2004 RISP International Workshop on Nonlinear
Circuit and Signal Processing (NCSP'04)
Hawaii, USA, Mar. 5-7, 2004.

Improvement of the modulation wavelet transform in ASR

Kenji Okada[1] Takayuki Arai[1] Noboru Kanedera[2] Kenji Asai[1]

[1] Dept. of Electrical and Electronics Eng., Sophia University, 7-1, Kioi-cho, Chiyoda-ku, Tokyo 102-8854 JAPAN
[2]Ishikawa National College of Technology, Tsubata-machi, Kahoku-gun, Ishikawa, JAPAN

Abstract

In this paper, we examine robust feature extraction methods for automatic speech recognition (ASR) in noise-distorted environments. Several perceptual experiments have shown that the range between 1 and 10 Hz of modulation frequency band is important for ASR. Combining the coefficients of multi-resolutional Fourier transform to split the important modulation frequency band for ASR into several bands especially increased recognition performance. We applied the wavelet transform to the feature extraction instead of multi-resolutional Fourier transform. We called this method of feature extraction "modulation wavelet transform" (MWT). The feature extraction of the previously proposed MWT covered the modulation frequency between 1 and 15 Hz. Therefore, we conducted speech recognition experiments using the MWT which covers the modulation frequency between 1 and 12 Hz by choosing the center frequencies of 2.5, 5.0, and 7.5 Hz. This new set of subbands yielded 3% increase in recognition accuracy compared to the previous results in several noise-distorted environments.

1. Introduction

In recent years the technology for automatic speech recognition (ASR) has been progressing. Still needed way is to extract feature which is effective in any kind of environment, even a noise-distorted environment. Kanedera et al. have reported that the band between 1 and 10 Hz is important for ASR[1]. They extracted the coefficients in a narrower modulation frequency band for lower frequency and a wider one for higher frequency. They called this method of feature extraction "modulation Fourier transform" (MFT)[2].

The wavelet transform allows us to carry out the MFT. We called this method of feature extraction "modulation wavelet transform" (MWT)[3]. The previously proposed MWT has extracted the modulation components between 1 and 15 Hz (MWT-0). In this study, we conducted speech recognition experiments using the MWT which covers the modulation frequency between 1 and 12 Hz (MWT-1).

We conducted speech recognition experiments using several mother wavelet with MWT. We compared the recognition accuracy of the wavelet transform with the conventional methods such as MFCC or PLP in both clean and noise-distorted environments.

The modulation wavelet which we used is described in Section 2, the experiment is described in Section 3. The result is described in Section 4.

2. Modulation wavelet

In MWT, multiple bands are extracted on the modulation frequency domain. Because wavelet transform has high-resolutional frequency characteristics for low frequencies and low-resolutional frequency characteristics for high frequencies, the wavelet transform works more effectively and efficiently than the multi-resolutional FFT.

Fig. 1 shows the frequency responses of the subbands for MWT-0. As shown in this figure, we previously covered the modulation frequency between 1 and 15 Hz.

Fig. 2 shows the frequency responses of the subbbands for MWT-1. In the current study, we cover the modulation frequency between 1 and 12 Hz.

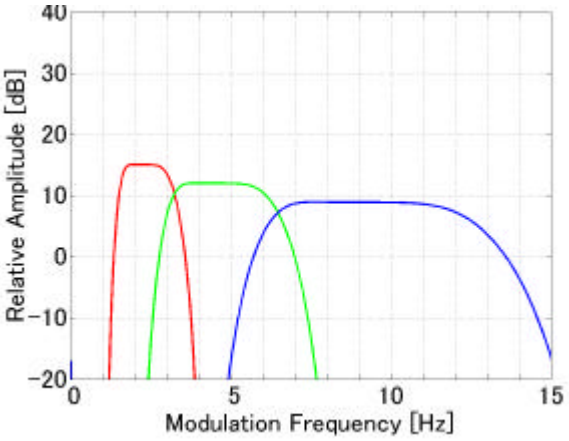

Fig. 1: Frequency responses of the three subbands (MWT-0).

