NOISE REDUCTION OF SPEECH BY WAVELET THRESHOLDING

Hunze Chen *and* ***Chung-Hsien Wu***

Department of Computer Science and Information Engineering,

National Cheng Kung University, Tainan, Taiwan, R.O.C.

Email: {chenht, chwu}@csie.ncku.edu.tw

ABSTRACT

The noise reduction of speech performed in wavelet domain is rather seldom used. The Fourier transform performs well upon linear time-invariant signal processing and is appropriate for many applications like stationary signal processing. Nevertheless it cannot deal with many transient characteristics of which speech signals possess. The wavelet expansion gives a more accurate local description and separation of signal characteristics. Wavelets can adjust themselves well to suit the signal and reveal its properties. Besides its time complexity is only $O(N)$. Though there are some theories regarding to the wavelet shrinkage, most of them focus on artificial signal or 2-D image signal processing. The determination of the threshold value for the wavelet shrinkage of speech signals is either unsuitable or lacks theoretical foundation. This paper will find out this value from the view of mean-squared-error perspective, and derive it step by step mathematically. Once the threshold could be set accurately, the noise reduction of speech through wavelet shrinkage could be successfully done and it might not alter the characteristics of the speech or distort the speech itself. Experimental results show that our approach gives encouraging results in noise reduction for speech signals.

I. INTRODUCTION

Digital signal processing of noisy speech in a background of noise is important as more applications of speech recording or recognition can not always be performed in a soundproof booth or made in an anechoic chamber. Speech recognition rate has dropped owing to the corrupted speech signals. As new technologies develop, voice will dominate to be the input media of the computer in the future. Just like in the films "star war" or "knight rider", talking to a computer will definitely be possible and get a fair response.

Speech signal processing is usually applied in the frequency domain through Discrete Fourier Transform (DFT). The time complexity of Fast Fourier Transform (FFT) is $O(N \log N)$, where N is the sample size, and is often a large number as the sample rate might be 8k, 22050, or 44100 Hz etc. Such algorithms cost CPU time for even seconds of speech signal.

There are some researches using Wavelet Transform (WT) techniques concerning noise removal either in one-dimensional artificial signal or in 2-D signal processing like image processing [1,2,3,4,5]. But few concern about and deal with speech signal. If we can make noise reduction and improve the quality of speech such as signal-to-noise ratio (SNR), the speech recognition rate will definitely increase and this will make computer voice input more reliable and less error prone.

Donoho and Johnstone [1] had proposed the universal threshold $th = \sigma\sqrt{2\log_e N}$ where σ is the standard deviation of Gaussian noise. For large N, this threshold seems unrelevant, and is against reasoning to have the threshold that depends upon the sample size. Soon et al. [6] treat the threshold as the expected value of the noise magnitude in the decomposed subband, but this value lacks sound theoretical proofs in wavelet arena.

The time complexity of discrete wavelet transform (DWT) is $O(N)$, i.e., linearly

proportional to the sample size. That makes noise reduction by DWT rather faster. There are basically two types of threshold rule, one is soft thresholding(η_s), and the other is hard thresholding(η_h):

$$\eta_s(w,th)=\begin{cases} w-th & w \ge th \\ 0 & |w|<th \\ w+th & |w| \le -th \end{cases} \quad (1)$$

$$\eta_h(w,th)=\begin{cases} w & |w| \ge th \\ 0 & |w|<th. \end{cases}$$

where w is wavelet coefficients, and th is the threshold.

Speech signals are different in a way from those artificial ones. It possesses its own nature. We thought it is one dimension of signal and image signal as two dimensional, but their characteristics are quite different. The research about acoustic theory of speech production starts from physics laws and the theory brings up waveform representations and parametric representations. In digital signal processing of speech through wavelet transform, not much research done concerns about finding the threshold value.

