Study of Overlapped Speech Detection for NIST SRE Summed Channel Speaker Recognition

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 studies the overlapped speech detection for improving the performance of the summed channel speaker recognition system in NIST Speaker Recognition Evaluation (SRE). The speaker recognition system includes four main modules: voice activity detection, speaker diarization, overlapped speaker detection and speaker recognition. We adopt a GMM based overlapped speaker detection system, by using entropy, MFCC and LPC features, to remove the overlapped segments in summed channel test condition. With the overlapped speech detection, the speaker diarization achieves a relative 18% diarization error rate reduction for the 2008 NIST SRE summed channel test set, and we obtain relative equal error rate reductions of 13.3% and 9.4% in speaker recognition on the 1conv-summed task and 8conv-summed task, respectively.

Index Terms: speaker recognition, speaker diarization, summed channel, overlapped speech.

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]. 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.

In the earlier studies [2, 3, 4], the SRE performance of summed channel condition is much worse than that of single channel condition. Even with the speaker diarization process which detects "who spoke when", the SRE performance of summed channel condition is still significantly lower. One of the major problems of current speaker diarization systems is their inability to deal with overlapped speech [2, 3, 4].

In recent years, several techniques have been proposed to reduce the overlapped speech errors in the speaker diarization system. Boakye [5] presented an overlapped speech detection system which uses a Hidden Markov Model (HMM) based detector to identify the overlapped speech segments. Several studies employ Entropy [5, 6], MFCC, and Linear Predictive Coding (LPC) features [5] to detect and extract usable speech from overlapped speech segments.

For NIST SRE evaluation, several papers [2, 3, 4] have been reports on the work how to use the speaker diarization method to improve the summed channel speaker recognition performance. In our previous paper [4], the experimental results show that the diarization method can significantly improve the summed channel speaker recognition performance. However, we did not handle overlapped speech problem while the overlapped speech segments were assigned to one of the two speakers in the conversation. Obviously, it introduced the imposter speech in the summed-condition speaker recognition. It is interesting and important to study how overlapped speech affects on both diarization error rates (DER) [7] and speaker recognition rates for NIST SRE summed channel telephony dataset.

In this paper, we make a study of overlapped speech detection to improve NIST SRE summed channel speaker recognition performance. We present an overlap speech detection system to remove overlapped speaker features for summed channel speaker recognition, and study how overlapped speech affects on both diarization error rate (DER) [7] and speaker recognition rate. The overlap speech detector is based on GMM modeling by using Entropy, MFCC and LPC features.

Similar to the previous paper [4], we are interested in speaker recognition of two of the summed-channel subtasks, 1conv-summed and 8conv-summed in the 2008 NIST SRE (SRE08) [1]. 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 the test data, summed, consist of one summed-channel telephone conversational excerpt of approximately 5 minutes of speech in a single summed channel.

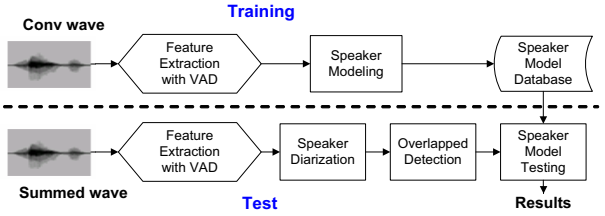

Figure 1. Diagram of speaker recognition system for NIST SRE on summed channel condition

Compared with the previous reported system [4], the modified summed channel speaker recognition system contains an extra overlapped speech detection section. Overall, we first use a spectral subtraction based voice activity detection (VAD) [4] to remove the non-speech frames. In speaker diarization, we employ an effective purification process [4, 8, 9] in combination with the Viterbi decoding algorithm to cluster the summed channel speech data into two separate channels. The GMM based overlapped speech detector has been applied to the diarized speech signal and the detected overlapped speech segments are removed from the speaker recognition. The speaker recognition is based on the GMM-SVM modeling technique [10]. The experiments are conducted on the English trials. Fig. 1 shows the flowchart of the system.

The paper is organized as follows. In Section 2 we describe the VAD for the speech/non-speech detection. The acoustic features used in the paper are presented in Section 3. The speaker diarization is described in Section 4. The overlapped speech detection method is introduced in Section 5. The experimental results are reported in Section 6. Finally, we conclude in Section 7.

