Speech Feature Extraction Based on Wavelet Modulation Scale for Robust Speech Recognition

Xin Ma[1], Weidong Zhou[1], Fang Ju[1], and Qi Jiang[2]

[1] College of Information Science and Engineering, Shandong University
Jinan, Shandong, 250100, P.R. China
{max, wdzhou, jufang}@sdu.edu.cn
[2] College of Control Science and Engineering, Shandong University
Jinan, Shandong, 250100, P.R. China
jiangqi@sdu.edu.cn

Abstract. An analysis based on wavelet modulation scales feature extraction is proposed. Considering human auditory perception and varieties of disturbances, instead of the frequency differences, wavelet modulation scales are adopted to reflect the dynamic features of speech in ASR. Experiments for the Chinese digit-string recognition show extracting the wavelet modulation scales as the dynamic features have good performance both in additional noises and convolutional noises environment.

Keywords: Feature extraction, Wavelet analysis, Modulation scales.

1 Introduction

Automatic recognition of speech (ASR) has good performance in clean environment, but when speech signal is distorted by noise, the performance of ASR will degrade. Usually the environmental noises are additional and convolutional noises. To eliminate the effects of noises, some methods, such as spectrum subtraction, noise compensation are often introduced and they can effectively suppress the noisy disturbance. But when the condition of environment changes, the results of recognition will become worse.

Noise can be suppressed by improving the robustness of features. As is well known in automatic speech recognition based MFCC features [2], difference and acceleration coefficients are often adopted as auxiliary features to improve the robustness against the noises [3], and they are good dynamic features of speech. Similarly, modulation spectrum is another feature that can well reflect the dynamic feature of speech. In area of modulation spectra, the components that irrelevant to the recognition can be easily separated from the speech features [4].

Usually we can get modulation spectrum by Fourier transform, however, some other studies [5] suggest human perception for modulation accords to a constant-Q property, directly applying Fourier transform can only get uniform distribution in frequency area. To mimic this constant-Q property of human perception, in this paper, we adopt the wavelet transform to get modulation scales as speech features and use

I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 499 – 505, 2006.

them in ASR, and use normalizing technique to improve the robustness of speech features against noises. Experiments for the Chinese digit-string recognition prove these approaches have good effects for recognition rate under noisy environments.

The paper is organized as follows: in section 2, first, the theory of modulation spectrum and wavelet modulation scale features are described, then the normalizing process are presented. Experiments and analysis of results are shown in section 3, finally the conclusions are given in Section 4.

2 Modulation Spectrum and Wavelet Modulation Scale

2.1 Theory of Modulation Spectrum

The actual modulation transform is based on the spectrogram, the spectrogram can be defined as

$$|S_x^{(\gamma)}(t,\omega)=STFT_x^{(\gamma)}(t,\omega)|^2 . \tag{1}$$

It complies with principle of quadratic superposition [6], if a signal can be expressed as $x(t)=c_1x_1(t)+c_2x_2(t)$, the spectrogram of $x(t)$ can be written as

$$T_x(t,f)=|c_1|^2T_{x1}(t,f)+|c_2|^2T_{x2}(t,f)+c_1c_2^*T_{x1,x2}(t,f)+c_1c_2^*T_{x2,x1}(t,f) . \tag{2}$$

From above equation, we can see spectrogram of a signal has distinct interference terms. Modulation spectrum can be calculated from spectrogram as follows:

$$M_x(\omega,\eta)=\int_{-\infty}^{+\infty} S_x(t,\omega)e^{-j\eta t}dt . \tag{3}$$

Where ω and η are the acoustic frequency and modulation frequency respectively. $M_x(\omega,\eta)$ can also be viewed as the two-dimensional transform of the instantaneous autocorrelation function, or the correlation function of a Fourier transform $X(\omega)$ [7], but in $M_x(\omega,\eta)$ there are still interference terms ,which can be attenuated by smoothing process using proper window function. Here we use $M^{SP}(\omega,\eta)$ standing for the smoothed $M(\omega,\eta)$, that is

$$M^{sp}(\omega,\eta)=M_W(\eta,\omega)*_{\omega}M_x(\eta,\omega) . \tag{4}$$

$M^{sp}(\omega,\eta)$ is the result of the convolution of $M_x(\eta,\omega)$ and $M_W(\eta,\omega)$ in ω ,the interference terms can be reduced evidently in smoothed modulation features [8]. This conclusion is the base of modulation spectrum using in robustness improvement. The usually steps are as follows: first we frame the speech signal using short windows, the short-time fourier transform is used to acquire the spectrogram, then the spectrogram is divided into subbands in which the modulation frequency transform is performed. As the most useful components for speech recognition in modulation spectrum is between 2-16Hz [9], we select proper bands of modulation frequencies as the speech features.

