ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

An Enhanced RASTA processing for speaker identification[1]

ZHEN Bin, WU Xihong, LIU Zhimin, CHI Huisheng

Center for Information Science, Peking University, Beijing, 100871

Abstract: In this paper, we propose an Enhanced RASTA (E_RASTA) technique for speaker identification. The new method consists of classical RASTA filtering in logarithmic spectrum domain following by another RASTA processing in spectrum domain. In this manner, both the channel distortion and additive noise are removed effectively. In isolated digit speaker identification experiment on TI46 database, we found that the E_RASTA performed equal or better than J_RASTA method. The new method does not need the estimation of speech SNR in order to determinate the optimal value of J and multi-templates in J_RASTA, and the information of how the speech degrades.

Ⅰ Introduction

A typical speaker identification (SID) system contains a feature extraction module followed by a classifier.[1] While the classifier is trained on training data, one of the main difficulties is the inconsistent of the training speech and recognition speech. Usually the speech is corrupted by additive noise and channel distortion in environment.[2]

For additive noise, if the noise is uncorrelated to the original speech, the noise component is additive in the power spectrum of the signal[2]. One accepted way to deal with the noise is spectral subtraction, in which the estimated noise power spectrum is subtracted from that of the speech.[3] At least two problems arise: 1) A speech detector is needed to determine the speech interval to obtain reliable noise estimate. 2) The subtraction may lead to negative power spectrum value. Channel distortion indicates different communication channel between the training and recognition speech, e.g., by switching to a new microphone.[2] If the channel distortion is fixed or only slowly varying in time, cepstral mean subtraction is one way of addressing the problem.[4] But the method does require a long-term average, which may be difficult to obtain in real-time implementation. Hermansky proposed to filter the logarithmic power spectrum through a RASTA filter to deal with the channel distortion.[5] If there are both additive noise and channel distortion, Hermansky proposed J_RASTA method to filter the channel distortion in high signal-to-noise ratio (SNR) in logarithmic power spectrum and the power spectrum in low SNR.[5] At least three problems arise: 1) Obviously, the J_RASTA method left the additive noise in high SNR and channel distortion in low SNR unprocessed. 2) A speech interval is required to estimate the SNR of the speech to determine the particular value of J. 3) Multi-template are required in classifier for the signal dependent value of J. All this make it difficult to be used in real time.

In this paper we proposed an enhanced RASTA processing for SID to deal with both additive noise and channel distortion. The paper is organized as follows. Section Ⅱ describe the principle of enhanced RASTA method. Section Ⅲ applied the enhanced RASTA method to SID. Section Ⅳ is the discussion and conclusion.

Ⅱ Principle of the enhanced RASTA method

The steps of RASTA performed on spectrum are as follows. For each analysis power spectrum frame, do the following[5]

1) Compute the auditory spectrum using filter bank.
2) Transform the spectral amplitude through a nonlinearly compression (e.g., logarithmic transform).
3) Filtering the time trajectory of the compressed spectral component with RASTA filter, whose transfer function is

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}} \qquad (1)$$

The J_RASTA method substitute for the logarithmic operation in step 2

$$y = \log(1 + Jx) \qquad (2)$$

where J is a signal-dependent positive constant and x is the power spectrum. The amplitude-warping transform (2) is linear-like for $J<<1$ and logarithmic-like for $J>>1$. Thus, the J_RASTA method may filter the uncorrelated additive noise component in low SNR and the channel distortion in high SNR.

