Denoising Speech Signals by Wavelet Transform

Slavy Georgiev Mihov, Ratcho Marinov Ivanov and Angel Nikolaev Popov

Abstract – **This paper investigates the use of wavelet transform for denoising speech signals contaminated with common noises. Shown are the basic principles of wavelet transform as an alternative to the Fourier transform. The practical results obtained are based on processing a large dedicated database of reference speech signals contaminated with various noises in several SNRs. This research tends to be an extension to the practical research for speech signal enhancement for the purposes of hearing-aid devices.**

Keywords – **wavelet denoising, speech signals**

I. Introduction

Here is presented an investigation of the use of wavelet theory for practical signal denoising. Studied are the potentials of wavelet transform for improving the hearing perception of humans to noise contaminated speech records. This study is a continuation of the research for speech enhancement for the needs of small portable devices and particularly hearing-aid devices [1].

Fourier transform based spectral analysis is the dominant analytical tool for frequency domain analysis. However, Fourier transform cannot provide any information of the spectrum changes with respect to time. Fourier transform assumes the signal is stationary, but speech signal is always non-stationary. To overcome this deficiency, a modified method-short time Fourier transform allows to represent the signal in both time and frequency domain through time windowing function. The window length determines a constant time and frequency resolution. Thus, a shorter time windowing is used in order to capture the transient behavior of a signal; we sacrifice the frequency resolution. The nature of the real speech signals is nonperiodic and transient; such signals cannot easily be analyzed by conventional transforms. So, an alternative mathematical tool – wavelet transform must be selected to extract the relevant time-amplitude information from a signal. In the meantime, we can improve the signal to noise ratio based on prior knowledge of the signal characteristics.

A. Wavelet Denoising

Wavelet denoising is considered a non-parametric method. Thus, it is distinct from parametric methods in which parameters must be estimated for a particular model that must be assumed a priori.

Assume that the observed data

$$X(t) = S(t) + N(t) \tag{1}$$

contains the true signal *S(t)* with additive noise *N(t)* as functions in time *t* to be sampled. Let $W(\cdot)$ and $W^{-1}(\cdot)$ denote the forward and inverse wavelet transform operators. Let $D(\cdot,\lambda)$ denote the denoising operator with soft threshold λ. We intend to wavelet denoise *X(t)* in order to recover $\hat{S}(t)$ as an estimate of *S(t)*. Then the three steps

$$Y = W(X) \tag{2}$$

$$Z = D(Y,\lambda) \tag{3}$$

$$\hat{S} = W^{-1}(Z) \tag{4}$$

summarize the procedure. Of course, this summary of principles does not reveal the details of implementing the operators *W* or *D*, or selection of the threshold λ.

B. Wavelet Transform

In this work, we stated only some keys equations and concepts of wavelet transform, more rigorous mathematical treatment of this subject can be found in [2, 3, 4, 5]. A continuous-time wavelet transform of *f(t)* is defined as:

$$CWT_{\Psi} f(a,b) = W_f(b,a) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(t) \overset{*}{\Psi}\left(\frac{t-b}{a}\right) dt \tag{5}$$

Here a, $b \in R$, $a \neq 0$ and they are dilating and translating coefficients, respectively. This multiplication of $|a|^{-1/2}$ is for energy normalization purposes so that the transformed signal will have the same energy at every scale. The analysis function $\Psi(t)$, the so-called mother wavelet, has to satisfy that it has a zero net area, which suggest that the transformation kernel of the wavelet transform is a compactly support function (localized in time).

One drawback of the CWT is that the representation of the signal is often redundant, since *a* and *b* are continuous over *R* (the real number). The original signal can be completely reconstructed by a sample version of $W_f\,(b,a)$. Typically, we sample $W_f\,(b,a)$ in dyadic grid, i.e., $a = 2^{-m}$ and $b = n2^{-m}$, $m,n \in Z+$. Substituting the last one into (5):

$$DWT_{\Psi} f(a,b) = \int_{-\infty}^{\infty} f(t) \overset{*}{\psi}(t) dt \tag{6}$$

where $\Psi_{m,n}(t) = 2^{-m}\Psi(2^{m}t-n)$ is the dilated and translated version of the mother wavelet $\Psi(t)$.

