Speech Noise Cancellation using Time Adaptive Threshold Value in Wavelet Transform

Chul-Hee Lee*, Ki-Hoon Lee*, Hyang-Ja Hwang**, In-Seob Moon*** and Chong-Kyo Kim*

* Dept. of Electronic Engineering, Chonbuk National University, Jeonju City, 561-756, Korea
Tel : +81-063-272-1177 Fax : +81-063-276-0928 E-mail: chlee@ssplab.chonbuk.ac.kr
**Radioactive Materials Regulation Dept. 2 of Radiation Safety Center, Korea Institute of Nuclear Safety, Daejeon City, 305-338, Korea
Tel : +81- 042-868-0656 E-mail: hhjbluesky@hotmail.com
***Dept. of Information & Communication Engineering, Chosun College of Science & Technology, Kwangju City, 501-759, Korea
Tel : +81-062-263-8336 E-mail: mis@mail.chosun-c.ac.kr

Abstract:

This paper proposes a new noise cancellation method for speech recognition in noise environments. We determine the time adaptive threshold value using standard deviations of wavelet coefficients after wavelet transform by frames. The time adaptive threshold value is set up by using sum of standard deviations of wavelet coefficients in cA3 and weighted cD1. cA3 coefficients represent the voiced sound with lower frequency components and cD1 coefficients represent the unvoiced sound with higher frequency components. In experiments, we removed noise after adding white Gaussian noise and colored noise to original speech. The proposed method improved SNR and MSE more than wavelet transform and wavelet packet transform does. As a result of speech recognition experiment using noise speech DB, recognition performance is improved by 2~4 %.

Keywords: Keywords : wavelet & wavelet packet transforms, soft threshold, noise cancellation, speech signal

1. INTRODUCTION

In speech recognition system, noise is one of the most seriously affected factors for reducing the recognition performance. Therefore, noise cancellation technique is very important to improve speech recognition perfor- mance.

Recently, many research projects have widely studied wavelet transform which offers variable mother functions to analysis as the components of frequency. Multi-resolu- tion features of wavelet transform by using variable mother functions and localization in time-frequency domain are useful for analysis of signals with statistical features unknown or unpredictable non-stationary signals [3][4]. In general noise canceling method, when we use wavelet transform for noisy signal, noise component for each scale is much less than speech signal components. Therefore, it is needed to use threshold value for removing the noise signal which is below threshold value.

In this paper, we proposed a method which is adopting threshold value for one of the most important noise canceling techniques by using wavelet transform. We proposed optimal time adaptive threshold value after implementing wavelet transform, by using low frequency and high frequency components which represent voiced region and unvoiced region respectively. In simulation, we used additive noisy speech that white noise and colored noise are added to original speech signal for noise canceling test. We conducted speech recognition experiment using noisy speech DB to apply practical speech recognition system. And we used SNR (Signal to Noise Ratio) and MSE (Mean Squared Error) for evaluation of performance and comparing between conventional wavelet transform and wavelet packet transform. This paper consists of 5 chapters. In chapter 2, we discussed noise canceling techniques using wavelet transform. Chapter 3 describes proposed time adaptive threshold value, we showed the simulation method and results in chapter 4, and finally we conclude in chapter 5.

2. NOISE CANCELLATION METHOD USING WAVELET TRANSFORM

The basic concept of thresholding method, when we perform wavelet transform to noisy speech corrupted by white Gaussian noise, is a noise canceling by using proper threshold value. Since noise components of each scale are relatively less than speech components, we remove the components below optimal threshold value and re- synthesize each scale. Then, we reduced efficiently noise signal from the noisy speech signal. The threshold value λ is determined by each level scale after we select mother wavelet and get coefficients of wavelet transform from noisy speech. We obtain the noise canceled speech after removing the components below threshold value of wavelet coefficients by soft thresholding.