2. Voice Activity Detection

The spectral subtraction process [11, 12] is adopted for noise reduction to assist the VAD process. We choose the spectral subtraction technique for its low computational cost and implementation complexity. 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 spectrum from the noisy signal. The standard deviation of cleaned signal is used to distinguish the speech frames and noisy frames. The details of the VAD algorithm can be found in [4]. The spectral subtracted speech signal is used for the frame selection of VAD as well as for the entropy feature extraction of overlapped speech segment detection. The acoustic features (MFCC and LPC) for speaker diarization and speaker recognition are still derived from the original speech signals.

3. Feature Extraction

The LPC features are used in the speaker diarization system [4]. The LPC features, the entropy feature in spectral domain and the frame energy [5] are adopted for the overlapped speech detection. The MFCC features are applied for the summed channel speaker recognition system [4].

3.1 Entropy Feature

The entropy was originally defined for information theory by Shannon [13]. It measures the average length of bit code per symbol under optimal coding and is expressed as:

$$I(S) = -\sum_{i=1}^{N} P(s(i))\log_2(P(s(i))) \quad (1)$$

where $S = [s(1),...,s(N)]$ indicates a source of N samples. $P(s(i))$ is the probability of emission of symbol i. The entropy $I(S)$ is maximal when all the symbols are equal probable and minimal while one symbol has a probability of one and the others are zeros.

To apply the entropy in the overlapped speech detection is based on the factor that the signal spectrum has higher entropy probability during the overlapped speech section than the non-overlapped speech section. Instead of the entropy measurement in time domain [5, 6], the measure of entropy is defined in the spectral energy domain can be expressed as:

$$E(|Y(w,t|^2) = -\sum_{w=1}^{\Omega} P(|Y(w,t)|^2)\log_2(P(|Y(|w,t)|^2) \quad (2)$$

$|Y(w,t)|^2$ is the spectral energy of the frame t, and $P(|Y(w,t)|^2)$ is the probability of the frequency band of frame t.

To reduce the effects of noise in the entropy estimation, we use the spectral subtracted signal, which are generated for the VAD frame selection, to extract the entropy feature.

3.2 LPC Feature

The linear predictive coding (LPC) coefficients of speech signal contain the formants of a speaker, which represent the spectral envelope of speech based on linear predictive model. We use LPC features in the speaker diarization as well as the overlapped speech detection. 12 LPC features are computed over a speech frame of 30 ms with a frame shift of 12.5ms.

3.3 MFCC Features

MFCC features are used for the summed channel speaker recognition system. A 16-dimenison MFCC features including C_0 are generated for each speech frame with a window of 30ms and a frame shift of 12.5ms. The C_0 feature will be also used in the overlapped speech modeling. 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 selected feature vectors are processed by RASTA filtering [14] and cepstral mean and variance normalizations (MVN).

4. Speaker Diarization

For the speaker diarization, we first conduct the initial speaker purification which is important to the subsequent speaker merging and clustering. A hybrid speaker diarization strategy for the summed channel audio segmentation is adopted to improve the initial clustering [4, 8, 9]. It consists of a GMM based progressive purification process and a Viterbi decoding process for clustering. Since there are only two speakers in the summed channel, the BIC criterion [8, 15] can be applied directly to merge the similar speech segments until two clusters are left. We summarize the clustering method [4] as follows:

1 Process the speech signal with 12-dimension LPC feature vectors and divide them into segments of 2 second, and then group them into $Q = 15$ initial clusters.

2 Perform the initial cluster purification via EM and MAP adaptation [16] as follows:

2.1 Train a Root GMM, λ_{Root} initially with 2 mixture components using all the clusters;

2.2 Train the cluster-dependent GMMs, $\lambda_1, \lambda_2, ..., \lambda_Q$, by adapting λ_{Root} based on MAP adaptation;

2.3 Evaluate all the segments against the cluster-dependent GMMs, $\lambda_1, \lambda_2, ..., \lambda_Q$, and relocate the segments into the clusters accordingly;

2.4 Repeat the steps 2.2 to 2.3 until no segment change is found;

2.5 Increase the mixture components of GMM model by 2 and repeat step 2.1 until the number of mixture is equal to 16.

3 Based on the initial purification, run Viterbi decoding to re-align and merge the cluster pair with the largest BIC score until the number of clusters is reduced to 2.

5. Overlapped Speech Detection

The overlapped speech detection consists of training and testing. The 1-conversation speech data from the NIST SRE06 and SRE05 databases are used to train overlapped speech models and fine tune the trained models respectively. These databases recorded two speaker voices on two separate

channels. To simulate a summed channel conversation, these two channels were summed into single channel.