The J_RASTA method indicates that if the additive noise is uncorrelated to the speech, we can remove the modulation component from white noise by simple RASTA-like filtering on spectrum domain. We refereed it additive RATSA for the RASTA like filtering performed on spectrum domain in the paper. Then, we have to determine the low and high cutoff frequency of the bandpass filter. The statistic information of speech temporal modulation across frequency channel can be compute as in Fig. 1

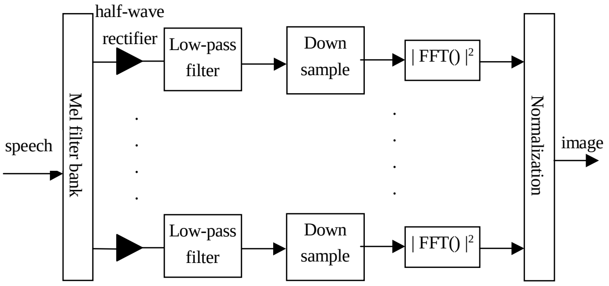

Fig. 1 Diagram of the processing of statistic temporal modulation

Incoming speech is analyzed into the Mel filter bank.[6] Within

[1] This work was supported by the key projects of national nature science funds grants 69635052 and postdoctoral fundation..

each channel the signal envelope is derived by half-wave rectification and low-pass filtering. After dowsampled, the modulations of the envelope signals are analyzed by FFT. Finally, the normalized squared magnitude of the FFTs from each channel is plotted into an image. Fig.2 is the statistic temporal modulation of Gaussion white noise and the speech from TI46 database. The x-axis is the normalized frequency and the y-axis is the index of frequency channel. The temporal modulation frequency of white noise is rather low in all frequency channels, while that of speech expend to a wider frequency range. This is not quite strange. Because the white noise is a stationary signal, its envelop will not change too much from frame to frame. But for speech signal, it is a non-stationary signal, and only can be regard as stationary within 10ms-30ms. Thus, the envelop of speech changes rapidly in frame scale. Then, in principle, filtering all slowly and rapidly varying components in the magnitude spectrum may not significantly harm information bearing components, but it may reduce some slowly or rapidly varying noise in the speech.[5]

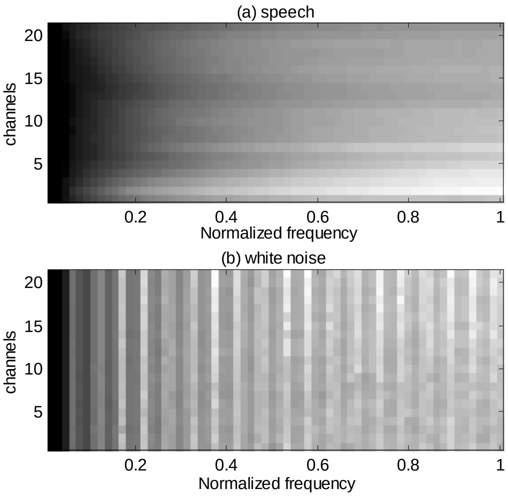

Fig. 2 Statistics information of the temporal modulation of (a) speech and (b) Gaussion white noise. The deeper the color, the bigger the value.

Thus, the total step of the enhanced RASTA filter, can be summarized as follows:

1) Compute the auditory spectrum using Gammatone filter bank.
2) Transform the spectral amplitude through a nonlinearly compression (e.g., logarithmic compression).
3) Filter the time trajectory of the compressed spectral component with convolution RASTA filter of Eq. 1.
4) Expand the filtered logarithmic spectral component nonlinearly to spectral domain
5) Filter the processed spectrum with another additive RASTA filter.

At last, transform the logarithmic absolute value of the filtered spectrum to cepstrum domain by Discrete Cosine Transform (DCT) to serve as feature vectors in classifier. Based on the Fig.2 and the recognition result, the transfer function of additive RASTA filter in spectrum domain is

$$H(z) = 0.33z^{4} * \frac{1-2z^{-2}+z^{-4}}{1-1.59z^{-1}+0.63z^{-2}-0.19z^{-3}+0.17z^{-4}} \tag{3}$$

whose 3dB passband is 1.5Hz~24Hz. Fig. 3 illustrates the frequency responses of the two RASTA filter.

