Robust Speech Recognition Features Based on Temporal Trajectory Filtering of Frequency Band Spectrum

Jia-lin Shen[1], Wen-liang Hwang[2] and Lin-shan Lee[1,2]
1. Dept. of Electrical Engineering, National Taiwan University
2. Institute of Information Science, Academia Sinica
Taipei, Taiwan, R.O.C.
e-mail : xshen@speech.ee.ntu.edu.tw

Abstract

This paper presents the use of a variety of filters in the temporal trajectories of frequency band spectrum to extract speech recognition features for environmental robustness. Three kind of filters for emphasizing the statistically important parts of speech are proposed. First, a bank of RASTA-like band-pass filters to fit the statistical peaks of modulation frequency band spectrum of speech are used. Secondly, a three-channel octave band-filter band with a smoothed rectangular window spline is applied. Thirdly, a data-driven filter is developed. Experimental results show that significant improvements for speech recognition using the proposed feature extraction approach under noisy environments can be achieved.

1 Introduction

The environmental robustness for a practical speech recognition system to real world applications is definitely very importment because mismatch between training and test environments will cause serious degradation on the speech recognition performance. A variety of approaches for robust speech recognition were proposed which can be classified into 4 categories : speech enhancement techniques, robust speech feature extraction techniques, model-based compensation approaches and robust distance measure criterions[1-4]. The RASTA (RelAtive SpecTrAl) method belonging to the second category was proposed to extract robust speech features for recognition by processing temporal trajectories of frequency band spectrum using a band-pass filter[5]. The principle of RASTA method comes from the human auditory perception which indicates the relative insensitivity of human hearing to slowly and quickly varying auditory stimuli[6]. Thus, the RASTA band-pass filter is designed with an IIR filter with a sharp spectral zero at the zero frequency in the modulation frequency domain. On the other hand, the low cut-off frequency of this band-pass filter suppresses the spectral components that change more quickly than the typical range of change of speech.

The most interesting point of RASTA method is to emphasize the important part of speech signal by human hearing perception which is definitely more immune to noise. Green[6] indicated that a greater sensitivity of human hearing to modulation frequencies around 4 Hz to lower (or higher) modulation frequencies. Instead, this paper presents the use of a variety of filters to replace the band-pass filter in RASTA method by the viewpoint of statistics of speech. First, we analyze the frequency response of temporal trajectories of frequency band spectrum for a large set of speech data to find the peak frequencies which are believed to be more insensitive to noise. Then a bank of band-pass filters with pass bands carefully adjusted to the peak frequencies of the modulation frequency band spectrum are selected to emphasize these peak frequencies. In this way, a bank of RASTA-like filters for each frequency band component are obtained by adjusting the parameters of the one-pole IIR filter in the RASTA method. Secondly, in order to systematically design the multiple band-pass filters, a three-channel octave band-filter band with a smoothed rectangular window spline is used. Thirdly, a data-driven filter based on the frequency response of speech signal in modulation frequency domain is developed.

The recognition of all the 1345 Mandarin syllables is taken as the example task for the experiments, which is the key problem for very-large-vocabulary Mandarin speech recognition. Typical channel distortion(convolutional noise) in addition to different levels of additive white Gaussian noise are included in the tests. The experimental results show that the error rates can be immediately reduced by 29.11%, 37.07%, and 29.12% for ∞, 30dB and 20dB of signal-to-noise ratio(SNR) respectively using the features processed by RASTA method and further reduced by 42.17%, 30.50%, and 19.64% respectively using the features

proposed here.

This paper is organized into 6 sections. Section 2 describes the robust feature extraction process discussed here. Section 3 analyzes the statistics of modulation frequency band spectrum of speech. In section 4, three approaches of using various filters in place of the RASTA filter are proposed. The experimental results are evaluated and discussed in section 5. Section 6 makes the concluding remarks.

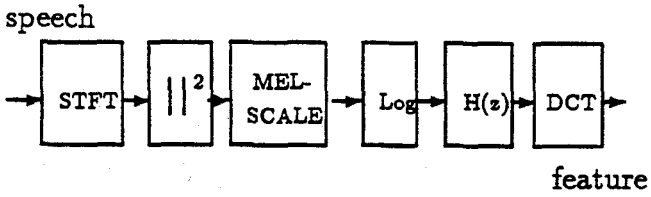

Figure 1: The block diagram of robust feature extraction process.

