Using the Modulation Complex Wavelet Transform for Feature Extraction in Automatic Speech Recognition

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

[1] Dept. of Electrical and Electronics Eng., Sophia University
7–1 Kioi-cho, Chiyoda-ku, Tokyo, JAPAN
http://www.splab.ee.sophia.ac.jp/
[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. Previous research showed that combining the coefficients of multi-resolutional modulation frequency band. We show that this multi-resolutional approach can be achieved using a wavelet transform instead of the Fourier transform. Taking the FFT phase into consideration, we applied the Gabor function, which is a complex function, as mother wavelet. This approach yielded a 1.7% increase in recognition accuracy compared to the FFT-based multi-resolutional approach.

1. Introduction

Feature extraction is essential for a robust ASR system. Recently modifications to conventional discrete Fourier transform (DFT) cepstral processing have taken into account psychophysical processes of human hearing. For example, the mel-frequency cepstral coefficients (MFCCs) incorporate a mel-scaled filterbank. Similarly, the perceptual linear predictive (PLP) analysis [1] incorporates a set of critical-band filters, and takes into account equal loudness pre-emphasis and intensity-loudness conversion.

Current ASR systems also take into account speech dynamics, such as delta (Δ) processing and RASTA (RelAtive SpecTrAl) processing [2], which yield more robust speech features for ASR. Low modulation frequency components (less than 1 Hz) of cepstral coefficients are especially affected by convolutional distortion. For human speech perception, on the other hand, those low modulation frequency components are not used. Arai et al. [3] conducted a study on syllable intelligibility tests focusing on the modutaion frequencies for human speech perception. The results indicated that the range between 1 and 16 Hz of the modulation frequency band is most used for human speech recognition.

Furthermore, Kanedera et al. [4] have reported that the band between 1 and 16 Hz modulation frequency, especially between 2 and 10 Hz, is useful to increase ASR performance. For feature extraction Kanedera et al. [5] proposed a modulation Fourier transform (MFT). In this method time trajectories of the PLP coefficients are extracted and their spectral components are computed by applying high-resolution and low-resolution FFTs to achieve two levels of resolution. The modulation frequency bands were centered at 2.5 Hz, 5.0 Hz and 7.5 Hz for the ASR experiments. The experimental results showed that recognition accuracy increased.

Okada et al. [6] substituted a wavelet transform for FFT in a similar multi-resolutional approach. This method of feature extraction will be referred to as a modulation wavelet transform (MWT). In their study the recognition performance was better than for the MFT in noise-distorted environments.

Recently, Kanedera et al. [7] conducted an ASR experiment showing that it is important to use phase information of the modulation spectrum as well as the magnitude. In the current study we use a complex function as the mother wavelet for time trajectories of the PLP coefficients in the MWT to take the modulation phase into account and examine how this affects ASR performance. This method of feature extraction will be referred to as the modulation complex wavelet transform (MCWT).

In Section 2 the MCWT is described in greater detail. The ASR experiment and results are presented in Section 3 and discussed in Section 4.

2. Modulation Complex Wavelet Transform

2.1. Modulation spectral analysis

A modulation spectrum is obtained by applying the Fourier transform to time trajectories of either spectral components, cepstral coefficients, or PLP coefficients. The modulation spectrum represents temporal patterns of speech dynamics, and the axis of abscissas is called the modulation frequency. It has been reported that the shapes of modulation spectra are not language dependent [8] and the peak of the modulation spectrum corresponds to the most sensitive modulation frequency region of the human auditory system [9].

2.2. MCWT for ASR

In a wavelet transform the basis function is called the mother wavelet. The mother wavelet has an elastic length called *scale* and is shifted by a value called *translate*. The combination of *scale* and *translate* can express a signal. The wavelet transform has high resolution frequency characteristics at low frequencies and high resolution time characteristics at high frequencies. Thus, the wavelet transform works more effectively and efficiently than the multi-resolutional FFT approach.

The MWT extracts several coefficients from the modulation frequency bands by applying different scale parameters of the wavelet transform. In this study we used Meyer, Haar, Mexican-hat, Morlet, Biorthogonal (3.7) as mother wavelets. These mother wavelets do not take phase components into consideration. Therefore, the Gabor function, which is a complex function, is used as a mother wavelet.

2.3. Gabor function

The exponential function, $e^{j\omega t}$, used in the Fourier transform as a basis function has an infinitely wide range in the time domain. In other words, applying the Fourier transform results a loss of time information from a signal. The window function $w(t)$ was used in the function $w(t)e^{-j\omega t}$ to gain temporal resolution. This is known as a short time Fourier transform (STFT).