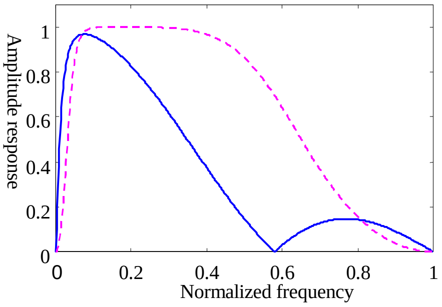

Fig. 3 The frequency response of the additive (dotted line) and convolution (solid line) RASTA filter

Ⅲ Experiments

3.1 Speech database

The speech database used in the experiments was the ten isolated digits of standard speech database TI46. Each digit was spoken by sixteen speakers (eight females and eight males). The data of these digits were divided into two sets (training and testing). In training set, each digit was repeated 10 times by each speaker in one session. In the testing set, each digit was repeated 16 times by each speaker in eight different sessions. In each session, two utterances were recorded.

The ten utterances from the training set were used for training and the sixteen utterances from the testing set were used for testing. During training phase, we use DTW to obtain one template for each digit from training sets. The three kinds of degraded speech shown in Fig. 4 are used to evaluated the performance, which were referred as additive degraded speech, additive-filtered degraded speech and filtered-additive degraded speech, respectively. The channel filter simulates the telephone channel and its 3dB passband is 300Hz-3300Hz as shown in Fig. 5. We also introduce the speech like additive noise to the speech. The speech like noise simulated the background noise in the noisy environment and can be generated by a Gaussion white noise passing through a filter whose frequency response is shown in Fig. 6. The SNR of a test utterance is defined as the ratio of the power of clean speech to the noise power.

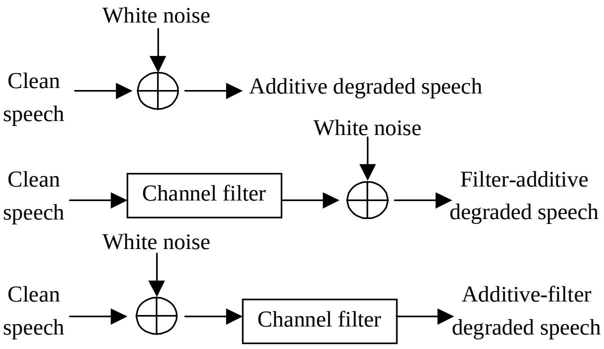

Fig.4 Schematic representation of the degrade speech

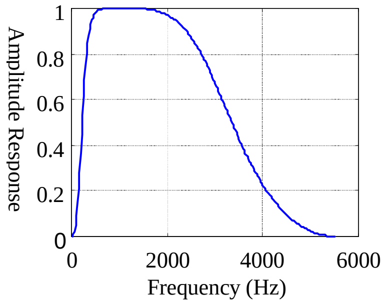

Fig. 5 Frequency response of the channel distortion filter.

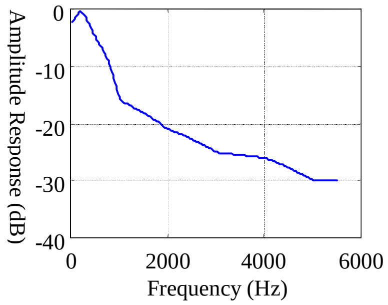

Fig 6. Frequency response of the speech like noise filter

3.2 Recognition result

In this section the results of text-dependent speaker identification are reported. To make a comparison, we also apply MFCC and Gammatone Cepstrum Coefficients (GCC) as traditional features.[7] Gammatone filter bank simulates the function of peripheral auditory system, and has been widely used.[10] The dimension is 19, which is almost the best value for the dimension for speaker identification using MFCC, and C_0 is discarded to normalize the average energy.[8] The SID was done using a simple-but-efficient DTW-based recognizer for comparing features. The recognizer was trained on the clean speech of the training set. We use one template for each digit.

Fig. 7 show the average recognition rate of different feature with additive noise degraded speech. The recognition rate is the average recognition rate of digit from 0 to 9 for 16 person. The GCC outperforms MFCC in low SNR and is nearly as good as MFCC in high SNR. The average recognition rates of J_RASTA are less than those of MFCC's and GCC's when SNR is higher than 30dB. The average recognition rates of additive RASTA are much better than those of MFCC's and those of GCC's in low SNR and are almost the same as those in high SNR. The additive RASTA performs better than J_RASTA when SNR is higher than 30dB, but worse than it when SNR is less than 30dB.