We show the soft threshold value in eq. (1). The threshold value λ is used in eq. (2) for wavelet transform and eq. (3) for wavelet packet transform.

$$T_{soft}(X)=\begin{cases}\mathrm{sgn}(X)(|X|-\lambda), & |X|\geq\lambda\\ 0 & ,\ |X|<\lambda\end{cases} \quad (1)$$

$$\lambda=\sigma\sqrt{2\log(N)} \quad (2)$$

$$\lambda=\sigma\sqrt{2\log(N\log_2 N)} \quad (3)$$

where X is noisy speech signal, N is number of samples and σ is average deviation of selected wavelet coefficients. Average deviation is calculated by using eq. (4) which uses median value of wavelet coefficients.

$$\sigma=\mathrm{MAD}/0.6745 \quad (4)$$

The quantity MAD is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution.

3. TIME ADAPTIVE THRESHOLD VALUE

From the result of wavelet transform, we have different wavelet coefficients and features between speech signal region and noise signal region. The standard deviation has a small value in noise signal region, but it has a comparatively large value in speech signal region. In wavelet transform, for smaller scale, it is more precisely detected in rapidly changed region. Hence, the standard deviation of wavelet transform coefficients becomes small. We assumed that there exists only noise signal in beginning and ending parts of noisy speech [2].

We divided noisy speech signal into frames, performed wavelet transform for each frame. and created time adaptive threshold value by using cD1 and cA3 under noisy environment. The plosive, fricative, and affricate sounds in the noisy speech have relatively small energy in comparison with voiced region and large energy in high frequency region in frequency domain. And the voiced region has more energy in low frequency region. The frame scale standard deviations of cD1 and cA3 are represented by C_{D1}^k and C_{A3}^k respectively, where k is the number of frames.

In this paper, we found that we can use C_{A3}^k as the time adaptive threshold value because of the similarity of speech signal "□□□□□" and output signal of C_{A3}^k as shown in Fig. 1.

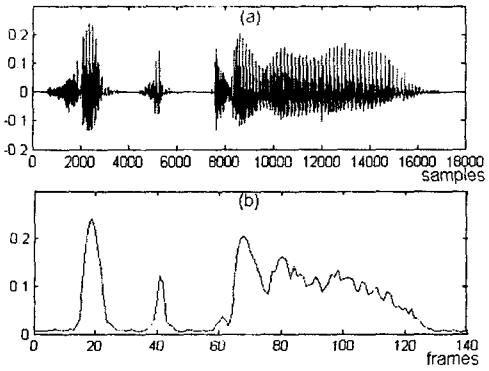

Fig. 1. (a) Original speech wave (b) C_{A3}^k wave.

However the plosive and fricative sounds with higher frequency components in speech wave can not represent C_{A3}^k useful for lower frequency components. Therefore, we can almost perfectly represent speech wave by using the linear combination of C_{D1}^k and C^k which contains the information of plosive and fricative sounds.

By using these properties, we defined parameters for detecting speech signal in wavelet domain as eq. (5) [2].

$$T^k=C_{A3}^k+\alpha C_{D1}^k \qquad k=1,\ldots,N \quad (5)$$

In eq. (5), k is the number of frames. The '6' is used as the weight value α in [2] to detect speech boundaries.

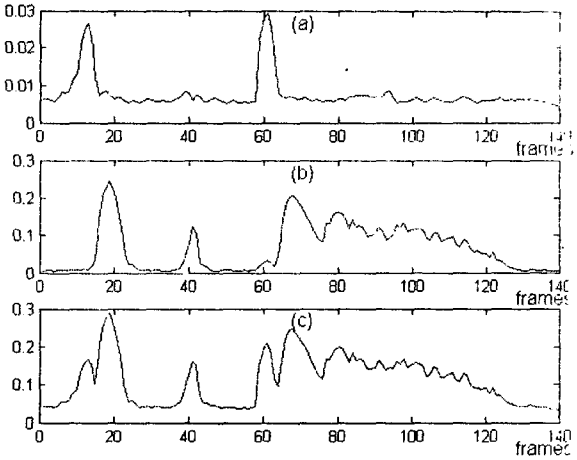

Fig. 2. (a) C_{D1}^k wave (b) C_{A3}^k wave (c) T^k wave.