5.1. Model Training

The GMM modeling is adopted in the overlapped speech detection. The GMM models are trained based on an EM via MAP adaptation [16] by using a combination of LPC, entropy and the MFCC C_0 features [5] described in Section 4. We first train a root GMM of 32 mixture components using all the speech features. The single speech model and overlapped speech model are adapted from the root model. The training dataset consists of 200 1-conversation recordings from SRE06 database. With the ASR transcriptions and time alignment information provided by NIST for ground-truth index of the two channels, these training data are classed into non-overlapped speech and overlapped speech for the model training.

5.2. Fine-tuning for Overlapping Speech Detection

The 1-conversion speech data from SRE05 are used as the development dataset for fine-tuning the classification threshold for non-overlapped and overlapped speech segments. The non-overlapped and overlapped speech datasets are generated as same as for SRE06 dataset.

The two sets of speech segments from SRE05 dataset are tested against the non-overlapped speech model and overlapped speech model generated based on SRE06 dataset. Given a speech segment *s,* it was scored against both the non-overlapped speech model $\lambda_{non-overlapped}$ and the overlapped speech model $\lambda_{overlapped}$ to estimate a likelihood score:

$$\Lambda(s) = \ln[P(s \mid \lambda_{overlapped})] - \ln[P(s \mid \lambda_{non-overlapped})] \quad (3)$$

Considering there might be a scoring bias between the two models, we decide the overlapped speech as:

$$\Lambda(s) = \begin{cases} \text{overlapped speech, if } \Lambda(s) > \mathrm{T}_{\text{thres}} \\ \text{single speech,} \qquad \text{otherwise} \end{cases} \quad (4)$$

where the bias threshold T_{thres} is computed to minimize DER based on the SRE05 development dataset:

$$(T_{thres})_{\min DER} = \arg\min_{T_{thres}} \{ | \frac{SE + MS + FA}{SPK} | \} \quad (5)$$

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 of speech segments in which missed overlapped speech detection happen, the false alarm speaker time (FA) is the total time of speech segments in which false overlapped speech detection happen, and scored speaker time (SPK) is the time of the whole conversation.

Finally, the above-mentioned two trained models and threshold are used to remove the overlapped speech segments in the summed channel speaker recognition.

6. Experiment Design

The summed channel speaker recognition experiments were conducted on the 1conv-summed and 8conv-summed subtasks of the NIST SRE08. The GMM-SVM speaker modeling with 46-dimension MFCC features were implemented. We evaluate the speaker recognition performance by both the Equal Error Rate (EER) and the Detection Cost Function (DCF) [1].

6.1. GMM-SVM based Speaker Recognition

We built the GMM-SVM speaker classifier using 46 MFCC features. 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 [16]. We formed a GMM supervector for each conversation which is normalized by its standard deviation and weighted by the squared root of the weights of the Gaussian mixtures. The SVMTorch [17] was used to train the SVM model. The channel normalization was conducted using Nuisance Attribute Projection (NAP) [10] to project out the nuisance subspace from the original supervector space. The rank of NAP is set to be 60.

The SRE04 corpus was used as the background training data set for UBM as well as the background speaker data set for SVM training. At the same time, the SRE04 corpus was also used to derive the NAP matrix. As for the score normalization, we reported the experimental results based on the TZnorm scores normalization [18]. In the experiments, the SRE05 1-side training data were used for training cohort models in Tnorm and the SRE04 data were used as imposter speech utterances in Znorm.

6.2. Speaker Diarization

NIST has provided the automatic speech recognition (ASR) transcripts for the evaluation data in SRE08. Similar to what is in [3, 4], we were able to obtain about 50% speech transcripts and alignments from the corresponding 4-wire ASR results for the SRE08 summed channel test set. These 4-wire ASR transcripts and alignments provide the ground-truth of the speakers' voice activity detection. We used these recordings as the evaluation data set to evaluate the overlapped speech detection method and speaker diarization performance on SRE08. The DER results for the development dataset and SRE08 summed channel evaluation subsets (about 50% of whole sets) are shown in Table 1.

From Table 1, we can see that the overlapped speech detection has successfully reducing the DER rates and achieved about 18.62% and 17.99% relative improvements on development dataset and evaluation dataset, respectively.

Table 1. DERs with or without overlapped speech detection.

DER	Development data sets	SRE08
Without Overlapped Speech Detection	13.21%	12.51%
Overlapped Speech Detection	10.75%	10.26%

6.3. Speaker Recognition

To evaluate the benefit from the speaker diarization and overlapped speech detection, we first conducted speaker recognition experiments with the summed speech data without separating the speakers, and then applied the speaker diarization to verify the speaker recognition performance. Finally, we used the overlapped speech detection to process the diarized signal and remove overlapped speech segments. The experimental results on the SRE08 1conv-summed and 8conv-summed tasks are shown in Table 2 and Table 3, respectively.