Gabor suggested using the Gauss function $e^{-(t/\sigma)^2}$ as the window function. Thus, the STFT of the Gabor function is expressed as

$$\hat{f}(\omega, b) = \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi}\sigma} e^{-\frac{(t-b)^2}{\sigma^2}} e^{-j\omega t} f(t) dt, \qquad (1)$$

where b is *translate*. The general formula for the Gabor function as a mother wavelet is

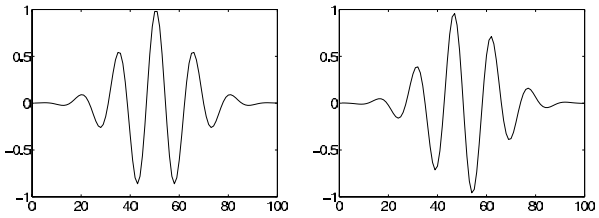
$$\psi(t) = \frac{1}{2\sqrt{\pi}\sigma} e^{-(\frac{t}{\sigma})^2} e^{-jt}. \qquad (2)$$

In this function the parameter σ controls the effective length of the window. Fig. 1 shows the corresponding Gabor function.

3. Experimental Setup

We conducted three speech recognition experiments using the MCWT to extract important modulation frequency bands for time trajectories of the PLP coefficients. We used PLP coefficients because the filtered time trajectories of the PLP coefficients outperformed that of the MFCC [4]. A 9th-order PLP analysis was used. The conditions are shown in Table 1.

The HMM ToolKit (HTK) [10] was used to train for six states, with two mixture components per state. We used



(a) real part. (b) imaginary part.

Figure 1: Gabor function.

babble noise from the NOISEX-92 database [11]. The test data were degraded by additive noise (babble, SNR 10dB).

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)

3.1. Experiment 1

We conducted the first experiment in order to see the effect of using a complex mother wavelet instead of only using a real one as in the previous studies both in a clean and a noise-distorted environment. We divided the modulation frequency band into 3 bands using the wavelet transform whose mother wavelet is the Gabor function. The scale parameters were selected so that the center modulation frequencies became 2.5 Hz, 5.0 Hz and 7.5 Hz. The scales are shown in Table 2. The three cases of the parameter σ in Eq. (2) was used; that is, 4, 8 and 16. These numbers were empirically selected by doing a preliminary experiment. Fig. 2 shows frequency characteristics of the wavelet transform with $\sigma = 8$.

Table 2: The scales used in Exp. 1 (center modulation frequencies are 2.5, 5.0 and 7.5 Hz).

# Bands	Scales
3	26.5 12.5 8.5

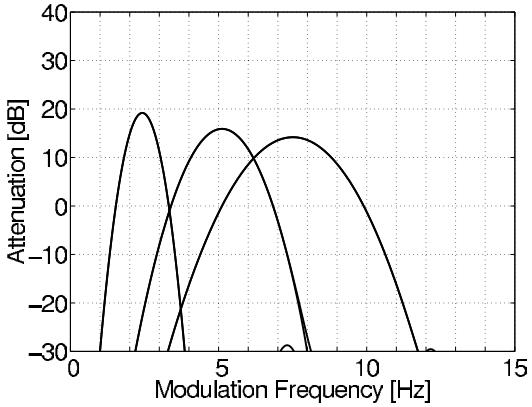


Figure 2: Division of the modulation frequency band into 3 bands (σ = 8).

3.2. Experiment 2

Next, the scales were selected to divide the important modulation frequency range (2 to 10 Hz), as is shown in Table 3. We conducted the second experiment to limit the maximum modulation frequency to 10 Hz.

Table 3: Exp. 2: The scales used in Exp. 2 (the maximum modulation frequency was limited to 10 Hz).

# Bands	Scales
2	28 10.5
3	28 17 10.5
4	28 22 15 10.5
5	28 22 18 13 10.5

Table 4: The scales used in Exp. 3 (center modulation frequencies are 2.5, 5.0 and 7.5 Hz).

Window function	Scales
Hanning	26 13 8.5
Blackman	26 13 8.5
Chebyshev	22 13 8.5
Kaiser	26 13 8.5
Alternative	26 13 8.5

3.3. Experiment 3

We tested the use of several window functions as possible alternatives to the Gauss function. The center modulation frequencies used in Experiment 1 (2.5 Hz, 5.0 Hz and 7.5 Hz) were also used in Experiment 3. The scales are shown in Table 4.

The alternative window in Table 4 was approximated by the following function:

$$w[n] = 0.21 + 0.37\cos\left(\frac{2\pi n}{N-1}\right) + 0.25\cos\left(\frac{4\pi n}{N-1}\right) + 0.12\cos\left(\frac{6\pi n}{N-1}\right) + 0.04\cos\left(\frac{8\pi n}{N-1}\right) + 0.01\cos\left(\frac{10\pi n}{N-1}\right)$$

Table 5: Summary of results for Experiment 1 (%).

