The IIR NIST SRE 2008 and 2010 Summed Channel Speaker Recognition Systems

Hanwu Sun, Bin Ma, Chien-Lin Huang, Trung Hieu Nguyen, Haizhou Li

Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore 138632

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

Abstract

This paper reports the IIR speaker recognition system for the summed channel evaluation tasks in the NIST SRE 2008 and 2010. The system includes three main modules: voice activity detection, speaker diarization and speaker recognition. The front-end process employs a voice activity detection algorithm for effective speech frame selection. The speaker diarization system that was developed for 2007 and 2009 NIST RT Evaluations is adopted for summed channel speech segmentation. A hybrid purifying and clustering algorithm is developed to segregate the summed channel speech by speakers. The GMM-SVM speaker recognition system is adopted to evaluate the performance with both MFCC and LPCC features. The system achieves an overall EER of 3.46% in the 1conv-summed task and 1.87% in the 8conv-summed task, respectively, where only all English trials are involved.

Index Terms: speaker recognition, speaker diarization, summed channel

1. Introduction

In the core test of the NIST Speaker Recognition Evaluations (SREs), each of the telephone trials contains one two-channel telephone conversational excerpt with the target speaker designated from one of the channels [1, 2]. The two-channel excerpt often refers to the four wires (4-wire) telephone recording. However, the 4-wire recording is not always available in many practical application scenarios. With a typical analogue telephone set at home or in office, the conversations in the two channels are usually summed to a single track. It is denoted as 2-wire or summed-channel recording.

Since 2005, the summed-channel speaker recognition has been one of the evaluation tasks in the NIST SREs, especially in 2008 and 2010 [1, 2]. In such a task, two voices from both side of the conversation are summed together. The challenge is to distinguish the voice of the intended target speaker 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 [3,4,5,6].

In this paper, we are interested in speaker recognition of two of the summed-channel tasks, 1conv-summed and 8conv-summed in the NIST SRE 2008 and 2010. The training data, 1conv and 8conv, consist of one and eight conversational excerpts of approximately 5 minutes of speech each excerpt, only involving the target speaker on the designated side, while other speaker voice is recorded on another side or channel; while the test data, summed, consist of one summed-channel telephone conversational excerpt of approximately 5 minutes of speech in a single summed channel.

We take advantage of our speaker diarization techniques developed in the 2007 and 2009 NIST Rich Transcription (RT) Meeting Recognition Evaluation (RT-07 and RT-09) [3] for summed-channel speech segregation. In this way, one can consider the proposed speaker recognition system to have a two-steps process - speaker diarization followed by speaker recognition, as illustrated in Figure 1.

Speaker diarization is a task to detect "who spoke when" in the meeting recordings. We first use a spectral subtraction based voice activity detection (VAD) [7] to remove the non-speech frames. In speaker diarization, we employ an effective purification process [3, 4] in combination with the Viterbi decoding algorithm to cluster the summed channel speech data into two separate channels. The speaker recognition is based on the GMM-SVM modeling technique [8]. While there are both English and non-English trials [1] in NIST SRE 2008 1conv- and 8conv-summed tasks, only English trials exist in the NIST SRE 2010 1conv- and 8conv-summed tasks [2]. We only report results of English trials in this paper.

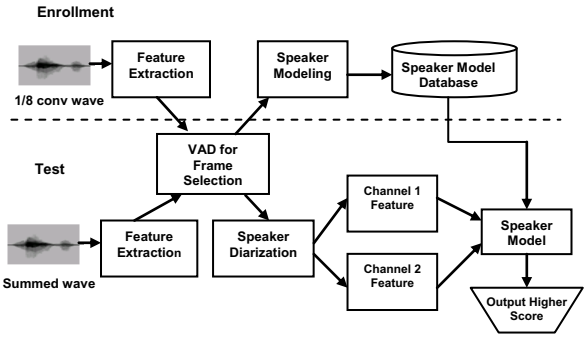

Figure 1. Diagram of speaker recognition system for NIST SRE 2008 and 2010.

The paper is organized as follows. In Section 2 we give an overview of the voice activity detection for the feature frame selection. The hybrid speaker diarization is introduced in Section 3. The experimental results are reported in Section 4. Finally, we conclude in Section 5.

2. Voice Activity Detection for Speech Frame Selection

We study a spectral subtraction process [9, 10] for noise reduction to assist the VAD process. We choose the spectral subtraction technique for its low computational cost and implementation complexity. The basic idea of the spectral subtraction is to suppress the additive noise in the corrupted speech signals. The estimate of the original and clean signal spectrum is obtained by subtracting an estimate of the noise

power (or magnitude) spectrum from the noisy signal. Detailed descriptions can be found in [7]. We apply an over-subtraction factor 1.2 in the noise reduction to facilitate the energy based speech frame selection, and a subtraction factor 1.0 for full noise subtraction.