2 Robust Feature Extraction

The block diagram of the robust feature extraction process discussed here is plotted in Fig. 1. The input speech is first processed by short-time fourier transform (STFT) and then the corresponding power spectrum for each frame is filtered by a set of 30 triangular band-pass filters spaced uniformly on a mel-frequency scale. A filter $H(z)$ is then applied to filter the temporal trajectories of each mel-frequency band component. Finally, the discrete cosine transform (DCT) is applied for each frame to achieve the cepstral coefficient. As listed in Table 1, a variety of features with different filters $H(z)$ are compared. Apparently, as $H(z) = 1$, the output features denote the well-known mel-frequency cepstral coefficient (MFCC). The RASTA-MFCC is derived using a band-pass filter where more slowly and quickly changing parts for each spectral component are suppressed. Also, if the one-pole IIR filter of $H(z)$ in RASTA-MFCC, i.e., $\frac{1}{1-0.98z^{-1}}$ is ignored, the derivation of the MFCC features is obtained which designates the delta-MFCC[7]. Fig. 2 shows the frequency responses and impulse responses of $H(z)$ used in RASTA-MFCC and delta-MFCC respectively. It can be noted that the frequency with peak frequency response is moved to lower position by the one-pole filter of $H(z)$ in RASTA-MFCC which supports the theory of Green[6] mentioned above. In addition, when $H(z)$ is designed as a high-pass filter which makes the long-term average of spectrum of mel-frequency band identically zero, the derived MFCC features are post-processed by the well-known cepstral mean subtraction (CMS) process.

feature	$H(z)$
MFCC	1
RASTA-MFCC	$\frac{0.1z^4(2+z^{-1}-z^{-3}-2z^{-4})}{1-0.98z^{-1}}$
delta-MFCC	$0.1z^4(2+z^{-1}-z^{-3}-2z^{-4})$
MFCC with CMS	high-pass filter

Table 1: Summary of various types of features with respect to $H(z)$

3 Modulation Frequency Band Spectrum of Speech

As mentioned previously, the RASTA band-pass filter is developed to fit the peak frequency of modulation frequency band spectrum by human hearing perception. In this section, we try to analyze the modulation frequency band spectrum of speech signal in statistics. In other words, the frequencies with peak frequency responses in modulation frequency domain are thus obtained by statistical analysis instead of human hearing perception discussed by Green[6]. Here all of the Mandarin syllables which are composed by 22 INITIAL's (consonants) and 41 FINAL's (vowels but including possible medial and nasal ending) are used for analysis. The frequency responses of temporal trajectories for the mel-frequency bands are plotted in Fig. 3, where only 4 bands out of all the 30 bands are included. It can be noteworthy that similar shapes of modulation frequency band spectrum are obtained for all the mel-frequency bands where there exist two peaks at around 0 Hz and 18Hz respectively. Also, the magnitudes of modulation frequency band spectrum for lower frequencies(slow change in the mel-frequency band) are usually much larger than that for higher frequencies which indicates a cut-off frequency around 40 Hz. This characteristic is consistent with the band-pass filter used in delta-MFCC or RASTA-MFCC. However, the effect of multiple peaks in the modulation frequency response is not considered in delta-MFCC or RASTA-MFCC. Therefore, we try to use a bank of filters to replace the band-pass filter in RASTA method by emphasizing the statistically peaks of speech in modulation frequency band spectrum to improve the recognition performance.

4 A Variety of Filters

4.1 Bank of RASTA-like filters

The general form of RASTA band-pass filter shown in Table 1 can be expressed as follows :

$$\frac{\sum_{k=-N}^{N} k \cdot z^{N+k}}{\sum_{k=-N}^{N} k^2 \cdot (1-\alpha z^{-1})} \quad (1)$$

