AN EFFICIENT FEATURE SELECTION METHOD FOR SPEAKER RECOGNITION

Hanwu Sun, Bin Ma and Haizhou Li
Institute for Infocomm Research
Agency for Science, Technology and Research (A*STAR), Singapore.
{hwsun, mabin, hli}@i2r.a-star.edu.sg

ABSTRACT

In this paper, a new feature selection method for speaker recognition is proposed to keep the high quality speech frames for speaker modelling and to remove noisy and corrupted speech frames. In order to obtain robust voice activity detection in variety of acoustic conditions, the spectral subtraction algorithm is adopted to estimate the frame power. An energy based frame selection algorithm is then applied to indicate the speech activity at the frame level. The eigenchannel based GMM-UBM speaker recognition system is used to evaluate this proposed method. The experiments are conducted on the 2006 NIST Speaker Recognition Evaluation core test condition (telephone channel) as well as microphone channel test condition. It demonstrates that this approach can provide an efficient way to select high quality speech frames in the noisy environment for speaker recognition.

Index term- speaker recognition, voice activity detection, feature selection, spectral subtraction, noise reduction

1. INTRODUCTION

Acoustic features based on short-term spectral analysis have been the dominant feature type adopted in speaker recognition systems, due to their computation efficiency and effectiveness for recognition performance [1]. One of the important issues for the acoustic feature extraction in speaker recognition is to explore efficient and robust feature selection for the modeling of speaker recognition systems. The feature selection aims to remove non-speech and choose useful speech signal for the speaker modeling with a voice activity detection (VAD) algorithm.

VAD is an outstanding problem for speech transmission, enhancement and recognition. The varying nature of speech and background noise makes it especially difficult. Besides the high quality telephone speech signals, speaker recognition has also to be conducted in the hand-free equipments and devices such as mobile phones, PDAs and other communications devices. The challenges for a robust VAD algorithm in these applications are various background noisy signals picked by various microphones, different recording distances between microphones and speakers, and different transmitted channels. There are wide ranges of speaker recognition applications in these scenarios, but the low signal-to-noise (SNR) of the captured signals hamper the speaker recognition accuracy in these situations.

The National Institute of Standard in Technology (NIST) [2, 3] has organized series of speaker recognition evaluations (SREs), focusing on the text-independent speaker recognition technology. In recent years, the measurement for the state-of-the-art approaches have been conduced with both high quality telephone conversational speech and low quality microphone channel speech data [2, 3].

A good VAD algorithm is crucial for achieving desired speaker recognition performance, by keeping efficient speech frames and removing the noise frames and corrupted speech frames [4]. As a result, only high quality speech frames are selected to model the speakers in speaker recognition system. In the noise environment, the voice activity detection is difficult to identify the speech frames due to the low SNR. In this paper, we propose a new feature selection method based on the spectral subtraction [5, 6]. The spectral subtraction is designed to improve the VAD capability for speech recognition in the adverse acoustic conditions, The selected features will be used to model speakers with universal background model (UBM) based Gaussian mixture modeling (GMM) method [7]. To compensate the channel effects, eigenchannel adaptation algorithm [8] has been applied.

This paper is organized as follows. The spectral subtraction algorithm is described in Section 2. The frame level feature selection method based on spectral subtraction is presented in Section 3. In Section 4, the speaker recognition experiments using the GMM-UBM speaker recognizer approach are conducted on the 2006 NIST SRE core test set (telephone channel) and microphone channel test set. Finally, we summarize our findings in Section 5.

2. SPECTRAL SUBTRACTION

In order to make an efficient and robust feature selection for speaker recognition in noise environment, we adopt the spectral subtraction process to assistant the voice activity

detection process. The main reasons we chose spectral subtraction for noise reduction is that the computational cost and implementation complexity are relatively low. The basic idea of the spectral subtraction method is to suppress the additive noise in the corrupt 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. It is well known that the proposed scheme makes use of a frequency domain-based technique, commonly known under the generic name of spectral subtraction [5]. The spectral subtraction algorithm can be summarized as follows.

The noisy signal is given by

$$z(t) = s(t) + n(t) \tag{1}$$

where $z(t)$, $s(t)$ and $n(t)$ denote noisy signal, clean signal and noise, respectively. That is, *s(t)* is the clean speech.