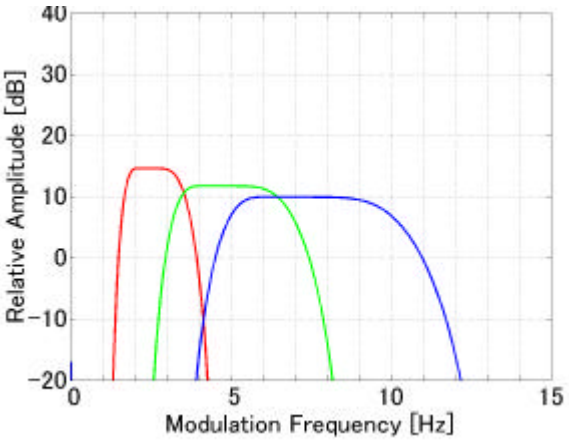

Fig. 2: Frequency responses of the three subbands (MWT-1)

3. Experimental setup

We conducted speech recognition experiments using the time trajectories of PLP coefficients[4]. The conditions are shown in Table 1. The HMM ToolKit (HTK)[5] was used to train for six states and two mixture components per state. We used a set of noise in the NOISEX-92 database[6]. The test data were degraded by additive noise (SNR 10dB). The center frequencies of subbands used in this study were listed in Table 2.

We used a set of noise (babble, buccaneer1, buccaneer2, destroyerengine, destroyerops, f16,factory1, factory2, hfchannel, leopard, m109, machinegun, pink, volvo, white) in the NOISEX-92 database. The test data were degraded by additive noise (SNR 10dB).

Table1: Conditions of ASR experiments

Task	Bellcore digit Database (0-9, oh, yes, no) 200 speakers, 13 words in each speaker
Sampling frequency	8 kHz
Frame period	25 ms
Window length	10 ms
Training	150 speakers (75 males and 75 females)
Test	50 speakers (25 males and 25 females)

Table 2: Center frequencies of subbands in the modulation frequency

Number of subbands	Center-frequency [Hz]
2	2.5 7.5
3	2.5 5 7.5
4	2.5 4.2 5.8 7.5
5	2.5 3.75 5 6.75 7.5

4. Experimental results

4.1. Comparing modulation wavelet and conventional methods

We conducted ASR experiments with the features extracted by MWT-0 and MWT-1. The modulation wavelet divided the important modulation frequency range (about 1 to 10 Hz) linearlly into 2, 3, 4, and 5 bands. For comparison, we also conducted experiments using MFCC + delta, PLP + delta as conventional method, MWT-0, and MWT-1.

The experiments were carried out in a clean environment and in a noise-distorted environment. In the noise-distorted environment we used all types of NOISEX-92 noise. The types of mother wavelets, we used, were 'Meyer', 'Morlet', 'Bior3.7'. 'Bior3.7' was 'Biorthogonal spline wavelets.'

The results are shown in Table 3 and Fig. 3. In clean environment conventional methods such as 'MFCC + delta', 'PLP + delta' gave a smaller error rate than MWT-1 method. In noise-distorted environments,

however, the tendency was oppsite; in other words, MWT-1 yielded better performance (e.g., WER was 14.8% for 3 bands).