σ		clean	babble noise
4	real	98.4	80.3
	complex	98.7	83.8
8	real	96.6	77.8
	complex	96.4	79.3
16	real	93.9	73.5
	complex	93.8	76.1

Table 6: Summary of results for Experiment 2 (%).

# Bands		σ=4	σ=8	σ=16
2	clean	98.2	95.2	91.8
	babble noise	79.9	76.5	72.3
3	clean	98.2	96.1	94.1
	babble noise	80.4	81.0	78.4
4	clean	98.0	95.7	93.8
	babble noise	79.3	80.2	78.3
5	clean	98.2	95.8	93.6
	babble noise	79.7	80.1	78.3

4. Results and Discussions

4.1. Exp. 1: Effect of complex function

The results from experiment 1 are shown in Table 5. Under the babble noise environment the complex function yielded better recognition accuracy than when only using the real part. This was the case for σ equals 4, 8, 16. This result indicates that the complex components are valuable for improving the ASR accuracy. The overall recognition accuracy was best when the parameter σ equals to 4.

4.2. Exp. 2: Effect of number of bands

The results for Experiment 2 are shown in Table 6. The recognition accuracies with 3 bands were always better than the other number of bands. The overall recognition accuracy was best when the parameter σ equals to 4.

Comparing Table 6 with Table 5, we see that the best recognition accuracy was obtained when 3 bands are centered at 2.5 Hz, 5.0 Hz and 7.5 Hz, respectively. This result is consistent with that of Kanedera et al. [4]. This suggests that including the modulation frequency components of 10 Hz is crucial.

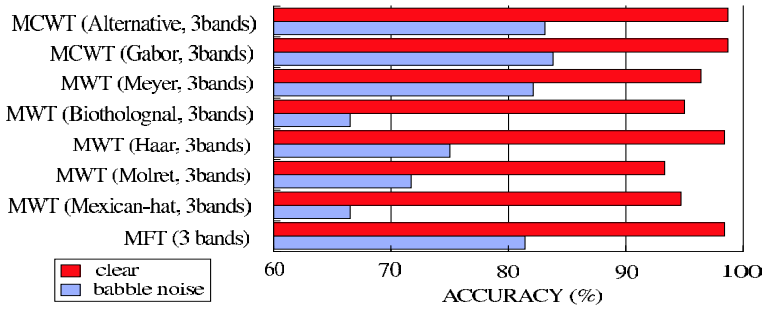


Figure 3: Comparison of the results with the previous studies.

Table 7: Summary of results for Experiment 3 (%).

Window function		Accuracy
Hanning	clean	95.7
	babble noise	78.6
Blackman	clean	96.4
	babble noise	79.5
Chebyshev	clean	97.0
	babble noise	80.1
Kaiser	clean	94.9
	babble noise	77.2
Alternative	clean	98.7
	babble noise	83.1

4.3. Exp. 3: Effect of window functions

The results from Experiment 3 are shown in Table 7. When we used the alternative window, the recognition accuracy was better than the other windows with the exception of the Gauss window used in Exp. 1.

5. Summary

In this study we examined a robust feature extraction method for ASR. We applied the MCWT using the Gabor function as a mother wavelet to time trajectories of PLP coefficients to take speech dynamics into consideration.

Fig. 3 shows a comparison of the methods used in the current study (MCWT) and previous studies. This approach yielded a 1.7% increase in recognition accuracy compared to the MWT (Meyer, 3bands) in noise distorted environments. Our results indicate that the proposed method using complex wavelets improves speech recognition accuracy and that modifying the shape of a window function might contribute to future improvements in robust speech recognition.

6. Acknowledgments

We would like to acknowledge Prof. Dawn M. Behne of the Norwegian University of Science and Technology for useful comments on this study.

7. References

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, Vol.87, No.4, pp.1738–1753, 1990.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, Vol.2, No.4, pp.578–589, 1994.

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

[4] 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.

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

[6] K. Okada, T. Arai, N. Kanedera, Y. Momomura and Y. Murahara, "Using the modulation wavelet transform for feature extrantion in automatic speech recognition," *Proc. of ICSLP*, Vol. I, pp.337–340, 2000.

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

[8] T. Arai and S. Greenberg, "The temporal properties of spoken Japanese are similar to those of English," *Proc. of Eurospeech*, Vol.2, pp.1011–1014, 1997.

[9] S. Greenberg, "Understanding speech understanding: towards a unified theory of speech perception," *Proc. of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Spech perception*, W. A. Ainsworth and S. Greenberg (eds.), Keele University, UK, pp.1–8, 1996.

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

[11] A. Vaiga 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.