We divide the sampled time series into 256 samples per frame with 160 shift samples. To ensure the signal continuity, each frame is windowed using Hamming window. Fast Fourier Transform (FFT) is applied on the *k*-th frame and we will obtain:

$$Z_w(f) = FFT\{z(t)w(t)\},\ t = 1,2,...,L \tag{2}$$

The noise spectral $|\overline{N}^{(k)}(f)|$ can be estimated and updated by:

$$|\overline{N}^{(k)}(f)| = \alpha\,|\overline{N}^{(k-1)}(f)| + (1-\alpha)\,|\overline{Z}_w(f)| \tag{3}$$

where $\overline{Z}_w(f)$ is the noise floor estimator using method proposed in [6]. The amplitude of the signal spectral $|\overline{S}^{(k)}(f)|$ can be estimated and updated by

$$|\overline{S}^{(k)}(f)| = \alpha\,|\overline{S}^{(k-1)}(f)| + (1-\alpha)\max\{[Z_w(f) - \beta\,|\overline{N}^{(k-1)}(f)|], 0\} \tag{4}$$

In Eq. (3) and (4), α denotes a constant ($0 < \alpha < 1$), is used to control the speed of signal and noise spectral amplitude tracking. The parameter β controls the amount of noise subtracted from the noisy signal $\overline{Z}_w(f)$. For full noise subtraction β is set to be 1 and for over-subtraction β can be set to a value greater than 1.

The noise reduction signal is obtained by

$$|\overline{S}(f)| = \rho\,|Z_w(f)| = \frac{|\overline{S}^{(k)}(f)|}{|\overline{S}(f)| + \beta\,|\overline{N}^{(k)}(f)|} \cdot |Z_w(f)| \tag{5}$$

$$\arg[\overline{S}(f)] = \arg[Z_w(f)] \tag{6}$$

where $|\overline{S}(f)|$ is an estimate of the clean signal spectrum magnitude $|S(f)|$. arg[] denotes the operation for phase operator.

Finally, we recover the signal by applying the Inverse Fast Fourier Transformation (IFFT) and overlapping the continuous frames properly:

$$S(f) = |\overline{S}(f)| \cdot \arg[Z_w(f)] \tag{7}$$

$$\overline{s}(t) = IFFT\{S(f)\} \tag{8}$$

The $\overline{s}(t)$ is the subtracted output signals.

3. FRAME SELECTION ALGORITHM

Based on the spectral subtraction signal, we proposed an energy based frame selection method for the speaker modeling in the speaker recognition system. The selected feature frames are obtained via the following energy based frame selection algorithm:

$$\text{Selected } \textit{Feature Frame} \Rightarrow \begin{cases} P(i) > \max[P(j)_{j=1,2,3,...,N}] - T_{SNR} \\ P(i) > T_{\min} \end{cases} \tag{9}$$

where *P(i)* is the power or standard deviation of *i*-th frame, the 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 parameter $T_{\min}$ may be set to be -60 dB if the input voice signal is in the normalized range 1 to -1. According to this frame selection scheme, the higher the T_{SNR} value, the more feature frames will be selected.

Obviously, there are two type frame powers *P(i)* which we can use. One is the frame power of original speech signals and another one is the frame power based on the spectral subtracted signal. The later one is adopted to make robust voice activity detection in noise voice signals. The spectral subtracted voice signal is only used for the frame selection purpose. The acoustic features for speaker recognition will still be extracted from the original voice signals. Figure 1 shows the flowchart of the proposed frame selection method in which the spectral subtraction process is used to indicate the selected speech frames.

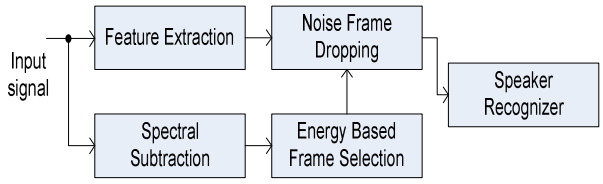

Figure 1: Flowchart of Spectral Subtraction Based Frame Selection.

4. EXPERIMENTS

4.1 Experiment Setup

In the experiments, we used the 2006 NIST SRE core test condition (1-conversation for training and 1-conversation

for test) and microphone test condition (1-conversation for training and 1-conversation for test) as our evaluation corpus. The results are presented in the terms of equal error rates (EERs) and minimum detection cost function (Min C-Det) defined by NIST speaker evaluation rule [2].

The GMM-UBM speaker recognition system with eigenchannel adaptation [7, 8] was chosen to conduct the experiments. We extracted 12 MFCCs, plus their first and second order derivatives, total 36 dimensions features with the frame length of 240 samples (30ms) and frame shift of 160 samples (20ms). The proposed energy-based feature selected scheme was applied to remove silence frames and to retain only the high quality speech frames. The feature vectors of the selected frames were processed by mean-variance-normalization (MVN) and RASTA filtering [9].