From Table 2, 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. We achieved an MFCC EER of 4.45% and a minimum DCF of 2.13%, representing a 48.2% relative improvement in EER, and 43.1% relative improvement in minimum DCF. Furthermore, with the overlapped speech detection, the performance in EER and DCF were further improved to 3.86% and 1.93%, about 13.3% relative improvement in EER and 9.4% improvement in DCF over the speaker diarization without overlapped speech detection, respectively.

In Table 3, we reported the performance on the SRE08 8conv-summed task. We observed similar EER reduction from 2.04%% to 1.79% by applying overlapped speech detection to diarized signal.

Table 2. EER and min DCF for the SRE08 1conv-summed subtasks (All English Trials) without and with speaker diarization and overlapped speech detection.

Data Conditions	Male		Female		All	
	EER %	DCF x100	EER %	DCF X100	EER %	DCF x100
Before diarization	6.98	3.01	7.67	2.94	7.47	2.98
With diar. Only	4.01	2.13	4.63	2.25	4.45	2.13
With both diar. and OL	3.27	1.87	4.21	1.90	3.86	1.93

Table 3. EER and min DCF for the SRE08 8conv-summed subtasks (All English Trials) without and with speaker diarization and overlapped speech detection.

Data Conditions	Male		Female		All	
	EER %	DCF x100	EER %	DCF x100	EER %	DCF x100
Before diarization	4.47	2.02	4.25	1.77	4.22	1.91
With diar. Only	2.58	1.23	1.70	0.77	2.04	1.03
With both diar. and OL	2.24	1.06	1.49	0.67	1.79	0.85

We also summarized the Detection Error Tradeoff (DET) [1] curves of SRE08 1conv-summed and 8conv-summed tasks in Figure 2.

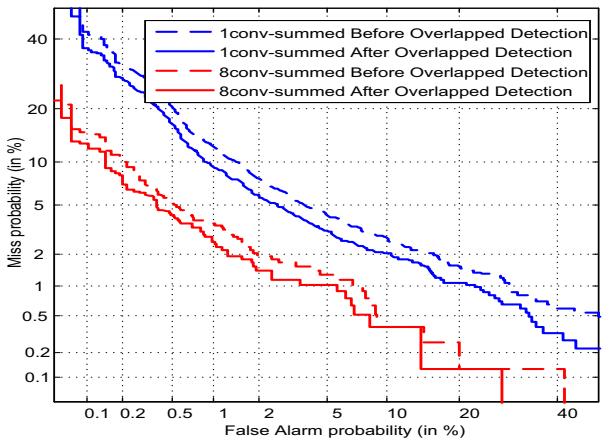

Figure 2. SRE08 1conv-summed and 8conv-summed subtask (All English Trials) DET curves with and without overlapped speech detection.

7. Conclusions

This paper studies the overlapped speech detection for SRE08 summed channel tasks. The results demonstrated that the overlapped speech detection has been successfully applied to the summed channel signal by achieving about 18% relative speaker diarization DER improvement on both the development dataset and SRE08 evaluation dataset. We have also achieved 13.3% relative EER improvement and 9.4% relative DCF improvement on 1conv-summed task. The results suggested that there is an obvious benefit by applying the overlapped speech detection for summed channel speaker recognition. Moving forward, we would like to test out the NIST speaker recognition tasks in which both the training data and test data are recorded in summed channels.

8. References

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

[2] 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.

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

[4] H. Sun, B. Ma, C. Huang, T. Nguyen, H. Li "The IIR NIST SRE 2008 and 2010 summed channel speaker recognition systems", in *Proc. Interspeech 2010*, pp.366-369, Japan, Sept. 2010.

[5] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. ICASSP 2008*, Las Vegas, 2008.

[6] O. Ben-Harush, I. Lapidot, H. Guterman, *"Entropy Based Overlapped Speech Detection as a Pre-Processing Stage for Speaker Diarization*", in *Proc. Interspeech 2009*, pp.916-919, Brighton, UK, Sept. 2009.

[7] "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.

[8] 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.

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

[10] 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.

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

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

[13] C. E. Shannon, "A mathematical theory of communication," *Tech. J.*, vol. 27, p. 379–423/623–656, 1948.

[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] 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.

[16] 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.

[17] 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.

[18] 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.