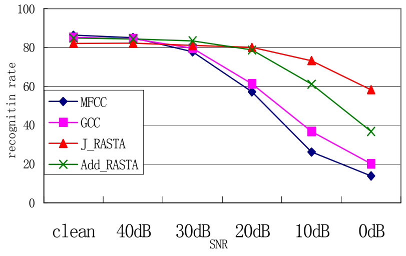

Fig. 7 The average speaker identification ratio of additive degraded speech. J_RASTA indicates GCC after J-RASTA filtering; Additive RASTA indicates GCC after additive RASTA filtering in spectrum domain.

As we know, the J_RASTA method is equal to DCT(SP) in low SNR condition, while the additive RASTA we proposed is DCT(log(SP)). Then, a question is how does the logarithmic affect recognition. We combined the RASTA filtering and logarithmic compression each other in low SNR and the recognition results are as follow:

Table 1. Average recognition rate of the combination of log() and RASTA(). All the features were DCT transformed before recognition. SP indicates power spectrum; Log(SP) indicates the logarithmic on power spectrum; RASTA(SP) indicates the classical RASTA filtering on spectrum; RASTA(log(SP)) indicates classical RASTA filtering on logarithmic spectrum.

	SP	log(SP)	RASTA(SP)	RASTA(log(SP))
10dB	63.26	36.72	64.96	30.97
0 dB	50.03	20.13	60.16	14.68

With logarithmic compression of power spectrum, the recognition rates decrease 30% or so whether there is RASTA filtering or not. While the RASTA filtering increase the recognition rates no more than 2%, or even decrease the recognition rate on spectrum domain. Thus, the better performance of J_RASTA benefits more from linear transform than from RASTA filtering, because the particular value of J influences the shape of logarithmic spectrum with different emphasis on peaks and valleys of the spectrum. It seems that the nonlinear compression plays an important role on the feature extraction.[5,9]

Fig. 8 and Fig. 9 is the recognition rate of degraded speech with both additive and convolution noise. The E_RASTA performs equal to the J_RASTA technique for filter-additive degraded speech and better than it for additive-filter degraded speech. For the two kinds of speech, the average error rate of all SNR condition of E_RASTA reduces 36.5% and 44.5% compared with that of MFCC's. Although the two degraded speeches were contaminated in different ways, it is not surprising for the enhanced RASTA has good performance for both kinds of degraded speech. As illustrated in Fig. 10, from the signal processing point of view, the additive noise in filter-additive degraded speech can be viewed as that the additive noise is added before the filter with another inverse filtering. The higher recognition rates of additive-filter degraded speeches than those of filter-additive degraded speech's are due to their relative higher SNR. Fig. 11 is the recognition rate of speech-like additive noise degraded speech. The average error rate of all SNR condition of E_RASTA reduces 28.2% compared with that of GCC's. The above results indicate that the E_RASTA can filter the noisy speech degraded by stationary noise no matter how it is contaminated.

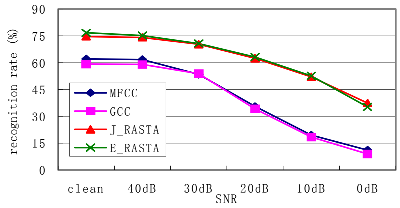

Fig. 8 The average speaker identification ratio of filter-additive degraded speech

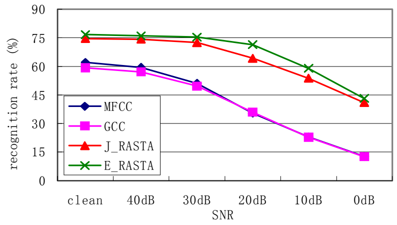

Fig. 9 The average speaker identification ratio with additive-filter degraded speech

