SPEECH ENHANCEMENT WITH REDUCTION OF NOISE COMPONENTS IN THE WAVELET DOMAIN

Jong Won Seok and Keun Sung Bae

School of Electronic and Electrical Engineering
Kyungpook National University, Taegu, Korea
won@mmirl3.kyungpook.ac.kr ksbae@ee.kyungpook.ac.kr

ABSTRACT

This paper describes a general problem of removing additive background noise from the noisy speech in the wavelet domain. A semisoft thresholding is used to remove noise components from the wavelet coefficients of noisy speech. To prevent the quality degradation of the unvoiced sounds during the denoising process, the unvoiced region is classified first and then thresholding is applied in a different way. Experimental results demonstrate that the speech enhancement algorithm using the wavelet transform is very promising.

1. INTRODUCTION

Degradation of the quality of speech caused by the acoustic background noise is common in most of speech processing applications such as mobile communication and speech recognition. Therefore, the problem of removing uncorrelated noise components from the noisy speech, i.e., speech enhancement, has been widely studied in the past and it is still remained as an important issue in the field of speech research.

Wavelet theory provides a unified framework for a number of techniques which has been developed independently for the various signal processing applications[1]. In particular, the wavelet transform is of interest for the analysis of nonstationary signal such as speech, sonar signal, etc. Recently a novel approach for noise reduction using the wavelet transform has been proposed by [2]. It employs the thresholding in the wavelet domain and has shown to have very broad asymptotic near-optimal properties for a wide class of signals corrupted by additive white Gaussian noise.

In this paper, we apply the thresholding technique in a wavelet domain to the speech signal to reduce the noise components while keeping the important information of speech. There should be, however, some considerations applying the thresholding method directly to speech signal since the speech signal in the unvoiced region contains relatively lots of high frequency components that can be eliminated during the thresholding process. Experimental results will be discussed with our findings.

2. WAVELET TRANSFORM

Wavelet theory is based on generating a set of filter by dilation and translation of a generating wavelet. All of the wavelets are scaled versions of the "mother wavelet". This requires that only one filter is designed and others will follow the scaling rules in both the time and frequency domain. The continuous wavelet transform is defined as

$$CWT(a,\tau) = \int f(t)\ \varphi^*_{a,\tau}\, dt \qquad (1)$$

$$\varphi_{a,\tau} = \frac{1}{\sqrt{a}}\ \varphi\left(\frac{t-\tau}{a}\right) \qquad (2)$$

where a and τ are real, and $\varphi(t)$ is the wavelet basis, i.e., mother wavelet. The wavelets are contracted($a<1$) or dilated($a>1$), and are moved over the signal to be analyzed by time shift τ. Contraction and dilation scale the frequency response to allow the set of wavelets to span the desired frequency range. At high frequency region of the signal, better time-resolution can be obtained using the contracted version of wavelet. On the other hand, better frequency-resolution can be obtained at low frequency range using the dilated version of wavelet. This kind of analysis is particularly useful for speech signal

processing where better time-resolution is needed at high frequency range to detect the rapid changing transient factor of the signal, while better frequency resolution is needed at low frequency range to track the slowly time-varying formants more precisely.

The discrete wavelet transform is given by

$$d_{j,k} = \int f(t)\ \varphi^*_{j,k}(t)\ dt \qquad (3)$$

$$\varphi_{j,k}(t) = a_0^{\frac{-j}{2}}\ \varphi(a_0^{-j}t - kT) \qquad (4)$$

If we set $a_0 \approx 1$ and sampling interval T small, it can be seen that equation (4) is the approximation of equation (2). Of particular interest is the discretization on a dyadic grid, which occurs for $a_0 = 2$. The dyadic form of the discrete wavelet transform is given as

$$d_{j,k} = \frac{1}{\sqrt{2^j}} \int f(t)\ \varphi^*\left(\frac{t}{2^j} - kT\right)\ dt \qquad (5)$$

The set of wavelets can be considered as a filter bank for speech analysis. Figure 1 shows a tree structured filter bank corresponding to the dyadic form of discrete wavelet transform on sequences. The halfband lowpass and highpass filters are $h_0(n)$ and $h_1(n)$, respectively, and $\downarrow 2$ means subsampling by the factor of 2.