The graph of T^k with α =6 is shown in Fig. 2. We found that the base line of T^k is increased from 0.01 to 0.06 by C_{D1}^k multiplied by '6' and added to C_{A3}^k. This increment can be used for end point detection of speech as described in [2]. But, in this paper, we performed following procedures to utilize the value of T^k as time adaptive threshold value.

1) Calculate the value of α by using the maximum ratio of C_{D1}^k and C_{A3}^k.

$$\alpha=\frac{\max(C_{A3}^k)}{\max(C_{C1}^k)} \quad (6)$$

2) After finding the value of T^k using α in eq. (5) and find the average value of T^k to remove base line then subtract the average value from C_{D1}^k (Fig. 3(a)).

$$T^{k'}=T^k-\beta\times mean(T^k) \quad (7)$$

where β is an adjustment parameter.

3) Normalize the maximum value of parameter to be '1' (Fig. 3(b)).

$$T^{k''} = T^{k'} \times \frac{1}{\max(T^{k'})} \quad (8)$$

4) Time adaptive threshold value is obtained for re-sampling (Fig. 3(c)). That is, because $T^{k''}$ is not the sample number of original signal but the number of frames, we can calculate the time threshold value by re-sampling. In this simulation, re-sampling is conducted by using re-sampling function in Matlab.

5) Use the time adaptive threshold value multiplied by λ from eq. (2) by using parameter $T^{k''}$ (Fig. 3(d)).

$$\lambda^{k'} = \delta\lambda(1 - T^{k''}) \quad (9)$$

where δ is an adjustment parameter.

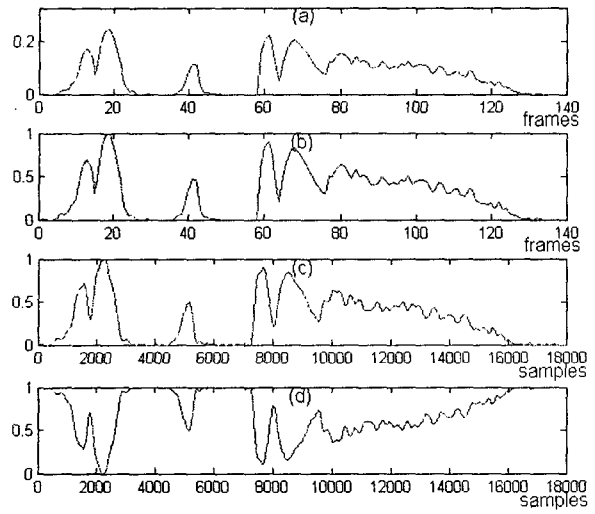

Fig. 3. (a) $T^{k'}$ wave (b) $T^{k''}$ wave (c) re-sampling (d) $\lambda^{k'}$ wave.

6) Finally, remove the noise by using soft thresholding by comparing the time adaptive threshold value of noisy speech signal.

4. EXPERIMENT AND RESULT

4.1. Experiment method

We performed an experiment under two kind of noisy speeches. Firstly, experiment-1 carried out removing noise from the noisy signal which white Gaussian noise and colored noise are added to original speech. Secondly, experiment-2 conducted speech recognition experiment using noisy speech DB collected from driving car.

In experiment-1, speech data for pronouncing "□□□□□" and "□□□" is sampled by 16kHz and quantized by 16bits. Noise signals are white Gaussian noise from 0dB to 20dB and colored noises : babble, f16, factory and pink noise in NOISEX-92 noise DB. Noisy speech is made by adding original speech to such noises. In order to evaluate objectively, we used SNR and MSE as following eq.s (10) and (11) [1].

$$SNR(\delta) = 10\log\frac{\sum_i \hat{s}_i^2}{\sum_i n_i^2} \quad (10)$$

$$MSE(\delta) = \frac{\sum_{i=0}^{N}(s_i - \hat{s}_i)^2}{N} \quad (11)$$

where s = original speech, n = noise, $\hat{s}$ = estimate of original speech, and N = number of samples.