2.2 Wavelet Modulation Scales

Considering of the constant-Q property of human perception for modulation, instead of fourier transform, we use the wavelet transform for every subbands and acquire the wavelet modulation scales representation. The detailed calculation of speech signal x(t) is as follows:

$$S_y(t,\omega)=\frac{1}{2\pi}|\int x(u)w^*(u-t)\ e^{-jwu}|^2 . \tag{5}$$

$S_y(t,\omega)$ is the spectrogram of $x(t)$, $w^*(t)$ is short-time window function. Along the time directions of $S_y(t,\omega)$,wavelet transform can be witten as

$$W_x(s,\zeta,\omega)=\frac{1}{s}\int S_x(t,\omega)\psi(\frac{t-\zeta}{s})dt \cdot \tag{6}$$

$\psi(t)$ is wavelet function, ζ is translation factor, W_x (s,ζ,ω) is the wavelet modulation scales representation of $x(t)$.

2.3 Modulation Scales Normalization

If a signal $x(t)$ was corrupted by additive noise $d(t)$ and convolutional noise $h(t)$, the noisy signal can be written as

$$y(t)=[x(t)+d(t)]*h(t) . \tag{7}$$

Here for convenience we let $s(t)=x(t)+d(t)$, then the spectrogram of $s(t)$ can be written as

$$S_y(t,\omega)=S_s(t,\omega)S_h(t,\omega) . \tag{8}$$

$S_s(t,\omega)$ and $S_h(t,\omega)$ are the spectrogram of $s(t)$ and $h(t)$, $S_y(t,\omega)$ is windowed along time scales and transformed by wavelet along time scales, the results are the wavelet scale representations of $y(t)$,

$$W_y(s,\zeta,\omega)=\frac{1}{s}\int S_x(t,\omega)W_L(t-B)\psi(\frac{t-\zeta}{s})dt . \tag{9}$$

$W_L(t)$ is window function used for not only avoiding the spectrum leakage but also smoothing the interference terms which was illustrated in equation (2), so here it is called smoothing window function.

If the frequency characteristic of convolutional noises can be thought as linear and time invariant over the smoothing window, we can get following approximate formula,

$$\mathrm{W_y(s,\zeta,\omega) \approx W_s(s,\zeta,\omega)W_h(\omega)}\,. \tag{10}$$

It can be normalized as

$$\mathrm{W_{y,norm}(s,\zeta,\omega)} = \frac{\mathrm{W_y(s,\zeta,\omega)}}{\int \mathrm{W_y(s,\zeta,\omega)ds}} = \frac{\mathrm{W_s(s,\zeta,\omega)W_h(\omega)}}{\int \mathrm{W_s(s,\zeta,\omega)W_h(\omega)ds}} = \frac{\mathrm{W_s(s,\zeta,\omega)}}{\int \mathrm{W_s(s,\zeta,\omega)ds}} = \mathrm{W_{s,norm}(s,\zeta,\omega)} \tag{11}$$

In actual applying the formula (9) to calculate the wavelet scales, the scale parameter s and translation factor ζ need to be discretized to $\mathrm{s_d}$ and ζ_n separately, we can write the discrete representation

$$\mathrm{W_{y,norm}(s_d,\zeta_n,\omega)} = \frac{\mathrm{W_s(s_d,\zeta_n,\omega)}}{\sum\limits_{\mathrm{s_d}} \mathrm{W_s(s_d,\zeta_n,\omega)}}\,. \tag{12}$$

Recently research about modulation spectrum manifests that the distributions of disturbances and the speech signal are different in the whole scales of modulation spectrum [3]. By select the proper scope of $\mathrm{s_d}$, interference terms made from noises can be attenuated, and formula (12) can be further approximated as

$$\mathrm{W_{y,norm}(s_d,\zeta_n,\omega)=W_{x,norm}(s_d,\zeta_n,\omega)}\,. \tag{13}$$

$\mathrm{W_{x,norm}(s_d,\zeta_n,\omega)}$ is normalized modulation scale representation of $\mathrm{x(t)}$.

3 Experiments and Analysis of Results

The speech signals was framed into 25ms (400 samples) per frame and windowed by hamming window with 8.75ms frame rate. This can acquire 128Hz sampling rate for modulation frequency. For extracting modulation scales features, the bior1.1 function was used as wavelet function. We use Mel subbands instead of uniform frequencies bands for complying the human perception. After we calculated the $\mathrm{S_x(t,\omega)}$, we need transform it to representations of the power spectrum under Mel scales. Here we divided the frequencies into 26 Mel subbands (k=26), every subband was framed and windowed by hamming window. For acquiring enough resolution, the frame should have enough length, here the 1s（128 frames）frame length was used, so every long frame include 128 short frame energy values E_n ($0 \le n < 128$). There are 2 dots overlaps because the length of bior1.1 filters are 2. Eight dyadic scales wavelet transforms was conducted to get the modulation scales vectors, the first two values of which should be discarded, like overlap-save method filtering quoted in [11]. Finally, the modulation scales features were normalized and filtered as quoted in section 2.3 of this paper. According to [1], only the third, forth, and fifth layer were saved as modulation scales parameters. So from every long frame we can acquire 3×128 wavelet scales matrix, every column in the matrix was used as parameters of corresponding short frame.