This paper proposes a theoretical method to determine the value of the threshold more soundly. This begins approximation of the noise-free signal by minimizing the mean squared error (MSE). The foundations of noise reduction by thresholding coefficients in transformed wavelet subbands are based upon the concentrating capability of the wavelet transform. If a signal has its energy concentrated on the small number of transformed coefficients, these coefficients tend to relatively large compared to the transformed coefficients of the noise in each subband. We assume the clean signal is independent of the noise. The shrinkage of the coefficients will remove the undesired signal (the noise) in the wavelet domain. And then the inverse of the wavelet transform will then retrieve the desired signal. Generally speaking, noise reduction is achieved by thresholding the wavelet coefficients of the wavelet transform of the noisy signal.

The rest of the paper is outlined as follows. The wavelet transform theory is described in section II. Section III provides the threshold selection. The experimental results and analysis are in section IV. Conclusions and references are then at final.

II. The Wavelet Transform Theory

Suppose the noisy signal y in time domain (t) is additive, i.e., $y[t] = f[t] + n[t]$, and is decomposed in an orthogonal basis,

$$\boldsymbol{G}=\{g_m\}_{L\le m\le J}: \langle y, g_m\rangle = \langle f, g_m\rangle + \langle n, g_m\rangle, \quad (2)$$

where f is the clean signal to be recovered, n is Gaussian noise of zero-mean and standard deviation σ.

When the sample size (S) of the noisy signal is equal to 2^{J+1}. The WT decomposition produces the scaling function coefficients $\{c_k, k = 1,\ldots, 2^L\}$ at (coarsest) scale level L, the wavelet coefficients $\{ w_{j,k}, k = 1,\ldots, 2^j ; j = L,\ldots,J\}$ at scale level L to J, and for a total of $2^L+2^L+2^{L+1}+2^{L+2}+\ldots+2^J = 2^{J+1} = S$ transform coefficients.

But for speech signal the sample size S is usually not in the form of 2's power. Each level of WT decomposition is actually achieved by convolving the data sequence with h_0 and down-sampling to obtain the scaling coefficients of half size, and also by convolving the data sequence with h_1 and down-sampling to get the wavelet coefficients of half size. h_0 is a low-pass filter and is the scaling function(φ) coefficients; whereas h_1 is a high-pass filter and is the wavelet (ψ) coefficients. Their relationship can be denoted as:

$$h_1(n)=(-1)^n h_0(1-n). \quad (3)$$

The procedure of splitting, filtering, and down-sampling is shown below in figure 1.

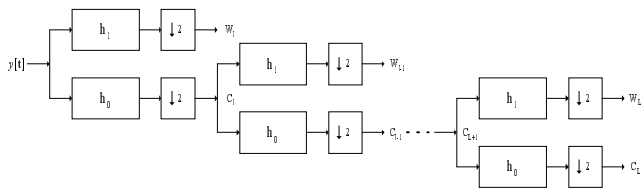

Figure 1. Scale level J-to-L Two-band Decomposition Tree

The thresholding is applied in wavelet coefficients in each subband while leaving

unchanged the scaling function coefficients $\{c_k, k = 1,\ldots, 2^L\}$ of scale level L (coarsest level).

The discrete wavelet thresholding procedure executed in our model is diagramed in figure 2. The synthesis of the clean speech and the noise is finished by the tool Cool Edit 2000.

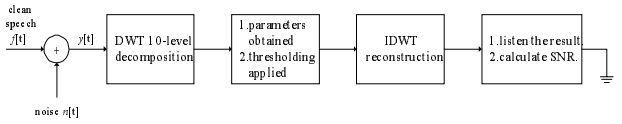

Figure 2. The DWT thresholding flowchart.

