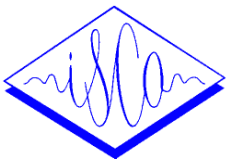

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

USING THE MODULATION WAVELET TRANSFORM FOR FEATURE EXTRACTION IN AUTOMATIC SPEECH RECOGNITION

Kenji Okada[1], *Takayuki Arai*[1], *Noboru Kanedera*[2], *Yasunori Momomura*[1], *and Yuji Murahara*[1]

[1] Dept. of Electrical and Electronics Engr., Sophia University
7–1 Kioi-cho, Chiyoda-ku, Tokyo, 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 16 Hz of modulation frequency band is important for human speech recognition. Furthermore it has been reported the same 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. Combining coefficients of a multi-resolutional Fourier transform corresponds to a wavelet transform. To test the effectiveness and efficiency of the wavelet transform, we, therefore, applied the wavelet transform to recognition experiments. This approach yielded an average of 3% increase in recognition accuracy compared to the standard approach using mel-frequency cepstral coefficients (MFCC) 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. Arai et al. [1] conducted perceptual experiments for syllable intelligibility. The result indicated that the range between 1 and 16 Hz of modulation frequency band is important for speech recognition.

Further, Kanedera et al. [2] have reported that the band between 1 and 16 Hz modulation frequency, especially between 2 and 10 Hz, is important for ASR. For feature extraction Kanedera et al. [3] calculated for the time trajectories of perceptual linear prediction (PLP) coefficients two levels of resolution coefficient of fast Fourier transform (FFT) using high-resolution and low-resolution FFTs. In their paper they reported that extracting coefficients from the high-resolutional corresponds to a modulation frequency band around 2.5 Hz and, around 5 and 7.5 Hz from the low-resolutional FFT coefficients increased the recognition accuracy. It yields a narrower modulation frequency band for lower frequency and a wider one for higher frequency. We call this method of feature extraction "modulation Fourier transform" (modulation FT).

The wavelet transform allows us to carry out the modulation FT. We compared the recognition accuracy of the wavelet transform with the standard approach using 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

Discrete Fourier transform (DFT) has sine and cosine as its base functions. The discrete cosine transform (DCT) has only the cosine as its base function. Whereas the DFT-based approach divides the modulation frequency linearly, the wavelet transform divides the modulation frequency band non-linearly in an effective and efficient way. In the wavelet transform we can define a suitable base function which can well express the signal. The base function is called the 'mother wavelet.'

In the wavelet transform the mother wavelet is elastic length called 'scale.' The mother wavelet is shifted by a value called 'translate'. The combination of 'scale' and 'translate' express the signal.

The general formula for on wavelet transform is:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)$$

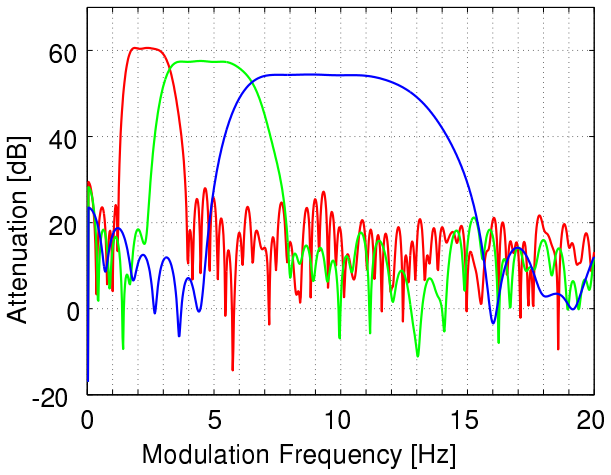

Figure 1: Dividing the modulation frequency band into 3 parts(meyer).

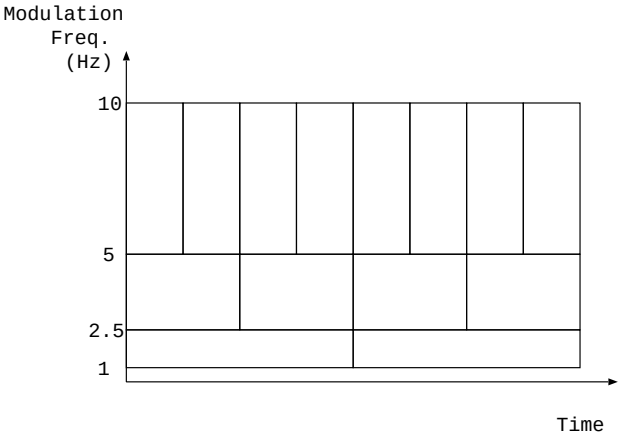

Figure 2: Dividing into the multi-resolutional band in logarithmically into 3 parts.