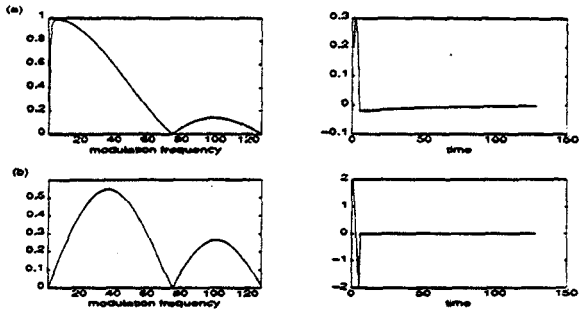

Figure 2: The frequency and impulse responses for (a)RASTA-MFCC and (b)delta-MFCC

where the order N of the FIR filter determines the cut-off frequency while the parameter α of the all-pole IIR filter can be adjusted to fit the frequency with peak spectra in the modulation frequency domain. Therefore, a bank of RASTA-like filters with parameters N and α carefully selected are used such that the designed filters can match the envelope shown in Fig. 3. Here two RASTA-like filters with different parameters α carefully adjusted to fit the peak frequency 4 Hz and 18 Hz mentioned previously are utilized. In addition, the order N in the FIR filter is chosen such that nearly 40Hz of cut-off frequency can be achieved. The derived features are called RASTA-FB-MFCC.

4.2 Three-channel Octave Band-filter Band

In the second approach, in order to systematically design the multiple band-pass filters, a three-channel octave band-filter band with a smoothed rectangular window spline is used[8]. To derive a real causal filter from the envelope of each bank of the three-channel octave filter banks, a transformation function is applied with the following form[9]:

$$H = Ae^{-i\Im log(A)} \qquad (2)$$

where A denotes the desired shape of the modulation frequency response, $\Im$ denotes the hilbert transform and H is the derived causal filter. In this way, the phase term corresponding to the envelope of the filter can be easily obtained. The derived features are called OB-MFCC.

4.3 Data-driven Filter

Instead of using the multiple band-pass filters discussed above, an interesting way to design the filter H(z) is to use directly the envelope of the spectra of modulation frequency band of speech shown in Fig. 3. Here the transformation function in eq.(2) is also applied to derive the corresponding phase term. The filter thus designed is called data-driven filter. Also, in order to disregard the steady-state channel noise such as microphone or transmission line which appears in the dc component of the modulation frequency band spectrum, the dc component of Fig.3 is set to zero.

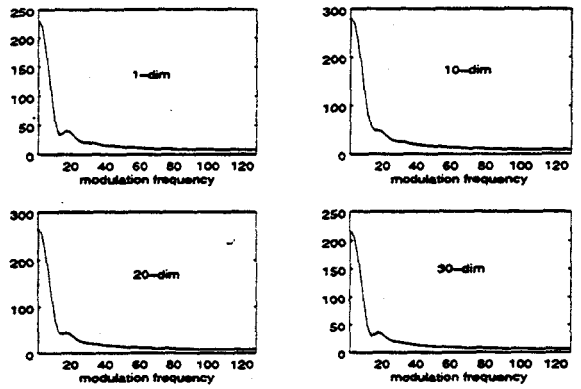

Figure 3: The frequency responses of mel-frequency band spectrum of speech signal

5 Experimental Results

The preliminary simulation experiments are performed for speaker dependent tests. A speech database produced by 3 speakers is used. For each speaker, 4 collections of all the 1345 Mandarin syllables for two types of microphones C410 and D3700 were produced respectively. C410 is a cross-talking and noise-cancelling capacitor microphone with a flat frequency response, while D3700 is a hand-held dynamic microphone. For each speaker, 3 collections of the 1345 Mandarin syllables are used for training and 1 collection is used for testing. The experimental results are average of the three speakers.

The experimental results for various types of features are shown in Table 2. In matched condition, the training and test data are produced using microphone C410, while in unmatched conditions, the test data are recorded by microphone D3700. Also, different levels of additive white Gaussian noise are included in the tests. In experiments 1-3, the experiments using the well-known MFCC features (stationary features) and delta-MFCC features (dynamic features) are evaluated. It is obvious that in matched conditions, the recognition rates using MFCC features outperform that using delta-MFCC, whereas the performance using MFCC features degrades much more seriously than that using delta-MFCC features in unmatched conditions. In experiment 4, the RASTA-MFCC features are used where the error rates can be largely reduced in unmatched conditions , especially in Top5 result-