III. THRESHOLD SELECTION

We observed the noisy signal $y[t] = f[t] + n[t] + V_{biased}$, where $n[t]$ is independent of $f[t]$ and identically distributed (iid) as normal probability density function (pdf): Normal(0, σ^2). V_{biased} is the dc offset and is a constant. The whole noisy signal could be shifted to zero horizontal x-axis, thus eliminating the bias. So the goal is to remove the noise $n[t]$, and to obtain an estimate $\hat{f}[t]$. The criterion of finding $\hat{f}[t]$ is to minimize the mean squared error (MSE),

$$\mathrm{MSE}(\hat{f}[t]) = \sum(\hat{f}[t]-f[t])^2 \qquad (4)$$

Let $Y = \boldsymbol{T}(y[t])$ be the operation of wavelet transform ($\boldsymbol{T}$) of noisy signal, similarly $F = \boldsymbol{T}(f[t])$, and $N = \boldsymbol{T}(n[t])$. Since the transform is orthogonal, N is also iid normal pdf: Normal(0, σ^2), with mean = 0, and variance = σ^2. So after DWT, we get a formula in wavelet domain as:

$$Y = F + N \qquad (5)$$

In digital processing of speech signals[7], the pdf of the amplitudes of speech in time domain is approximately Laplacian distribution (the double exponential distribution) [8,9]. Because of DWT applied is orthonormal, so is F Laplacian distribution. Thus the objective is to find a soft-threshold value th which minimizes the risk,

$$\begin{aligned}\mathrm{Risk}(th) &= \mathrm{Exp}\{\hat{F}(th)-F\}^2 \\ &= \mathrm{Exp}_F\,\mathrm{Exp}_{Y|F}\{\hat{F}(th)-F\}^2\end{aligned} \qquad (6)$$

where Exp is the expectation of probability, $F \sim$ Laplacian(0, σ_F^2), with zero mean, and variance = σ_F^2.

$Y|F \sim$ Normal(f, σ^2). $Y|F$ is normal distribution and has the same mean as f, which is equal to zero after eliminating the bias. $Y|F$ also has variance equal to σ^2.

$$\hat{F}(th) = \eta_s(Y) = \begin{cases}\mathrm{sgn}(Y)(|Y|-th), & |Y| \ge th \\ 0, & |Y| < th\end{cases} \qquad (7)$$

The risk (6) calculation now is to solve the Bayesian probability [10,11,12] and it could be derived as follows:

$$\mathrm{Exp}_F\mathrm{Exp}_{Y|F}\{\hat{F}(th)-F\}^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\eta_s(y)-x)^2 p(y|x)p(x)dydx \qquad (8)$$

$$= \sigma^2 \cdot q(\frac{\sigma_F^2}{\sigma^2}, \frac{th}{\sigma}) \quad , \text{ where } q \text{ is a function.}$$

$$= \sigma^2\cdot[\sigma_F^2 + 2\cdot(th^2+1-\sigma_F^2)\cdot\bar{\Phi}(\frac{th}{\sqrt{1+\sigma_F^2}}) - 2\cdot th\cdot(1+\sigma_F^2)\cdot\phi(th,1+\sigma_F^2)]$$

where

$$\phi(th,1+\sigma_F^2) = \frac{1}{\sqrt{2\pi(1+\sigma_F^2)}}\exp\{-\frac{th^2}{2(1+\sigma_F^2)}\} \qquad (9)$$

$$\bar{\Phi}(\mu) = \int_{\mu}^{\infty}\phi(x,1)dx\,. \qquad (10)$$

From the above formulas, we can get the optimal value of th :

$$th^* = \arg\min_{th} risk(th) \qquad (11)$$

The th^* obtained is not in closed form and must be computed through numerical integration from formulas (6) - (11) to find this value. Nevertheless it is interesting to point out that if we set $th = \frac{\sigma^2}{\sigma_F}$ for each subband (where σ^2 is the noise variance in each subband, and σ_F is the standard deviation of signal in each subband), the deviation from th^* (the theoretical value) will be controlled within certain range. It has not only the simplicity, but is also fast to set adaptively the threshold in each detailed subband, i.e., except the coarsest level subband of scaling function coefficients.

IV. Experimental Results and Analysis

The wavelet bases selected for performing DWT are Daubechies Wavelets, which are compactly supported and orthonormal, and the lengths of filter coefficients are 4, 8, 14, and 20 respectively. The clean signal is the clear and noise-free speech, named as 'ktv.wav' and 30 seconds long. The noise 'babble' 30 seconds long is a mumble speech of people. As opposed to 'babble' noise, the noise 'f16', also 30 seconds long, is the screaming of jet fighter engine. Before adding the noise, the amplitude of it is scaled 50% and 25% respectively as high. The DWT decomposition and synthesis is 10-level. The sampling rate is 22050 Hz. So the sample size is indeed a huge number. The whole experiments are all done and programmed in Matlab (v. 5.2) environment.