A speech frame *i* that meets the following energy criteria is selected for further processing.

$$\begin{aligned} 1 &\Rightarrow \quad P(i) > \max(P(j)_{j=1,2,\ldots,N}) - T_{SNR} \\ 2 &\Rightarrow \quad P(i) > T_{\min} \end{aligned} \tag{1}$$

where $P(i)$ is the power in decibel scale (dB) or standard deviation of the *i*-th frame after spectral subtraction, max(.) is an operator to find the maximum power level in all the frames in the current utterance, T_{SNR} is the cutoff SNR threshold from the maximum power level, and the $T_{\min}$ is the minimum power threshold required for the frames. The full scale input voice signal is in the range [-1, 1]. The parameter T_{SNR} is set to be 48 dB and $T_{\min}$ to be -75 dB in this study. Under this frame selection scheme, the higher the value of T_{SNR} is, the more feature frames are selected. The spectral subtracted speech signal is used just for frame selection. The acoustic features for speaker recognition and speaker diarization are still derived from the original speech signals.

3. Speaker Diarization

During speaker diarization, the initial speaker purification might affect the subsequent speaker merging and clustering. We propose a hybrid speaker diarization strategy for the summed channel audio segmentation to improve the initial clusters. This is inspired by findings in RT-07 an RT-09 speaker diarization evaluations [3, 4]. The strategy consists of a GMM based progressive purification process and a Viterbi decoding process for clustering. Since we have already known that there are only two speakers in each summed channel, we can apply BIC criterion [4, 11] directly to merge the similar speech segments until two clusters are left. We summarized the proposed hybrid clustering method in Algorithm 1.

Algorithm 1: Speaker Diarization Algorithm.

Step 1. Identify the speech and non-speech frames using the voice activity detection algorithm described in Section 2.

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

Step 3. Divide the speech frames into segments of 2 second in length and uniformly group them into 15 initial clusters.

Step 4. Perform the initial cluster purification via EM and MAP adaptation [12] as follows:
- 4a. Train a Root GMM with 2 mixture components using all the clusters;
- 4b. Train all the cluster-dependent GMMs by adapting the Root GMM based on MAP;
- 4c. Evaluate all the segments against the cluster-dependent GMMs 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 with 10 mixture components using all the clusters;
- 5b. Retrain the cluster GMMs by adapting from the Root GMM;
- 5c. Conduct Viterbi decoding to re-segment the recordings;
- 5d. Repeat steps 5a ~ 5d for several times until segmentation 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. Retrain step 5a ~ 5f until the number of clusters is reduced to 2.

4. Speaker Recognition

The summed channel speaker recognition experiments were conducted on the 1conv-summed and 8conv-summed subtasks of the NIST SRE 2008 (SRE08) and NIST SRE 2010 (SRE10). We implemented the GMM-SVM speaker models with MFCC and LPCC features in the experiments.

4.1. GMM-SVM Speaker Recognition

Two acoustic features MFCC and LPCC were used. A 16-dimenison MFCC features were 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. 46-dimension LPCCs were generated in the same way. The VAD described in Section 2 was used to select the speech frame and remove non-speech frames. The selected feature vectors were processed by RASTA filtering [14] and cepstral mean and variance normalizations (MVN).

We built two classifiers GMM-SVM-MFCC and GMM-SVM-LPCC using two different features [8]. A gender-independent universal background model (UBM) with 1024 Gaussian mixture components was first built, and the speaker GMM models were adapted from the UBM via a MAP algorithm [12]. We formed a GMM supervector for a conversation which is normalized by its standard deviation and weighted by the squared root of the weights of the Gaussian mixtures. The SVMTorch [15] is used to train the SVM model. The channel normalization was conducted using Nuisance Attribute Projection (NAP) [8] to project out the nuisance subspace from the original supervector space. The rank of NAP is set to be 60.

The NIST SRE 2004 corpus was used as the background training data set for UBM as well as the background speaker

set for SVM training. At the same time, the NIST SRE 2004 corpus is also used to derive the NAP matrix.

For the speaker recognition, a variety of score normalization approaches [16] have been proposed for a robust decision. We compared the Tnorm, Znorm, TZnorm and ZTnorm, and found that the TZnorm score normalization gave an overall better performance than others in this GMM-SVM speaker system. Hence, we only report the experimental results based on the TZnorm scores normalization. In the experiment, the NIST SRE 2005 1-side training data were used for training cohort models in Tnorm and the NIST SRE 2004 data were used as imposter speech utterances in Znorm.