In experiment-2, we performed speech recognition experiment using noisy speech DB collected for 10 words of car commands under driving car at 80km/h.

4.2. Results of experiment

It is important that how determine each frame size and how determine wavelet mother functions. In this paper, we use 256 samples per frame and symlets 3tap as a wavelet mother function.

4.2.1. Result of experiment-1

In experiment-1, we performed noise reduction from noisy speech with white Gaussian and colored noises added to original speech. Fig. 4 shows noise reduction result of noisy speech "□□□□□" with 5dB white Gaussian noise added.

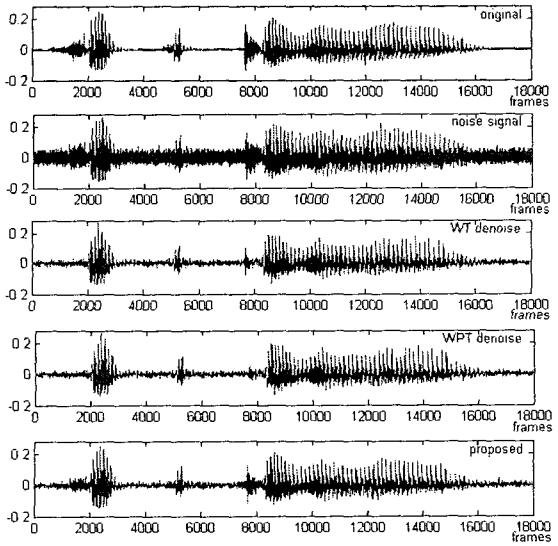

Fig. 4. Noise cancellation of noisy speech with 5dB white Gaussian noise added to "□□□□□" speech.

In WT and WPT, a fricative sound '□'(1800 samples duration) and a plosive sound '□'(8000 samples duration) are not separated from noise efficiently. However, by the proposed method, noise can be efficiently removed by classifying noise and speech in fricative and plosive sound region. Tables 1 and 2 show SNR and MSE of each result. It is found that the performance is improved comparing with conventional WT and WPT.

Table 1 SNR for "□□□□□" speech.

Noise scale	WT	WPT	Proposed
0 dB	6.8495	6.7832	6.258
5 dB	9.3705	9.0316	9.825
10 dB	11.3744	10.1672	12.880
15 dB	13.110	10.571	15.832
20 dB	15.011	10.701	18.440

Table 2 MSE for "□□□□□" speech.

Noise scale	WT	WPT	Proposed
0 dB	2.331	2.367	2.672
5 dB	1.305	1.411	1.177
10 dB	0.822	1.086	0.583
15 dB	0.551	0.989	0.296
20 dB	0.355	0.960	0.162

Fig. 5 is a result of noise reduction from "□□□" with f16 noise of NOISEX-92 DB added. Tables 3 and 4 represent SNR and MSE of each noise data that are babble, f16, factory and pink noise.

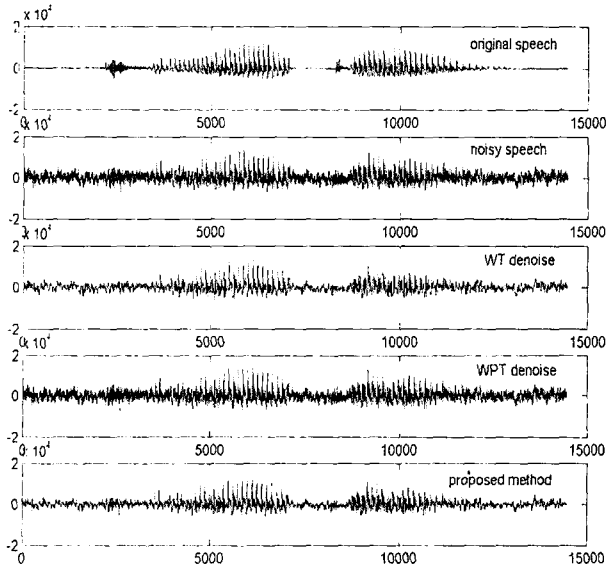

Fig. 5. Noise cancellation of noisy speech with f16 noise added to "□□□" speech.