s. Furthermore, when the RASTA-FB-MFCC features are used, the Top1(Top5) error rates are reduced by 59.01%(67.70%), 56.26%(77.38%), 43.57%(64.61%) for ∞, 30dB and 20dB of SNR respectively in unmatched conditions in comparison with the results using MFCC features. Moreover, the recognition accuracy is increased from 87.29% to 90.19% in matched conditions. However, in comparison with the results using combined MFCC with delta-MFCC features, the recognition rates are reduced by around 2% in matched condition while they can also be increased by 3-17% in unmatched conditions. Experiments 6-8 are the experimental results using the 3-channel octave band-filter band, where lower 1, 2 and 3 filters out of the filterbanks are used separately. It can be found that comparable results can be obtained with that using RASTA-MFCC features when only 1 filter is used. However, the recognition rates using 2 filters are lower than that using RASTA-FB-MFCC features. This is because the RASTA-FB-MFCC features are derived by emphasizing the most important parts of speech. Note that almost identical results are obtained when 2 and 3 filters are used. It proves the point that the high frequency parts in modulation frequency domain are intensively not important. In the last experiment, the data-driven filter is used where the performance degrades more seriously in matched condition and unmatched condition with convolutional noise only as compared to that using RASTA-MFCC features. However, better recognition rates in noisy environments with higher level of additive noise can be obtained especially in Top5 results using the data-driven filter.

6 Conclusion

In this paper, we intend to derive robust speech features for recognition by filtering the temporal trajectories of frequency band spectrum. The statistically peaks of speech in modulation frequency band spectrum are first obtained by analyzing a large set of speech data. Then a variety of filters are proposed to emphasize these statistical peaks such that significant improvements in accuracy under noisy environments can be achieved.

References

[1] O. Ghitza, "Auditory Nerve Representation as a Front-end for Speech Recognition in a Noisy Environment", *Computer Speech and language*, 1 (2); 109-130, Dec. 1986.

[2] A. Aero, "Environmental Robustness in Automatic Speech Recognition", *ICASSP*, pp. 849-852, 1990.

[3] M. J. F. Gales and S. J. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination", *Computer Speech and language*, pp. 289-307, Sep. 1995.

[4] D. Mansour and B.H. Juang, "A family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition", *IEEE Trans. ASSP*, pp. 1659-1671, Nov. 1989.

[5] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE. Trans. on Speech and Audio Processing*, Vol.2, No.4, Oct. 1994.

[6] G. Green, "Temporal Aspects of Audition", Ph.D. Thesis, Oxford, 1976.

[7] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-29, pp. 254-272, Apr. 1981.

[8] B.A. Dautrich, L.R. Rabiner, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition", *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-31 (4): 793-807, Aug. 1983.

[9] John H. Benedetto and Anthony Teolist, "A Wavelet Model and Data Compression", *Applied and Computational Harmonic Analysis*, pp. 3-28, 1993.

features	match	mismatch		
		∞	30dB	20dB
1. MFCC (14)	87.29 (98.74)	70.63 (95.17)	43.05 (75.99)	18.81 (44.76)
2. delta-MFCC (14)	81.71 (98.29)	76.51 (96.36)	58.07 (82.38)	38.22 (58.74)
3. MFCC+delta -MFCC(28)	92.04 (98.96)	84.83 (97.10)	64.25 (82.97)	36.13 (57.25)
4.RASTA-MFCC (14)	82.97 (97.84)	79.18 (97.55)	64.16 (90.86)	42.45 (72.64)
5. RASTA-FB -MFCC(28)	90.19 (98.59)	87.96 (98.44)	75.09 (94.57)	53.75 (80.45)
6. OB-MFCC(1) (14)	82.97 (97.99)	79.93 (97.32)	65.28 (90.56)	40.30 (73.09)
7. OB-MFCC(2) (28)	89.81 (98.44)	86.39 (97.84)	70.63 (91.67)	47.43 (79.18)
8. OB-MFCC(3) (42)	90.33 (98.29)	87.81 (97.70)	70.33 (92.04)	47.50 (81.21)
9. data-driven (14)	77.62 (97.47)	76.65 (97.03)	67.29 (93.16)	45.06 (81.12)

Table 2: The Top1(Top5) recognition results for experiments 1-9 using various types of features (also shown the number of feature dimensions).