4.2. System Fusion

The speaker recognition system is a score fusion of the two GMM-SVM classifiers, GMM-SVM-MFCC and GMM-SVM-LPCC:

$$s_i = \sum_{f=1}^{N} w_f s(f,i) + b \tag{2}$$

where N=2 is the number of classifiers and $s(f,i)$ is the score of either GMM-SVM-MFCC or GMM-SVM-LPCC of the i-th trial. The fusion parameters consist of the classifier specific weights and a global bias b. We used FoCal toolkit [13] to optimize the detection cost and find the fusion parameters. The detection cost function (DCF) is a weighed sum of miss detection and false alarm rates defined in the NIST SRE evaluation plans [1, 2], and is given as follows:

$$\begin{aligned} DCF &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &+ C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target}) \end{aligned} \tag{3}$$

where $C_{\text{Miss}} = 10$, $P_{\text{Target}} = 1$ and $C_{\text{FalseAlarm}} = 0.01$. We report the speaker recognition performance by both the equal error rate (EER) and *DCF*.

4.3. Summed Channel Test

Since NIST provided all the automatic speech recognition (ASR) transcripts for the evaluation data in the speaker recognition. Similar to what is in [6], 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 12.51% diarization error rate (DER) with overlapping speakers [17] and 4.75% DER without overlapping speakers. The DER is computed as follows [17].

$$DER = \frac{SE + MS + FA}{SPK} \times 100(\%) \tag{4}$$

where the speaker error time (SE) is the total time that is attributed to the wrong speaker, the missed speaker time (MS) is the total time in which less speakers are detected than what is correct, the false alarm speaker time (FA) is the total time in which more speakers are detected than what is correct, and scored speaker time (SPK) is the sum of every speaker's utterance time as indicated in the 4-wire ASR file.

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

4.4. Experiment Results

To appreciate how we benefit from the speaker diarization process, we first conducted speaker recognition experiments with the summed speech data without separating the speakers. Then, we applied the speaker diarization before the speaker recognition to verify the performance. The experimental results for SRE08 1conv-summed and 8conv-summed tasks are shown in Table 1 and Table 2, respectively.

In Table 1, it can be seen that the speaker diarization method significantly improves the summed channel speaker recognition performance in terms of both EER and minimum DCF. Fusing the GMM-SVM-MFCC and GMM-SVM-LPCC classifiers, giving 4.45% and 4.51% EER respectively, we obtained an EER of 3.46%, with 22% relative improvement. This suggests that MFCC and LPCC features are complementary. With the help of speaker diarization, we achieved an EER of 3.46% and a minimum DCF of 1.56%, representing a 48.2% relative improvement in EER, and 43.1% relative improvement in minimum DCF.

Table 1. EER and min DCF for the SRE08 1conv-summed subtasks (All English Trials) before and after diarization.

Seg.	Feature	Male		Female		All	
		EER %	DCF x100	EER %	DCF x100	EER %	DCF x100
Before	MFCC	6.98	3.01	7.67	2.94	7.47	2.98
	LPCC	6.84	3.15	7.79	3.10	7.50	3.18
	Fusion	**6.04**	**2.76**	**6.97**	**2.69**	**6.68**	**2.72**
After	MFCC	4.01	2.13	4.63	2.25	4.45	2.13
	LPCC	3.57	2.02	4.81	2.12	4.51	2.22
	Fusion	**3.32**	**1.55**	**3.70**	**1.58**	**3.46**	**1.56**

Table 2. EER and min DCF for the SRE08 8conv-summed subtasks (All English Trials) before and after diarization.

Seg.	Feature	Male		Female		All	
		EER %	DCF x100	EER %	DCF x100	EER %	DCF x100
Before	MFCC	4.47	2.02	4.25	1.77	4.22	1.91
	LPCC	4.32	2.10	4.25	1.83	4.21	2.00
	Fusion	**3.83**	**1.91**	**4.16**	**1.61**	**4.08**	**1.77**
After	MFCC	2.58	1.23	1.70	0.77	2.04	1.03
	LPCC	2.44	1.34	1.77	0.79	2.11	1.04
	Fusion	**2.00**	**0.63**	**1.50**	**0.65**	**1.87**	**0.65**

In Table 2, we report the performance on the SRE08 8conv-summed task. We observe a high EER reduction from 4.08% to 1.87% by applying speaker diarization.

We also summarize the Detection Error Tradeoff (DET) curves of SRE08 1conv-summed and 8conv-summed tasks in Figure 2 and Figure 3, respectively, where minimum DCF is marked with red cycle.