Due to the orthonormal properties, there is no information redundancy in the discrete wavelet transform. In addition, with this choice of *a* and *b*, there exists the multiresolution analysis (MRA) algorithm, which decompose a signal into scales with different time and frequency resolution.

The differences between different mother wavelet functions (e.g. Haar, Daubechies, Coiflets, Symlet, Biorthogonal and etc.) consist in how these scaling signals and the wavelets are defined. The choice of wavelet determines the final waveform shape; likewise, for Fourier transform, the decomposed waveforms are always sinusoid.

The wavelet decomposition results in levels of approximated and detailed coefficients. The algorithm of wavelet signal decomposition and reconstruction of the signal from the wavelet transform is illustrated in numerous sources [2, 3, 4].

S. Mihov, R. Ivanov, A. Popov are with the Department of Electronics, Faculty of Electronic Engineering and Technologies, Technical University – Sofia, 8 Kliment Ohridski blvd., 1000 Sofia, Bulgaria, e-mail: smihov@tu-sofia.bg

C. Thresholding

How is the threshold λ determined? Let's say that the data has sample size n if it has been sampled at n points t_i such that $X_i \equiv X(t_i)$. Then for an orthogonal W, there will also be n transform coefficients Yj. If we prefer to use a threshold (such as the minimax threshold or the universal threshold) that depends only on n, then λ can be predetermined and we can use the three-step denoising procedure already described. However, if we prefer to use a data-adaptive threshold

$$\lambda = d(U) \tag{7}$$

(such as the threshold selected by Stein's Unbiased Risk Estimator (SURE)) that depends not just on n but on U (which again represents the data in any generic domain), then we must use a four-step procedure: (2), (7), (3), (4).

There are four common rules for selecting the threshold λ in practice: heursure, minimax, rigsure and sqtwolog. Minimax and SURE threshold selection rules are more conservative and would be more convenient when small details of the signal lie near the noise range. The two other rules remove the noise more efficiently.

On the other side, thresholding can be done in different ways. Most popular techniques are hard thresholding and soft thresholding (Fig. 1). Hard thresholding is the simplest method but soft one has nice mathematical properties.

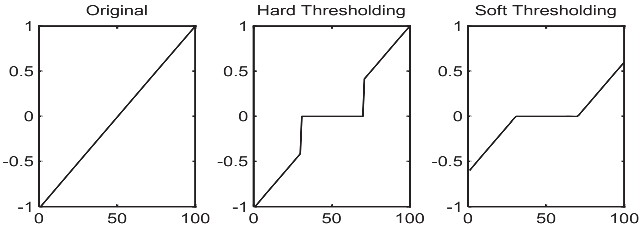

Fig. 1. – hard and soft thresholding

Hard thresholding can be described as the usual process of setting to zero the elements whose absolute values are lower than the threshold. The hard threshold signal is x if $x > \lambda$, and is 0 if $x <= \lambda$.

Soft thresholding is an extension of hard thresholding, first setting to zero the elements whose absolute values are lower than the threshold, and then shrinking the nonzero coefficients towards 0. The soft threshold signal is $sign(x)(x - \lambda)$ if $x > \lambda$ and is 0 if $x <= \lambda$.

II. Experiments

A wavelet transform must be specified by its analysis and synthesis wavelet filter banks, single-level convolutions and boundary treatment, and the total number L of iterated multiresolution levels. Thus, we can generate many different kinds of wavelet shrinkage denoising procedures by combining different choices for $W(\cdot)$ and $d(\cdot)$. If we let D denote more generally either the soft thresholding operator D_s or the hard D_h, then by combining choices for $W(\cdot)$, $D(\cdot,\cdot)$, and $d(\cdot)$, we can generate even more different kinds of wavelet-based denoising.

A. Signal Database

The test database contains 720 sentences from the IEEE corpus [6, 7, 8] produced by a male speaker. The sentences (*.wav files) were sampled at 25 kHz.

A subpart of 30 records has a narrowband duplicate sampled at 8 kHz. They are contaminated with different types of noise from a couple of common noisy environments, (listed in Table 1). The noise is added to the speech signal in four particular SRNs: (15 dB, 10 dB, 5 dB, 0 dB). The so produced pairs of reference and noisy signals are used for evaluating wavelet transform denoising.