Even if the perfect reconstruction of DWT, we pass the clean speech to a 2-level decomposition and synthesis model to estimate the maximum SNR obtainable, i.e., no thresholding applied. Though the speech not altered a little and as clear, the maximum obtained is 16.809 dB as computed. This could be attributed to the computer round-off and the huge of sample size. The SNR results are listed in table 1. Besides the above reasons, the SNR value upgraded within a certain limit is due to one-time estimation of the parameters. If the noisy speech is segmented, and each segment, for example, 2-3 seconds long is processed individually, it will make elevating SNR value higher possible and is left for further research.

Signal & Noise type	SNR before the experiments	SNR after the experiments & wavelet selected			
		D4	D8	D14	D20
ktv + (50%) babble noise	8.102	9.471	9.556	9.622	9.664
ktv + (25%) babble noise	14.123	14.735	14.796	14.838	14.869
ktv + (50%) f16 noise	7.319	9.121	9.292	9.393	9.440
ktv + (25%) f16 noise	13.339	14.234	14.406	14.485	14.524

Tabel 1. SNRs (in dB)
Note. D#: Daubechies wavelet of length #.

V. CONCLUSION

From the above, we may observed that the larger the length of filter coefficients used, the better SNR value obtained. This means that larger Daubechies wavelet has better filtering effects. The wavelet shrinkage for noise reduction of speech is possible, but we should first estimate the parameters correctly. This is the main purpose of the paper to provide the theoretical threshold for noise reduction. One of the characteristics of wavelet is that it is well suited to transient signals like speech, and it can separate the noise from signal. This make noise reduction possible by thresholding in wavelet domain. Whereas Fourier analysis deals mostly with stationary signals and is appropriate for periodic signals whose statistical characteristics don't change with time, so it can not remove some types of noises in speech like the 'babble' noise.

Even though the huge sample size, the wavelet thresholding is rather faster in performing noise reduction. This is because the time complexity of it is $\boldsymbol{O}(N)$, rather than $\boldsymbol{O}(N \log N)$ which is needed in Fourier transform.

References:

[1]David L. Donoho, and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," Biometrika, vol. 81, pp. 425-455, 1994.

[2] David L. Donoho, and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," Journal of American Statistical Association, vol. 90, No. 432, pp. 1200-1220, Dec. 1995.

[3]S. Mallat, a Wavelet tour of signal processing, Academic Press, 1998.

[4]S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," IEEE Trans. on Patt. Recog. and Mach. Intell., vol. 11, no. 7, pp.674-693, July 1989.

[5]David. L. Donoho, "De-noising by soft-thresholding," IEEE Trans. On Information Theory, vol. 41, no. 3, pp. 613-627, May 1995.

[6]I.Y. Soon, S.N. Koh, and C.K. Yeo, "Wavelet

for speech Denoising," IEEE TENCON – Speech and Image Technologies for Computing and Telecommunications, 1997.
[7]L.R. Rabiner, and R.W. Schafer, Digital Processing of Speech Signals, Prentice Hall, 1978.
[8]W. B. Davenport, "An Experimental Study of Speech-wave Probability Distributions," J. Acoust. Soc. Am., vol. 24, pp. 390-399, July 1952.
[9]M. D. Paez, and T. H. Glisson, "Minimum Mean Squared-Error Quantization in Speech," IEEE Trans. Comm., vol. Com-20, pp. 225-230, April 1972.
[10]S. M. Ross, Introduction to Probability Models, 5th Ed., Academic Press, 1993.
[11]E. Kreyszig, Advanced Engineering Mathematics, 7th Ed., John Wiley & Sons, 1993.
[12]R. V. Hogg, and A. T. Craig, Introduction to Mathematical Statistics, 4th Ed., Macmillan Publishing, 1978.