaker diarization evaluations [2,3]. The strategy consists of a GMM based progressive purification process and a Viterbi decoding process for clustering. Since there are only two speakers in each summed channel in most case, we apply the BIC criterion [2, 11] directly to merge the similar speech segments until two clusters are left. All the summed-channel excerpts for both training and test are first diarized into two speaker dependent clusters. We summarized this hybrid clustering method in Algorithm 1.

Algorithm 1: Speaker Diarization Algorithm.

Step 1. Identify the speech and non-speech frames using a energy based voice activity detection algorithm [3,7].

Step 2. Extract a MFCC feature vector for each of the speech frames from the summed channel. Unlike in speaker recognition, feature normalization is not applied inn speaker diarization.

Step 3. Divide the speech frames into segments of 2 second in length and uniformly group them into Q initial clusters.

Step 4. Perform the initial cluster purification via EM and MAP adaptation [9] as follows:

- 4a. Train a Root GMM, λ_{Root}, with 2 mixture components using all the clusters;
- 4b. Train all the cluster-dependent GMMs, $\lambda_1, \lambda_2, \ldots, \lambda_Q$, by adapting the Root GMM, λ_{Root}, via MAP,
- 4c. Evaluate all the segments against the cluster-dependent GMMs, $\lambda_1, \lambda_2, \ldots, \lambda_Q$, and relocate the segments into the GMMs, accordingly;
- 4d. Repeat the steps 4b and 4c until no segment changes is found;
- 4e. Increase the size of GMM model by 2 and repeat step 4a) until GMM model size is equal to 16.

Step 5. Based on the initial purification, we apply Viterbi Decoding, MAP adaptation and BIC approaches to re-segment the recordings, and purify and merge the clusters.

- 5a. Train the Root GMM, λ_{Root}, with 10 mixture components using all the clusters;
- 5b. Retrain the cluster GMMs ($\lambda_1, \lambda_2, \ldots, \lambda_Q$) by adapting from the Root GMM, λ_{Root};
- 5c. Conduct Viterbi decoding to re-segment the recordings;
- 5d. Repeat steps 5b ~ 5c for several times for segment convergence;
- 5e. Compute the BIC score for each pair of the clusters;
- 5f. Find the pair with the largest BIC score and merge the pair of clusters;
- 5g. Repeat step 5a ~ 5f until the number of clusters Q is reduced to 2.

4. Common Speaker Modeling

In the NIST SRE summed channel training and test evaluation task, there are multiple summed-channel excerpts, in which a common desired speaker exists. For the 2008 NIST SRE 3summed-summed task, there are 3 summed-channel excerpts for training the speaker model.

After the speaker segregation process described in Section 3, each summed-channel excerpt is separated into two speaker dependent clusters. Suppose there are *N* summed-channel excerpts for training, we will obtain 2*N* speaker clusters. From these 2*N* speaker clusters, *N* clusters have to be selected for training the common speaker model and no any two selected clusters are from the same excerpt. Obviously, it is crucial to identify the correct *N* common speaker clusters, which contain a desired speaker's speech, to train the target speaker model. We propose a two-stage processes to conduct this task. The cluster BIC scores among these 2*N* speaker clusters are first used to find *N* common clusters as the desired speaker clusters and the remained *N* clusters as the undesired speaker clusters. Then, we build two gender models for male and female to further check whether the selected *N* common clusters are consistent with the labeled gender provided by NIST.

4.1. Common Speaker Clusters Selection

For *N* summed channel training condition, 2*N* speaker clusters from the *N* summed-channel training excerpts will be generated based on the results of speaker segregation.

We conduct the common speaker clusters selection from 2*N* segregated clusters in the following steps.

a. Define $C_{i,m} \quad i=1,...,N, \quad m=1,2$ as the 2*N* segregated clusters.
b. Calculate the cross-cluster BIC [2, 11] scores:
$$B_{i,j}(m,n) = BIC(C_{i,m}, C_{j,n}) \quad i \neq j \; i,j=1,...,N, \quad m,n=1,2$$
c. Calculate the overall BIC scores for all the cross-cluster groups with 2*N* clusters.
d. Denote the cross-cluster group with the highest score as *P0* (containing *N* clusters) and the remaining *N* clusters as *P1*.

In certain cases, there are some special combinations for the *N* summed training utterances. For example, there are only two speakers involved in all the *N* summed excerpts. When one speaker is male and another speaker is female, we can make use of the gender information provided by NIST for deciding which speaker cluster is the desired one, although it is hard to make the decision based on the scores of the cross-cluster groups.