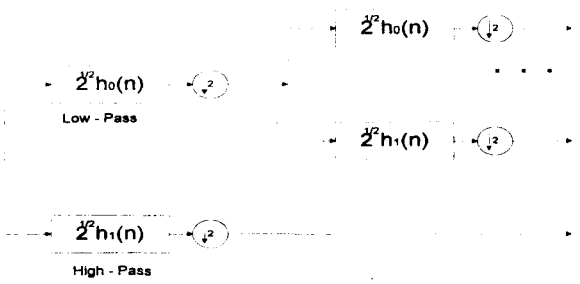

Figure 1. Dyadic octave analysis filter bank for discrete wavelet transform

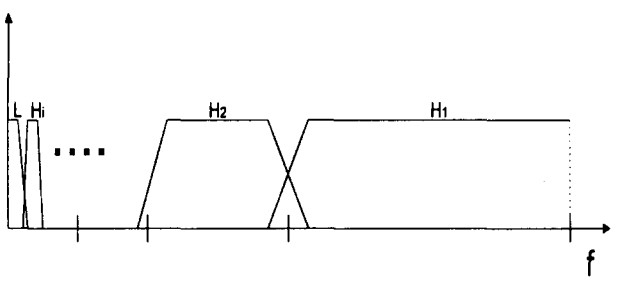

Figure 2. Frequency scheme of an octave filter bank

In the first stage, the spectrum of the input signal is split into an upper halfband and a lower halfband. The lower halfband is then split into two halves in the next stage successively. Figure 2 shows the resulting spectral characteristics of the filer bank given in Figure 1.

3. NOISE REDUCTION WITH SEMISOFT THRESHOLDING

Removing noise components by thresholding the wavelet coefficients[2] is based on the observation that limited number of wavelet coefficients in the lower bands are sufficient to reconstruct the original signal.

Let y be a finite length observation sequence of the signal x which is corrupted by zero mean, white Gaussian noise n with standard deviation ε.

$$y = x + \varepsilon n \qquad (6)$$

The goal is to recover the signal x from the noisy observations y. Let W denote a wavelet transform matrix for discrete wavelet transform. Then equation (7) can be written in the wavelet domain as

$$Y = X + N \qquad (7)$$

where capital letters indicate variables in the transformed domain, i.e., $Y = Wy$. If we assume the orthogonal wavelet transform, then inverse transform matrix M exists and can be written as

$$MW = I \qquad (8)$$

where I represents the identity matrix.

Let $\hat{X}$ denote an estimate of X based on Y which is noisy signal in the wavelet domain. Then the estimate $\hat{x}$ can be obtained from

$$\hat{x} = M\hat{X} = M\hat{Y} \qquad (9)$$

where $\hat{Y}$ represents the wavelet coefficients after applying the thresholding method. The estimate $\hat{X}$ is obtained by simply shrinking or killing the individual wavelet coefficients. The principle is that noise contributes to the most of coefficients but main feature of original signal contributes to only a few coefficients in the lower bands. Hence, by setting the smaller coefficients to zero, we can nearly optimally eliminate noise while preserving the important information of original signal.

For thresholding in the wavelet domain, we use a semisoft threshold function[3] that showed the advantages over hard and soft threshold function with respect to variance and bias of the estimated value. The semisoft threshold function is given by

$$THR(Y) = \begin{cases} 0 & |Y| \le \lambda_1 \\ sgn(Y)\left(\frac{\lambda_2|Y|-\lambda_1}{\lambda_2-\lambda_1}\right), & \lambda_1 < |Y| \le \lambda_2 \\ Y, & |Y| > \lambda_2 \end{cases} \quad (10)$$

where $THR(Y)$ represents the output value after thresholding the wavelet coefficients and λ_1 and λ_2 denote lower and upper threshold, respectively.