3.1 Recognition Experiments Under Clean Environments

Firstly, we used test set to perform speech recognition experiment under clean environment (no convolutional and additional noise), and assumed that both the training set and the test set are recorded under same channel conditions. The recognition errors on the test set are shown in Table 1.

Table 1. Recognition rate for clean speech

MFCC	MOD	NORM_MOD
7.92%	7.8%	8.6%

From table1, we can see that the performance of above three methods is similar under clean environment.

3.2 Recognition Experiments in Additive Noises

Noises signal n(k) superposed over the clean speech signals as additive disturbance, were extracted from NoiseX-92 database, signal-noise ratio can be determined as

$$SNR=10\log(\sum_k |s(k)|^2/\sum_k |n(k)|^2). \tag{14}$$

Table 2 shows recognition error rate of three methods for test set speech signal corrupted by white, pink, and babble noises. THE SNR of all test speech is 10dB.

Table 2. Recognition error rate of three methods for additive noisy speech

noise / method	white	Babble	pink
MFCC	28.35	32.44	36.12
Mod	20.24	23.61	25.61
Norm_Mod	22.51	27.68	26.44

From Table 2, we can see that the modulation scales features show better robustness under additive noisy environments. Normalized modulation features have good resistance to color noise.

3.3 Recognition Experiments in Convolutional Noises

The environment for a practical recognizer not only has additive noise but may have convolutional disturbance such as telephone network. For simulating the convolutional distortion to the speech, we use a telephone channel impulse response signal to corrupt the tested speech signal. The telephone channel impulse response signal was obtained from a real telephone channels, and its response feature curve was plotted in figure1.

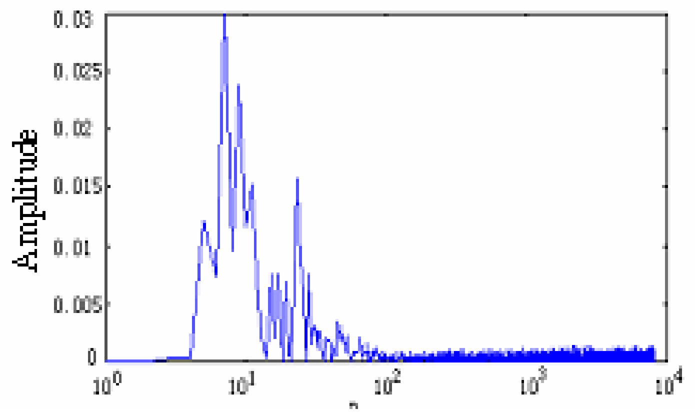

Fig. 1. Channel impulse was obtained from a real telephone channels

The recognition errors on the test set are shown in table 3.

Table 3. Recognition error rate of three methods for convolutional noisy speech

MFCC	MOD	NORM_MOD
21.55%	20.85%	10.4%

The recognition results of MFCC, modulation scales and normalized modulation scales are showed in Table 3. From Table 3, we can see the recognition result of un-normalized wavelet modulation scales features is not very good. However, after normalized, the wavelet modulation scales have good performances under convolutive environments.

4 Conclusion

Modulation spectrum is another way to reflect dynamic features of speech signals. The results of experiments for the Chinese digit-string recognitions show the new method has positive efforts in improving the robustness of speech recognition system. Further research will be done to exploit the modes and extents of its contributions for large vocabulary continuous speech recognition.

References

1. H. Hermansky, Human Speech Perception: Some Lesson From Automatic Speech Recognition (TSD'01, Zelezna Ruda, Czech Republic, Sep, 2001 in TSD'01[DB/OL], Zelezna Ruda)
2. LR.Rabiner and BH.Juang, Fundementals of Speech Recognition (Prentice Hall, Englewood Cliffs, NJ, USA, 1993:194-200)

3. S. Boll. Suppression, of acoustic noise in speech using spectral subtraction, IEEE and Signal Processing, April 1979:113-120
4. H. Hermansky, The Modulation Spectrum in Automatic Recognition of Speech IEEE Workshop on Automatic Speech Recognition and Understanding, 1997:140-147
5. E. R. Kandel, J. H. Schwartz, and T. M. Jessell, ed, Principles of Neural Science Third Edition, Chapter32 Hearing, Elsevier Science Publishing Co., Inc., 1991: 481-498
6. ZhangXian-da, Modern Signal Processing. the Tsinghua University press1995 : 456-457
7. Somsak Sukittanon, Les E. Atlas, Channel Compensation of Modulation Spectral Features, in Proceedings of the 2003 IEEE ISCAS, 2003
8. S. Sukittanon and L. E. Atlas, Modulation Frequency Features for Audio Fingerprinting Proc. of ICASSP'2002, 2002:1173-76
9. Takayuki Arai, Misha Pavel, Hynek Hermansky, and Carlos Avendano, Intelligibility of speech with filtered time trajectories of spectral envelopes Proc. ICSLP-96,Philadelphia, October 1996:2490-2493
10. http://htk.eng.cam.ac.uk/
11. Oppenheim A V, Schafer R W., Digital signal Processing Prentice Hall, Inc.,(1975) 85-86