The 2004 NIST SRE 1-conversation data were used for training gender-dependent UBMs with 512 mixtures, as well as for the eigenchannel adaptation for the 2006 NIST SRE telephone channel experiments. The 2005 NIST SRE microphone data were used for the eigenchannel adaptation for the 2006 NIST SRE microphone channel experiments. The number of eigenchannels was set to be 30. The scores are normalized with Tnorm algorithm [10]. The 2005 NIST SRE 1-conversation training data was used for training cohort models in Tnorm.

For the spectral subtraction, we set α to be 0.9. α is used to control the speed of signal and noise spectral amplitude tracking. Since the power of subtracted frames is only used to indicate the voice activity, we set β to be over-subtraction value 1.2 for the easy frame selection.

4.2. Statistics of the Selected Frames

The percentages (%) of selected frames from the whole utterances for both telephone (core test) and microphone channels are shown in Figure 2, according to the feature selection method in (9), under different cutoff SNR threshold T_{SNR}.

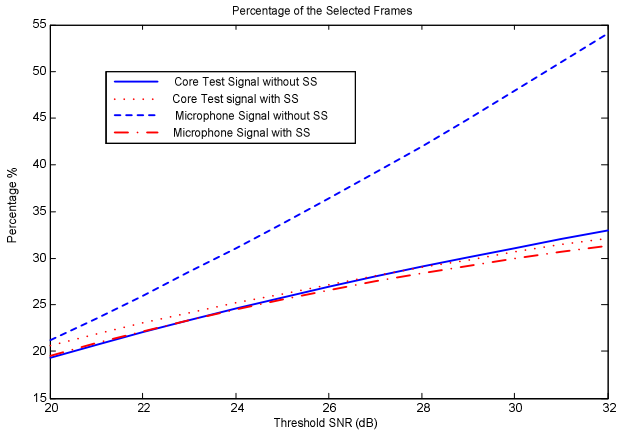

Figure 2: Percentages of the Selected Frames with or without Spectral Subtraction (SS) at Different Cutoff SNR Thresholds.

From Figure 2, we can see that for the telephone channel signals, the percentages of the selected frame with and without spectral subtraction process are consistent and have the similar trend. The reason is that the telephone channel signals in the 2006 NIST SRE core test set have high signal-to-noise ratio (generally larger than 30dB). The spectral subtraction method does not have big effect on the frames selection process as in (9).

For the microphone channel signals, which were recorded using distant microphone, have poor SNR ratios. From Figure 2, we can see that there are big differences in selected percentages with and without spectral subtraction for the microphone channel signals. It is not surprised that the percentages of selected features without spectral subtraction are much higher than that of the telephone channel signals. It is also noticed that the trend of the selected frame percentages for the spectral subtracted microphone channel signals is consistent with that of telephone channel signal, which is expected.

4.3. GMM-UBM Experiments

The speaker recognition experiments were conducted with the eigenchannel adaptation based GMM-UBM speaker recognition system, using both telephone channel core test set and microphone channel test set in the 2006 NIST SRE corpus, under different cutoff SNR thresholds. We would like to evaluate the speaker recognition performances of spectral subtraction process, as well as the cutoff SNR threshold setups for feature selection.

We chose the cutoff SNR threshold T_{SNR} from 24 dB to 31 dB and applied the feature selection based on these thresholds to all the training and test data of speaker recognition experiments. The recognition results in both EER (%) and Min C-Det are shown in Table 1 for telephone channel test set, and in Table 2 for microphone channel test set. The experiment results at the cutoff SNR as 30 dB, for both telephone and microphone channel test set, with and without spectral subtraction process are also shown in Figure 3.