Table 1. Common Noisy Environments

Environment	Noise Specifics
Airport	music, babble, aircraft engine
Babble	people speaking in the background
Car	car engine roar
Exhibition	music, babble, camera clicks
Restaurant	several levels of background speech,
Station	megaphone speech, footsteps
Street	engine roar, beeps, horns
Train	rail track noise

B. Work Environment

For the purpose of this research of the potentials of wavelet transform in denoising speech signals are carried several experiments. The speech signals from the database are being processed by denoising algorithms and the obtained denoised records are stored and measured. The processing algorithm and the ones used to give objective estimate of the obtained quality are done in Matlab. As a supplemental tool in the development of the script codes is used Wavelet Toolbox from Matlab IDE (Fig. 2).

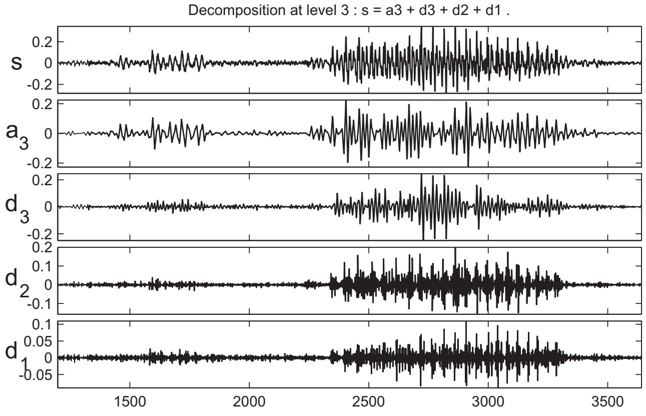

Fig. 2. – Wavelet Toolbox (Matlab)

C. Metrics

The theoretical claims of optimality and generality pertain to a wide range of local and global measures of error, such as SNR measured in decibels. In fact, varying results can be obtained with different experimental conditions (signal classes, noise levels, sample sizes, wavelet transform parameters) and error measures as well as the SNR (measured in standard deviations and in decibels). Which measure of error is most relevant? What about other "figures of merit"?

In our evaluation of the wavelet denoising performance, two metrics were used [9, 10]. The first is Signal-to-Noise Ratio (SNR), which gives the proportion of the wanted and unwanted signals. For classification of the frame as "signal" or "noise", the reference channel was used. The SNR is the proportion of the averaged energy during the "signal" and "noise" frames. This metric gives an indirect estimate of the sound quality.

Mean Opinion Score (MOS) – ITU-T P.800 was used as a primary metric for the quality of the output signal after processing. This is a dimensionless quantity with values ranging from 1 to 5. It gives an estimate of human perception of sound quality. Estimating MOS with real humans is long and expensive procedure, involving many humans listening to the records and giving their subjective opinion. For this reason, MOS is not suitable for use during the stage of algorithm development. We used objective Perceptual Evaluation of Sound Quality (PESQ) – ITU-T P.862. It produces similar results to MOS results in the same scale (1 to 5) to give an estimate of human perception of sound quality too. We used the Matlab implementation of PESQ algorithm [6] which requires reference channel.

III. Results

A. Threshold Type

One of the completed experiments makes a comparison between the types of thresholding: hard, soft. Since Qian thresholding [11] produces results somewhat in between these ones, it is out of scope of this particular experiment. Fig. 3 and Fig. 4 show the average results in both metrics (SNR and MOS) from wavelet denoising using soft and hard thresholding, obtained with four different methods of threshold estimation (heursure, rigsure (local thresholds λ_l estimated adaptively for each level l) , minimax (global threshold), sqtwog (global $\lambda = sqrt(2 \log n)$)).

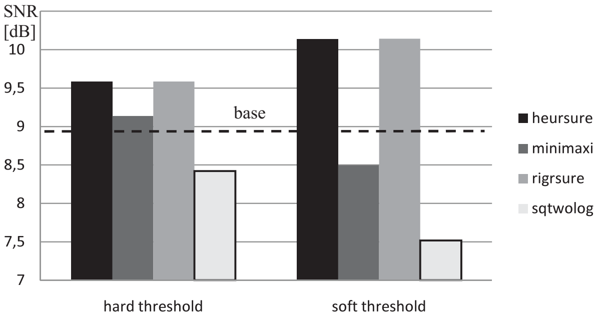

Fig. 3. – Threshold Types (SNR)

Processing the contaminated speech signals with minimax and sqtwog criterion for threshold estimation does not improve the perception quality of the signals. Both results in SNR and MSE metrics show degradation in the signal being processed by wavelet transform.