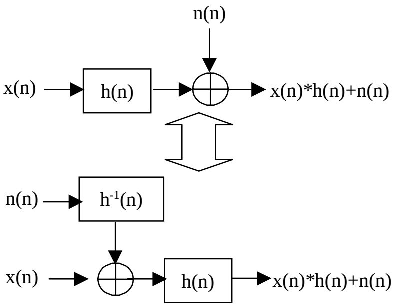

Fig 10 The equivalent of additive-filter degraded speech and filter-additive degraded speech from the signal processing point of view

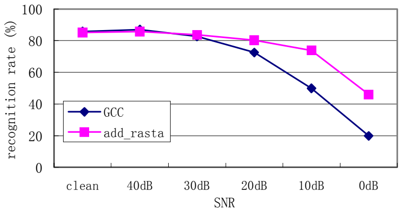

Fig 11 The average speaker identification ratio with additive speech-like noise degraded speech

Although the RASTA technique turned out to be rather successful, the SNR dependent value of J may be one of the factors prevent it further application in speech processing. Explicit estimation of the noise level in nonspeech interval is cumbersome, error prone and may not be necessary. Because the additive RASTA filter is only clean speech dependent, it does not use any information of the noise. Thus, the E_RATSA does not need to estimate the SNR of the speech.

Obviously, the E_RASTA filtering can be used in speech recognition easily. Fig. 12 is the speech recognition rate of additive-filter degraded speech. The average error rate of all SNR conditions of E_RASTA reduces 35.7% compared with that of J_RATSA’s.

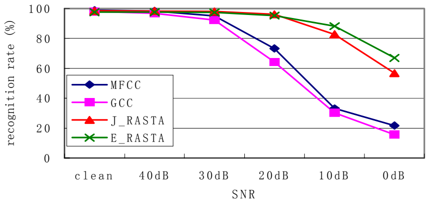

Fig 12 The average speech recognition with additive-filter degraded speech

Ⅳ Conclusion

In conclusion, we have proposed a new E_RASTA method. In our proposal, the classical RASTA filter on logarithmic spectrum domain is following by another additive RASTA filter on spectrum domain. In this manner the left additive noise is effectively removed. Experiments using DTW for isolated digit speaker identification on TI46 database suggests that the E_RASTA can outperform the classical RASTA when there is involving a combination of channel distortion and additive noise. The E_RASTA does not need the estimation of speech SNR in order to determinate the optimal value of J and multi-templates in J_RASTA, and it requires no knowledge of how the speech degrades. In addition, we found that the nonlinear compression has significant effects upon the feature extraction of speech.

References

1. J. P. Campbell, “Speaker recognition: a tutorial”, Proc. IEEE, vol. 85, no. 9, pp.1437-1462, 1997
2. L. Rabiner and B. H. Juang, “Fundamental of speech recognition”, Prentice Hall, 1993
3. S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, Proc. IEEE ASSP-27, pp. 113-120, 1979
4. R. Schwaz et al., “Comparative experiments on large vocabulary speech recognition”, Proc. ARPA Workshop Human Language Technol., (Plainsboro,NJ), 1993
5. H. Hermansky and N. Morgan, “RASTA processing of speech”, IEEE Trans. Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, 1994
6. R. D. Patterson and J. Holdsworth, “A functional model of neural activity patterns and auditory images,” Advances in speech, hearing and language processing, vol. 3, ed. W. A. Ainsworth, JAI Press, London, 1990
7. B, Xiang, “Speech feature extractor based on auditory model ant its application in speaker identification”, Master thesis of Peking University, 1998
8. B, Xiang, K. Chen and H. Chi, “Exploration of perceptually-based features for speaker recognition: an empirical study”, Chinese J. Electronics, vol. 5, no. 1, pp. 56-63, 1996
9. B. Zhen, X. Wu, Z. Liu, H. Chi, “A SNR-driven speech feature extractor”, Chinese J. Acoust. (submitted)
10. R.D. Patterson, M. Allerhand, and C. Giguere, "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," J. Acoust. Soc. Am. 98, 1890-1894, 1995