This formula was translated by the mother wavelet $\psi(\frac{x-b}{a})$. In this formula a is 'scale' and b is 'translate'.

We divide the modulation frequency band using the wavelet transform. In this study, we used scales that divide the important modulation frequency band logarithmically into two to five segments.

Fig. 1 shows an example of dividing the modulation frequency band in logarithmically into three parts. The wavelet transform has a high-resolutional frequency characteristic for low frequency and a low-resolutional frequency characteristic for high frequency. Thus the wavelet transform works more effectively and efficiently than the multi-resolutional FFT(Fig. 2).

Table 1: Conditions of ASR experiments

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

3. EXPERIMENTAL SETUP

We conducted speech recognition experiments using the wavelet transform to extract important modulation frequency bands for the time trajectories of PLP coefficients. We used PLP coefficients because the feature filtered time trajectories of PLP coefficients outperformed that of MFCC [4]. The conditions are shown in Table 1.

The scales were selected to divide logarithmically the range for 2 to 10 Hz of modulation frequency logarithmically as shown in Table 2. The HMM ToolKit (HTK [6]) was used to train for six states and two mixture components per state.

We used a set of noise (babble, buccaneer1, buccaneer2, destroyerengine, destroyerops, f16, factory1, factory2, hfcannel, leopard, m109, machinegun, pink, volvo, white) in the NOISEX-92 database [5]. The test data were degraded by additive noise (SNR 10dB).

At first we conducted ASR experiments totally splitting into 2, 3, 4, 5 bands. In the recognition experiment of modulation wavelet, the type of mother wavelet used was 'meyer'. The type of noise was 'babble' noise.

Next, we conducted the experiments applying several types of the mother wavelets. The types of mother wavelets were 'mexican hat', 'haar', and 'meyer'. Three bands were used to divide the important modulation-frequency band.

Table 2: The scales for divided modulation-frequency bands

The number of divided modulation-frequency bands	scales
2	16 8
3	32 16 8
4	64 32 16 8
5	128 64 32 16 8

4. EXPERIMENTAL RESULTS

4.1. Comparing modulation wavelet and standard methods

We conducted ASR experiments to compare modulation wavelet method with standard methods in a clean environment and in noise-distorted environment. In the noise-distorted environment 'babble' noise was used. For comparison, we also conducted experiments using MFCC + delta, PLP + delta, and modulation FT. In the recognition experiment using modulation wavelet, the type of mother modulation wavelet used was 'meyer'. The modulation wavelet divided the important modulation frequency range (about 2 to 10 Hz) logarithmically into 2, 3, 4, and 5 bands.

The results are shown in Table 3 and Figure 3. Under the babble noise environment, the modulation wavelet divided into 3 bands gave a smaller error rate than conventional feature extraction methods such as MFCC + delta, PLP + delta, and modulation FT. The modulation wavelet method gave 17.9% error rate. This rate is better than that of any other standard methods (MFCC + delta, PLP + delta, and modulation FT). For the modulation wavelet, we used the modulation bands around 2, 4, and 8 Hz on the center frequency, while we used the modulation bands centered at 2.5, 5, and 7.5 Hz for the modulation FT. The results indicates that the modulation wavelet outperforms modulation FT methods in a noisy environment. These results also indicates that three bands are necessary in the modulation frequency domain.

4.2. Mother wavelets

We applied several types of mother wavelets to 2-band and 3-band division in clean and 'babble' noise-distorted environments. The types of mother wavelets, we used, were 'morlet', 'mexican hat' , 'haar', and 'biorthologna l (3.7)'. The effect of the mother wavelets is shown in Tables 4, 5, and Figure 4.

Table 3: Comparison between modulation wavelet and standard approach(Word error rate [%]).

	clean	babble noise
MFCC + delta	1.65	21.5
PLP + delta	1.42	27.7
modulation FT	1.61	18.6
modulation wavelet (2 bands)	4.6	21.7
modulation wavelet (3 bands)	3.6	17.9
modulation wavelet (4 bands)	5.0	28.1
modulation wavelet (5 bands)	7.3	36.3

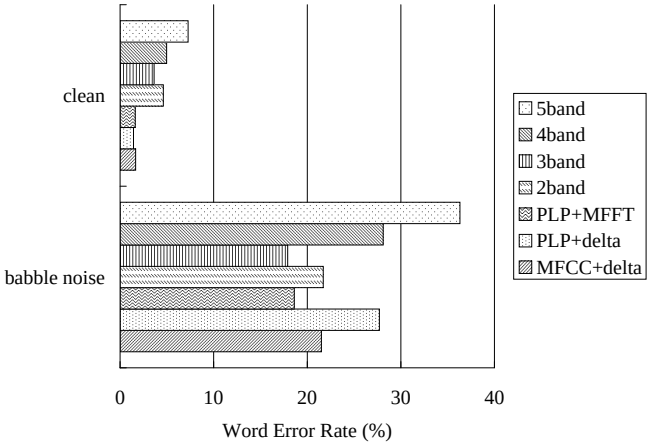

Figure 3: The standard approach and modulation wavelet.