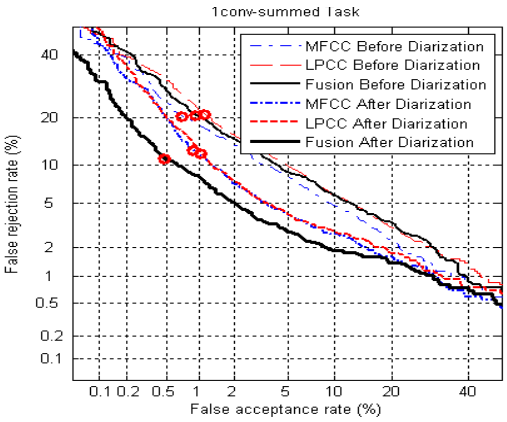

Figure 2. SRE08 1conv-summed channel subtask DET curves with and without diarization (All English Trials).

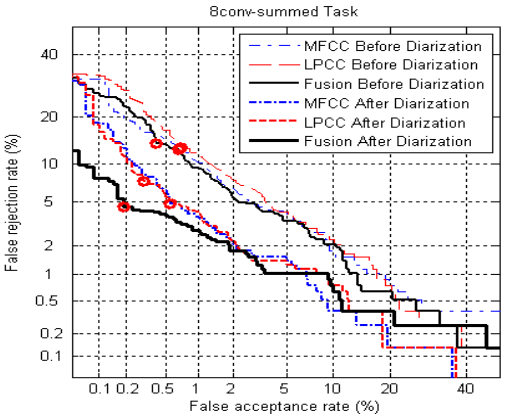

Figure 3. SRE08 8conv-summed channel subtask DET curves with and without diarization (All English Trials).

We applied the same technique in SRE10 evaluation task. Table 3 and Figure 4 summarize the fusion results and DET curves of SRE10. We only report SRE10 results with speaker diarization pre-processing, as submitted in NIST SRE 2010 by IIR site.

Table 3. EER and min DCF for the SRE10 1conv-summed and 8conv-summed subtasks after diarization.

After Segmentation and fusion	Male		Female		All	
	EER %	DCF x100	EER %	DCF x100	EER %	DCF x100
1conv-summed	3.96	1.60	4.75	2.06	4.26	1.96
8conv-summed	1.69	0.79	2.45	0.83	2.36	0.87

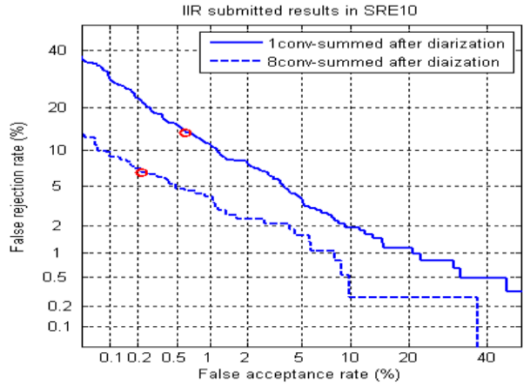

Figure 4. DET curves of IIR submission in SRE10 for 1conv-summed and 8conv-summed subtask.

5. Conclusions

This paper presented the speaker recognition system for the NIST SRE 2008 and 2010 summed-channel tasks. The hybrid speaker diarization was proposed to segregate the speech in the single channel by speakers. The speaker diarization has reduced the speaker recognition EER by 48.2% and 54.2% on the NIST SRE 2008 1conv-summed and 8conv-summed tasks, respectively. The experiments also suggest that there is an obvious benefit fusing two GMM-SVM speaker recognition classifiers based on MFCC and LPCC features. The experiments show that the proposed method performs consistently in both SRE08 and SRE10 summed channel tasks. Moving forward, we would like to test out the system on NIST speaker recognition tasks in which both the training data and test data are recorded in a summed channel.

6. References

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

[2] NIST 2010 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig//tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.

[3] H. Sun, B. Ma, Z. Swe. and H. Li., "Speaker Diarization System for FT07 and RT09 Meeting Room Audio," in *Proc. ICASSP*, pp.4982–4985, 2010.

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

[5] 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.

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

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

[8] 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.

[9] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE. Trans. Acoustics, Speech, Signal Processing*, vol. 27, pp. 113–120, 1979.

[10] R. Martin "Spectral Subtraction Based on Minimum Statistics," in *Proc. EUSPICO*, vol. 2, pp.1182–1185, 1994.

[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] 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.

[13] N. Brummer, Focal toobox (online) http://niko.brummer.googlepages.com/focal.

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

[15] R. Collobert and S. Bengio, "SVMTorch: support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143-160, 2001.

[16] 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.

[17] "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.