4.2. Gender Mapping

Based on the above common cluster group selection, we get the desired *P0*'s *N* clusters and the remained *P1*'s *N* clusters. We use the NIST SRE04 training data set to train two gender dependent models and test the *P0* and *P1* group features against these two models. Male and Female were modeled by two separate 64 mixture GMMs via EM algorithm [9], λ_M and λ_F. A gender decision can be made for the *P0* and *P1* cluster features as.

$$p(\lambda_M \mid Pi) \geq p(\lambda_F \mid Pi) \rightarrow Male \quad i=0 \text{ or } 1 \qquad (6)$$

$$p(\lambda_M \mid Pi) < p(\lambda_F \mid Pi) \rightarrow Female \quad i=0 \text{ or } 1 \qquad (7)$$

We can further correct the common speaker group using the gender information as:

$$\begin{cases} \text{if } P0_{gender} \neq Label_{gender} \text{ and } P1_{gender} = Label_{gender} & \Rightarrow P1 \\ Otherwise & \Rightarrow P0 \end{cases} \qquad (8)$$

Here *Label* means the gender label provided by NIST. Obviously, we choose *P1*'s *N* clusters as the desired common speaker clusters only under such condition: the *P0*'s gender is mismatched with the label gender while *P1* gender is the same with the label gender. Otherwise, *P0* is our common speaker group.

5. Speaker Recognition Experiment

The summed-channel speaker recognition experiments were conducted on the 3summed-summed subtask of the 2008 NIST SRE (SRE08) [1] based on the GMM-SVM-NAP algorithm with MFCC features. We report the speaker recognition performance by both the equal error rate (EER) and the detection cost function (DCF) defined in the SRE08 evaluation plan [1].

5.1. Summed Channel Training and Testing

Since NIST provided all the automatic speech recognition (ASR) transcripts for the evaluation data in the speaker recognition. Similar to what is described in [5], we are able to recover about 50% 2-wire transcripts from the corresponding 4-wire ASR files in the SRE08 summed channel test set. These 4-wire ASR transcripts provide the speakers' voice activity information. We have used these recordings as the development data set to evaluate the speaker diarization performance on SRE08 test set.

Based on the development data set, we achieved 13.12% diarization error rate (DER) with overlapping speakers [12] and 4.95% DER without overlapping speakers.

By applying the speaker diarization, each summed channel recording was clustered into two separate speaker clusters, each containing the speech of a single unknown speaker. Since the designated speaker is unknown, both of the two speaker clusters evaluated against the target speaker model, and we selected the higher speaker recognition score as the matching result.

5.2. Experiment Results

To evaluate how we benefit from the speaker diarization process, we first conducted speaker recognition experiments with the summed speech data without separating the speakers as baseline. Then, we applied the speaker diarization to the training and test excerpts to verify the performance under different training and test combinations. The experimental results for SRE08 3summed-summed tasks are shown in Table 1.

In Table 1, it can be seen that the speaker diarization process significantly improves the summed-channel speaker recognition performance in terms of both EER and minimum DCF. We have a baseline without diarization 17.85% EER for all trials and 17.65% EER for English trials, respectively. After applying the speaker diarization to training and test excerpts and using the gender detection for cluster mapping, both EER and DCF are reduced significantly, finally giving

7.82% EER for all trials and 4.19% EER for English trials, representing a 56.19% (all trial) and 76.26% (English trials) relative improvement in EER, and 43.46% (all trial) and 64.32% (English trials) relative improvement in minimum DCF. Meanwhile, we also observe that further benefit of both the EER and minimum DCF improvement is achieved by applying the gender detection to the training cluster selection.

Table 1: EER and min DCF for the SRE08 3summed-summed subtasks before and after diarization.

Train/Test Condition	All Trials		English Trials Only	
	EER%	DCFx100	EER%	DCFx100
Summed / Summed	**17.85**	**8.56**	**17.65**	**7.09**
Summed / Seg.	**14.41**	**7.74**	**12.18**	**6.25**
Seg. / Summed	**11.79**	**7.15**	**9.18**	**4.29**
Seg. / Seg.	**8.64**	**5.22**	**5.24**	**3.22**
Seg. / Seg. With Gender Detection	**7.82**	**4.84**	**4.19**	**2.53**

We also summarize the Detection Error Tradeoff (DET) curves of SRE08 3summed-summed all trials and English trials in Figure 2 and 3, respectively, where minimum DCF is marked with red cycle.