Table 3: Comparison between conventional approach, MWT-0, and MWT-1 (Word error rate [%])

	clean	noise
MFCC + delta	1.65	20.6
PLP + delta	1.42	25.4
MWT-0	3.6	17.8
MWT-1 (2 bands)	5.2	18.7
MWT-1 (3 bands)	3.9	14.8
MWT-1 (4 bands)	3.7	14.4
MWT-1 (5 bands)	3.6	14.3

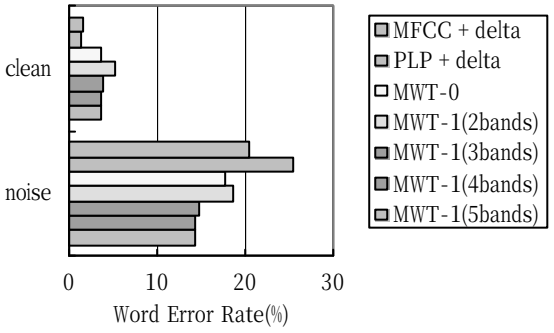

Fig. 3: Comparison between conventional approach, MWT-0, and MWT-1

These results indicate that MWT-1 outperforms MWT-0, and furthermore, MWT-1 with 3 or more bands performed better than MWT-1 with two bands. The differences among 3, 4, and 5 bands in MWT-1 were small.

Because smaller feature dimension needs smaller computational resources, we will use MWT-1 with 3 bands in the next section on comparison among different mother wavelets.

4.2. Mother wavelets

The experiments are carried out in a clean environment and in a noise-distorted environment. We used 'Meyer', 'Morlet', 'Bior3.7' mother wavelet. In the noise-distorted environment we used all types of NOISEX-92 noise.

Table 4: Comparison among several types of mother wavelet (Word error rate [%])

	clean	noise
Meyer (MWT-1 3bands)	3.8	14.1
Morlet (MWT-1 3bands)	4.5	15.0
Bior3.7 (MWT-1 3bands)	3.5	15.3

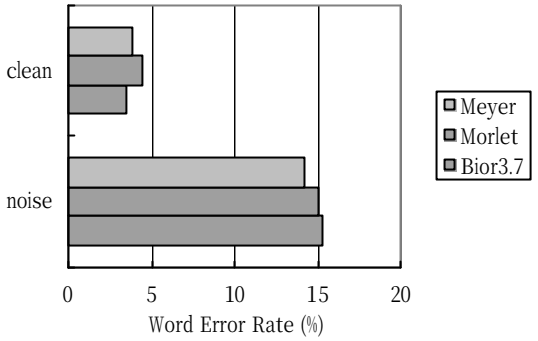

Fig. 4: Comparison between several types of mother wavelet

The results are shown in Table 4 and Fig. 4. In clean environment conventional methods such as 'MFCC + delta', 'PLP + delta' gave a smaller error rate than MWT-1. 'Meyer' used as mother wavelet gave 14.1% error rate. This rate is better than 'Morlet' or 'Bior3.7' used as mother wavelet.

This result indicates that 'Meyer' dividing 3 bands gave the best recognition accuracy under noise-distorted environment. We conjectured that this was due to the character of the 'Meyer' mother wavelet and the number of subbands.

5. Conclusions

For feature extraction in ASR we examined robust feature extraction methods. We compared the improved modulation wavelet (MWT-1) with previously proposed modulation wavelet. The improved modulation wavelet gave better recognition than previously proposed approach.

References

[1] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, “On the importance of various modulation frequencies for speech recognition." *Proc. of Eurospeech,* pp. 1079 - 1082, 1997.

[2] N. Kanedera, T. Arai, H Hermansky, M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition." *Speech Communication* Vol. 28, pp. 43 - 55, 1999.

[3] K. Okada, T. Arai, N. Kanedera, Y. Momomura and Y. Murahara, "Using the modulation wavelet transform for feature extraction in automatic speech recognition," *ICSLP 2000*, Vol. 1, pp. 337 - 340, 2000.

[4] N. Kanedera, H. Hermansky and T. Arai, "On properties of modulation spectrum for robust automatic speech recognition," *Proc. IEEE ICASSP*, pp. II-613 - II-616, 1998.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "*The HTK Book,''* Ver. 2.2, Entropic, 1999.

[6] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, Vol. 12, No. 3, pp. 247 - 251, 1993.

[7] T. Arai, M. Pavel, H Hermansky, C Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, Vol. 105, No.5, pp. 2738 - 2791, 1999.