Table 1: Effect of the Spectral Subtraction on the 2006 NIST SRE Core Test Set (telephone channel)

	VAD without Spectral Subtraction		VAD with Spectral Subtraction	
T_{SNR} (dB)	*EER%*	*Min C-Det*	*EER%*	*Min C-Det*
24	5.42	0.0248	5.27	0.0241
25	5.28	0.0241	5.17	0.0240
26	5.25	0.0234	5.13	0.0237
27	5.06	0.0233	4.97	0.0232
28	4.95	0.0232	4.90	0.0232
29	4.95	0.0231	4.85	0.0231
30	4.91	0.0230	4.92	0.0230
31	4.98	0.0231	4.99	0.0231

Table 2: Effect of the Spectral Subtraction on the 2006 NIST SRE Microphone Channel Test Set

	VAD without Spectral Subtraction		VAD with Spectral Subtraction	
T_{SNR} (dB)	*EER%*	*Min C-Det*	*EER%*	*Min C-Det*
24	6.38	0.0231	5.46	0.0210
25	6.21	0.0227	5.43	0.0209
26	6.05	0.0220	5.36	0.0205
27	5.98	0.0222	5.37	0.0201
28	6.09	0.0217	5.40	0.0201
29	6.04	0.0222	5.47	0.0199
30	6.12	0.0223	5.27	0.0197
31	6.16	0.0225	5.32	0.0200

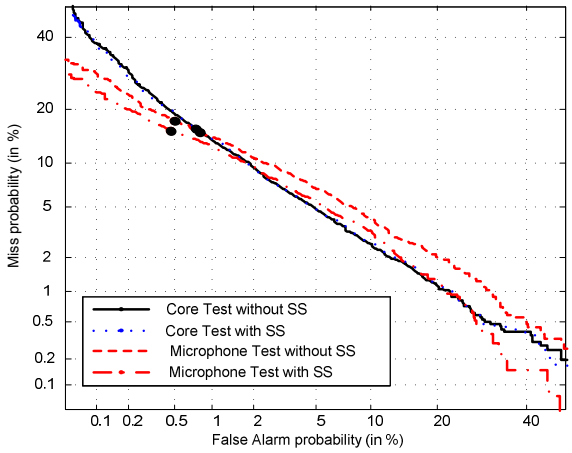

Figure 3: Effect of the VAD with or without Spectral Subtraction at the T_{SNR} =30 dB.

From Table 1, it can be seen that the results with spectral subtraction are slight better than those without spectral subtraction, but this improvement is very small. For the feature selection, it might not be necessary to apply the spectral subtraction to the clear voice signal.

From Table 2 and Figure 3, we can see that the frame selection with the spectral subtraction process gives impressive improvements of both EER and Min C-Det for the microphone channel test set. The relative EER improvement is between 11% up to 16%. It is crucial to apply the spectral subtraction for a better feature selection on the noisy voice signals in speaker recognition.

4.4. Effect of the Cutoff SNR Threshold

From the results with spectral subtraction under various cutoff SNR thresholds T_{SNR} shown in Table 1 and 2, we can see that the EER rate and C-Det are improved as the increase of the cutoff SNR threshold until around 30dB. At the same time, we also notice that the performances (both EER and C-Det) with our feature selection scheme under the cutoff SNR thresholds T_{SNR} from 27 to 31 dB are almost constant and have little variation. It is reasonable to set the SNR threshold T_{SNR} between 27 to 31 dB to conduct the feature selection. Of course, we observed that the best result for both EER and Min C-Det for the telephone channel and microphone channel test sets happened at around 30 dB. We may conclude that 30 dB cutoff SNR threshold is good choice for the frame based feature selections in our GMM-UBM speaker recognition system.

5. CONCLUSIONS

A spectral subtraction based frame selection method for speaker recognition for noisy condition has been presented in this paper. The speaker recognition experiments were conducted on both the clear 2006 NIST SRE telephone channel and noisy 2006 NIST SRE microphone channel conditions. The experiment results demonstrated that the proposed method can effective select the efficient feature frames for speaker recognition under both clear and noisy conditions. The significant EER improvement was obtained for the microphone channel test using this frame selection method. We also found that the 30 dB cutoff SNR threshold is good choice for the frame selection based speaker recognition in the eigenchannel adaptation based GMM-UBM speaker recognition system.

REFERENCES

[1] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. van. Karafiat, D. A. Leeuwen, P. Matejka, P. Schwarz and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15 no. 7, pp. 2072–2084, 2007.

[2] NIST 2006 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf.

[3] NIST Speaker Recognition Evaluation Home page: http://www.nist.gov/speech/tests/spk/index.htm.

[4] Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the ELISA consortium research activities," in *2001: a Speaker Odyssey*, pp.67-72 Jun. 2001.

[5] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. ASSP*, vol. 27, pp. 113-120, 1979.

[6] R. Martin, "Spectral subtraction based on minimum statistics", *EUSPICO, Proc,* vol. 2, pp.1182-1185, 1994.

[7] 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.

[8] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp.1435-1447, 2007.

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

[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.