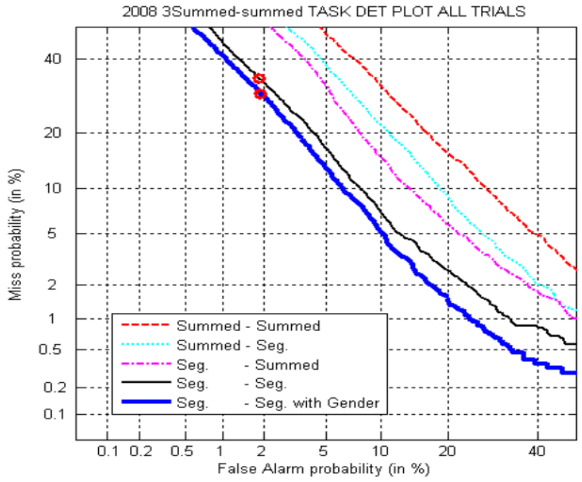

Figure 2: SRE08 3summed-summed channel subtask DET curves with and without diarization (all trials).

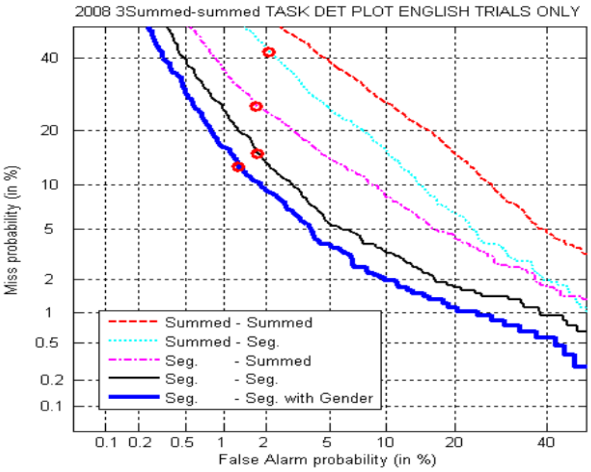

Figure 3: SRE08 3summed-summed channel subtask DET curves with and without diarization (English trials).

6. Conclusions

This paper presents an improved speaker recognition system for the NIST SRE 2008 summed-channel training and testing task. The hybrid speaker diarization was adopted to segregate the speech in the single channel by speakers. A mapping method was proposed to select the common desired speaker clusters for the model training. The speaker diarization and proposed cluster selection method have reduced the speaker recognition EER by 56.19% for all trails and 76.26% for English trials on the NIST SRE 2008 3summed-summed task, respectively. The experiments also suggest that there is an obvious benefit by applying the gender detection to the training cluster selection. The experiments show that the proposed method performs consistently in both all trials and English trials summed channel tasks. Moving forward, we would like to study the effects of summed channel overlapped or double talking sections on speaker recognition system.

7. References

[1] NIST 2008 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf.

[2] T.L. Ma, H. Sun, B. Ma and H. Li, "Speaker Clustering and Cluster Purification Methods for RT07 and RT09 Evaluation Meeting Data", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 2, pp.461-473, 2012.

[3] H., Sun, B. Ma, C. Huang, T. Nguyen and H. LI, "The IIR NIST SRE 2008 and 2010 Summed Channel Speaker Recognition System", in *Proc. Interspeech 2010*, pp. 366-369, Makuhari, Japan 2010.

[4] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Loquendo - Politecnico di Torino's 2006 NIST Speaker. Recognition Evaluation System," in *Proc. Interspeech*, pp.1238–1241, Belgium, 2007.

[5] D. Reynolds, P. Kenny and F. Castaldo, "A study of new approaches to speaker diarization," in *Proc. Interspeech*, pp. 6–10, Brighton, 2009.

[6] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, pp. 97–100, 2006.

[7] H. Sun, B. Ma and H. Li, "An Efficient Feature Selection Method for Speaker Recognition," in *Proc. ISCSLP*, pp. 181–184, 2008.

[8] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, 10(1):19-41, 2000.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10, no 1-3, pp. 42–54, Jan 2000.

[11] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," In Proc. NIST MLMI Meeting Recognition Workshop, Edinburgh, 2005.

[12] "Spring 2007 (RT-07) Rich Transcription meeting recognition evaluation plan," http://www.nist.gov/speech/tests/rt/rt2007/docs/rt07-meeting-eval-plan-v2.pdf.