Table 3 SNR for "□□□" speech.

Noise	Noisy speech	WT	WPT	Proposed
babble	2.99	3.17	3.07	3.43
f16	3.94	3.95	3.99	4.67
factory	8.85	9.00	9.06	10.34
pink	2.61	2.73	2.84	3.00

Table 4 MSE for "□□□" speech.

Noise	Noisy speech	WT	WPT	Proposed
babble	1.40	1.89	1.89	1.26
f16	1.13	1.78	2.27	0.95
factory	0.36	0.38	0.73	0.25
pink	3.20	4.99	6.74	3.12

4.2.1. Result of experiment-2

In experiment-2, we use 10 words of car commands collected from driving car at 80km/h. Noisy speech data is sampled by 11.025 kHz, and quantized by 16bits resolution. Noisy speech data consists of 210 words pronounced by 21 speakers. 16 speakers' data are used for training and 5 speakers' data are used for testing. We conducted speech recognition experiment using HTK ver. 3.2.1. Table 5 represents pronounced word list.

Table 5 Pronounced word list

	Pronounced words
1	□□□ □
2	□□□ □
3	□□ □□
4	□□ □□
5	□□ □□□
6	□□□ □□□
7	□□□ □□
8	□□□ □□
9	□□ □□
10	□□ □□

Table 6 represents speech recognition result for noisy speech using WT, WPT and the proposed method. The proposed method gives an improvement of about 4% comparing noisy speech data, and about 2% rather than WPT. The proposed method demonstrates the efficiency of speech recognition.

Table 6 Speech recognition ratio using noisy speech DB(%).

	Noisy speech	WT	WPT	Proposed
recognition ratio	76.67	73.33	78.34	80.00

5. CONCLUSION

In this paper, we proposed a noise canceling method for recognition of noisy speech, by using time adaptive threshold value in wavelet transform which is performed for noisy speech signal by frames. The threshold value is calculated by using standard deviation of wavelet coefficients. Considering the characteristics of speech signal, we set the threshold value in detail signal of the first scale for unvoiced speech with higher frequency components and in approximation signal of the third scale for voiced signal with lower frequency components.

We use two types of noisy speech DB's to reduce noise using proposed time adaptive threshold value. One is white Gaussian noise from 0dB to 20 dB and another is colored noises : babble, f16, factory and pink noise in NOISEX-92 for addition to original speech. As a result of experiment, we found that WT and WPT severely reduced speech component for plosive, fricative, and affricate sounds. However, in the proposed method, original speech signal is almost restored. For removing noise by proposed time adaptive threshold for noisy speech with white Gaussian and colored noises added, SNR and MSE are improved more than WT and WPT. And the recognition performance is improved by about 2~4% comparing conventional WT and WPT.

References

[1] □□□, □□□, □□□, "□□ □□□□ □□ □□□□□ □□□□ □□□□ □□□ □□□□", □□□□□□□ □□□, □4□, □4□, pp. 306-315, 2001.

[2] □□□, "Wavelet Transform-Based Speech Signal Processing: Speech Enhancement and Endpoint Detection", □□□□□□, □□□□□, 1999.

[3] Daubechies, I., "The Wavelet transform time frequency localization and signal analysis" IEEE.Tran. on information theory , vol. 36, no. 5, pp. 961-1005, 1990.

[4] Michel Misiti, Yves Mistis, Georges Oppenheim, and Jean- Michel Poggi, *Wavelet Toolboox for Use with MATLAB, The Math Work Inc*, 2001.

[5] Bahoura, M., Fouat, J., "Wavelet speech enhancement based on the teager energy operator," IEEE Signal Process. Lett. 8 (1), pp. 10-12, 2001.

[6] Chang, S., Kwon, Y., Yang S., Kim I., "Speech Enhance- ment for Non-stationary Noise Environment by adaptive Wavelet Packet," IEEE Trans. ASSP, vol. 1, pp. I-561-I-564, 2002.