In applying the thresholding method to speech signal, it is important not to harm the unvoiced sound in the speech signal. Since the unvoiced sound contains lots of noise-like high frequency components, eliminating them in the wavelet domain can cause severe degradation of intelligibility in the reconstructed signal. Therefore, we first separate unvoiced region from the noisy speech, and then apply the thresholding method in a different way from other regions.

For separation of unvoiced region, the distribution of wavelet coefficients in each band is examined. First we divide input speech into four different bands according to Figure 2, and the average energy of each band in the wavelet domain is calculated. The input speech segment is then classified as the unvoiced sound if the following two conditions are satisfied:(1)The highest band energy in the wavelet domain is greater than that of other bands. (2)The ratio of lowest band energy and highest band energy is less than 0.9.

The proposed speech enhancement algorithm is shown in Figure 3, and is summarized as follows

i) Compute the discrete wavelet transform(DWT).
ii) Normalize the wavelet coefficients with the highest frequency band(H_1 in Figure 2) energy.
iii) Determine unvoiced region observing the distribution of wavelet coefficients.
iv) If determined as unvoiced region, thresholding the wavelet coefficients of the highest frequency band only. Otherwise thresholding all the wavelet coefficients.
v) Compute the inverse DWT.

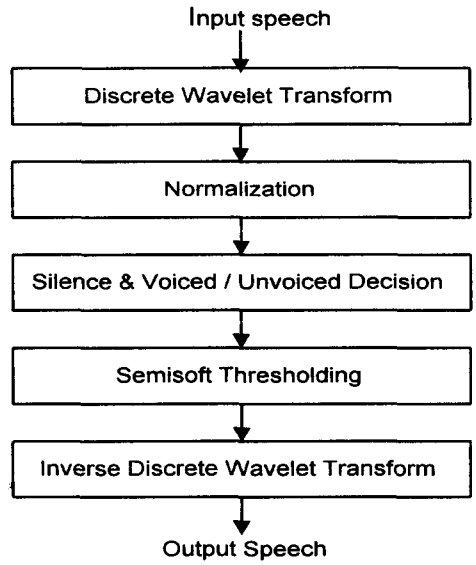

Figure 3. Block diagram of the speech enhancement algorithm

4. EXPERIMENTAL RESULTS

The proposed speech enhancement algorithm has been tested on the spoken Korean and English passages. The speech signals corrupted by additive white Gaussian noise with various global signal to noise ratio(SNR) from 10dB to -10dB were used for performance evaluation. Speech signal has the 8 ㎑ sampling rate and 256 sample of frame size for analysis. The Daubechies(D8) filter and pyramid algorithm were used to get the wavelet transformed signal. Thresholding value λ_1 was determined by [2], and λ_2 was set $\sqrt{2}\lambda_1$ empirically.

Figure 4 shows the clean speech waveform of voiced region, noisy speech of 10dB SNR and enhanced speech signal. The LPC spectra for the speech signal given in Figure 4 are shown in Figure 5. From these figures, it can be seen that the enhanced speech becomes very close to the original clean speech both in time domain and frequency domain.

For the quantitative evaluation, we measured the average signal-to-noise ratios and cepstral distance of the noisy and enhanced speech with respect to the clean speech. Table 1 shows the results for the English sentence of "Should we chase those cowboys?" At the input SNR of 0dB, the processed speech showed about 7dB improvement. The Korean sentence also showed similar results. Figures 6 and 7

show noisy and enhanced speech waveform for English and Korean sentences, respectively.

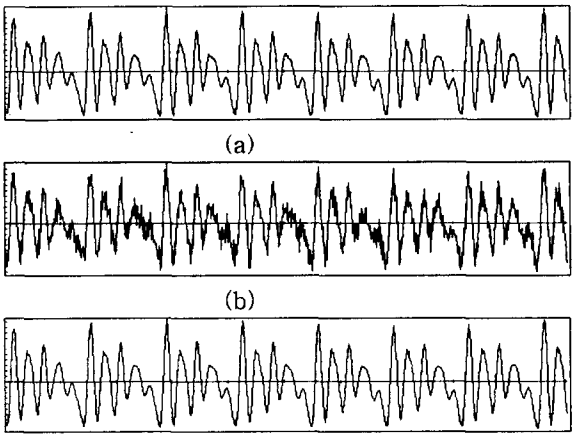

Figure 4. Speech waveform of voiced sound
(a) original speech (b) noisy speech(SNR=10dB)
(c) enhanced speech