Further there was a mother wavelet which gave a better recognition accuracy than the 'meyer' wavelet gave in a clean environment. But no type of mother wavelet gave a better recognition accuracy than "meyer" in a 'babble' noise-distorted environment. We conjectured that this was due to the character of the 'meyer' mother wavelet and the 'babble' noise.

4.3. Various noise environments

We applied several types of noise to the modulation wavelet. The types of mother wavelets were 'mexican hat', 'haar', and 'meyer.' Three bands were used to divide the important modulation-freqency band. These results are shown in Table 5. This approach yielded an average of 3% increase in recognition experiments.

Table 4: Word error rates by wavelet transform with several mother wavelets. (the case of dividing into 2 bands).

	2 bands		3 bands	
	clean	babble	clean	babble
meyer	4.61	21.7	3.6	17.9
morlet	23.4	57.3	6.7	28.3
mexican hat	3.1	22.8	5.0	33.5
haar	2.4	23.2	2.0	25.0
biorthologna1 (3.7)	15.9	50.2	5.0	27.3

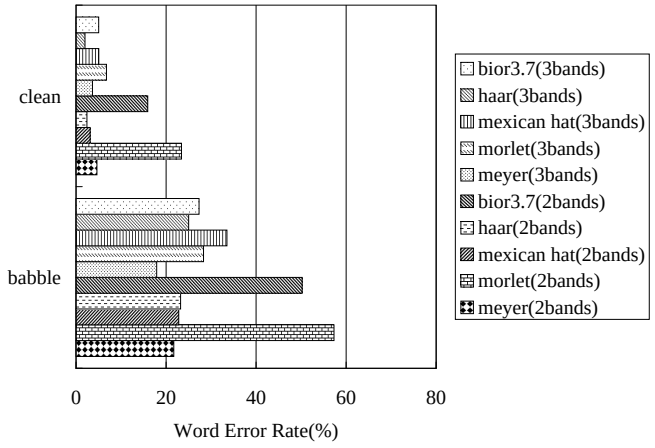

Figure 4: The modulation wavelet with several wavelet(2, 3bands)

5. CONCLUSIONS

For feature extraction in ASR we examined robust feature extraction methods. We compared the modulation wavelet with standard methods such as MFCC, PLP and modulation FT. The new method gave better recognition than the standard approach.

6. ACKNOWLEDGEMENTS

A part of this research was supported by Japan Science and Technology Corporation under Regional Science Promotion Program.

7. REFERENCES

[1] T. Arai, Misha Pavel, Hynek Hermansky, Carlos Avendano, "'Syllable intelligibility for temporally filtered LPC cepstral trajectories." *J. Acoust. Soc. Am.*, Vol. **105**, No.5, pp 2738 – 2791, 1999.

[2] N. Kanedera, Takayuki Arai, Hynek Hermansky and Misha Pavel, "On the importance of various modulation frequencies for speech recognition." *Proc. of Eurospeech*, pp 1079–1082, 1997.

[3] N. Kanedera, Takayuki Arai, Hynek Hermansky, Misha Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition." *Speech Communication* **28**, pp. 43–55, 1999.

[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] 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.

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

Table 5: Comparing modulation wavelet('meyer', 'mexican hat', 'haar') and MFCC with several types of noise(Word error rate [%]).

	Modulation Wavelet			
noise	meyer	mexican hat	haar	MFCC
babble	17.9	33.5	25.0	21.5
buccaneer1	19.8	21.1	53.4	21.7
buccaneer2	19.0	21.6	18.7	21.8
destroyerengine	16.9	21.0	38.5	19.0
destroyerops	16.8	28.9	65.1	16.9
f16	17.0	23.1	17.6	21.5
factory1	18.6	23.5	18.8	20.9
factory2	13.1	22.7	14.0	16.0
hfchannel	15.8	16.7	13.6	23.1
leopard	13.4	33.5	17.5	15.5
m109	14.6	26.2	15.6	15.8
machinegun	41.5	49.3	50.8	50.2
pink	16.2	16.4	14.1	19.0
volvo	9.5	21.4	10.6	7.0
white	17.3	16.4	13.7	19.6
mean	17.8	25.0	25.8	20.6