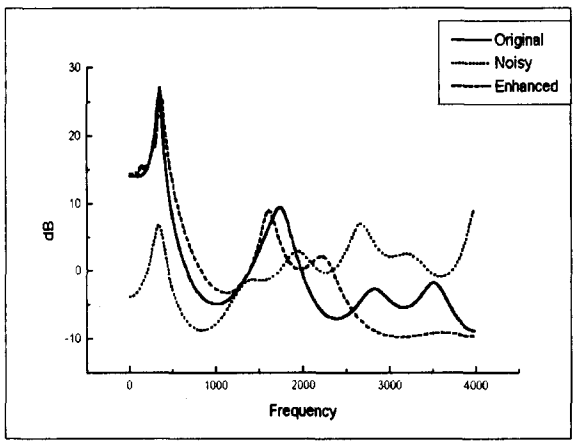

Figure 5. LPC spectrum of voiced sound

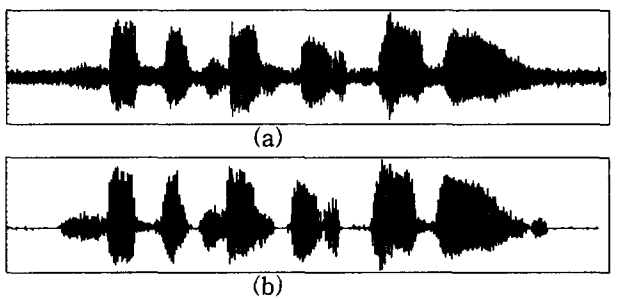

Figure 6. Speech waveform of English sentence
(a) noisy speech(SNR=10dB) ("Should we ....")
(b) enhanced speech

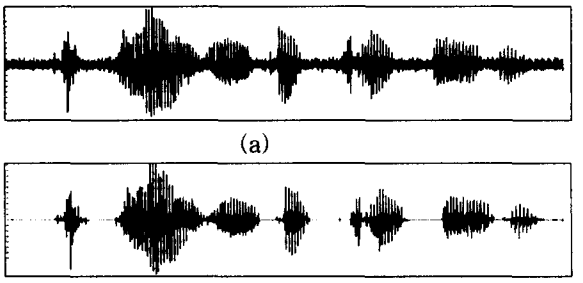

Figure 7. Speech waveform of Korean sentence
(a) noisy speech(SNR=10dB)
(b) enhanced speech

Table 1. Comparison of SNR and cepstral distance

SNR(dB)		Cepstral Distance	
Noisy	Enhanced	Noisy	Enhanced
-10	0.93	3.78	2.09
- 5	3.42	3.02	1.87
0	7.12	2.87	1.55
5	11.34	2.68	1.32
10	13.92	2.34	1.03

5. CONCLUSION

A denoising approach in the wavelet domain has been investigated for speech enhancement. A semisoft threshold function was used to remove noise components from the wavelet coefficients of noisy speech signal. In order to preserve the high frequency components of unvoiced sounds during the denoising process, thresholding was applied only to the highest band in the wavelet domain for unvoiced sounds. Experimental results showed considerable improvements in the SNR as well as the spectral distance. Considering the simplicity of the proposed speech enhancement algorithm, the results have proven to be very promising.

REFERENCE

[1] O. Rioul and M. Vetterli, "Wavelet and Signal Processing", IEEE Signal Processing Magazine, pp. 14-38, 1991.

[2] D. L. Donoho. "Denosing by soft thresholding", IEEE Trans. on Information Theory, pp. 961-1005, 1994.

[3] H. Y. Gao "WaveShrink with Semisoft Shrinkage", Research Report of MathSoft, September, 1995.

[4] I. Daubechies, Ten Lectures on Wavelets, SIAM, 1992.

[5] M. Vetterli, "Multi-dimensional Sub-band Coding," IEEE Signal Processing, Vol. 6, No. 2, pp. 97-112, 1984.