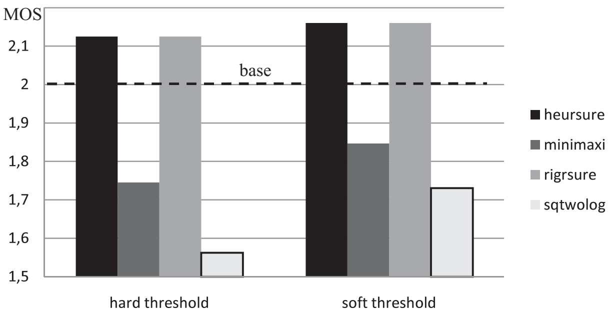

Fig. 4. – Threshold Types (MOS)

Signal quality improves by wavelet denoising when the threshold values are estimated according heursure and rigsure criteria. Since these two ones are quite similar algorithm, so are the results. Quite significant improvement is achieved using soft thresholding than using hard thresholding, which is quite unexpected and can partially be explained by the statistical parameters of noise and speech signals in the record.

B. Level of Decomposition

An experiment is done to evaluate the performance of wavelet transform (forward and reverse) with different levels of decomposition. The results in both metrics are shown on Fig. 5 and Fig. 6. The horizontal axis denotes the levels of decomposition and the vertical – SNR and MOS respectively. Given are several graphics corresponding to two different scaling algorithms (mln, sln) used for scaling the signal before thresholding (hard or soft) in wavelet transform. "Base" denotes the results for the initial contaminated speech signal (no processing).

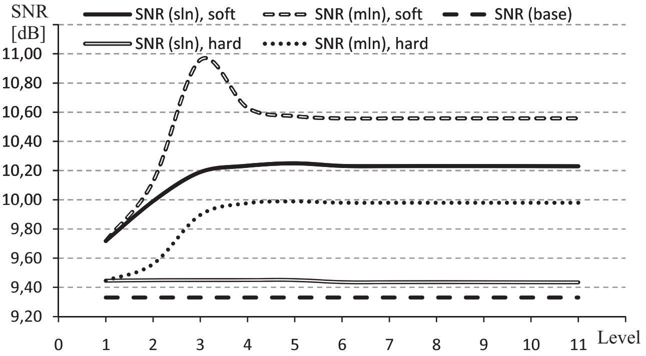

Fig. 5. – Levels of Decomposition (SNR)

Increasing the levels of decomposition increases the computational complexity of the wavelet denoising algorithm. The graphics show that this does not give sensible improvement in signal quality. For practical reasons it is pointless to evaluate large levels of wavelet decomposition so the levels of decomposition should be limited to no more than 5.

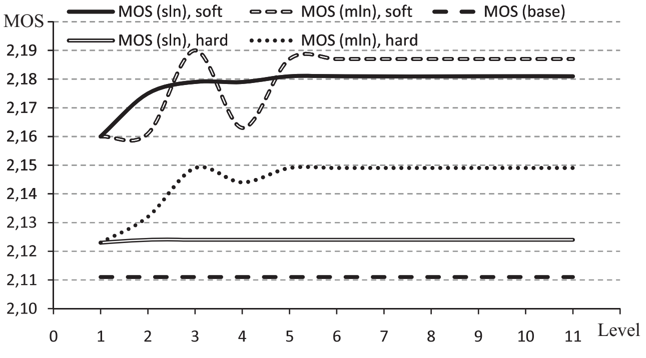

Fig. 6. – Levels of Decomposition (MOS)

The graphical results obtained from sln scaling algorithm are smoother than the ones from the mln algorithm. The results from the second one (mln) tend to achieve better scores in SNR and MOS metrics. However, these two metrics are just an objective measure of the subjective human perception for sound quality. The human estimate of the results of both algorithms shows that sln denoised signal sounds quite a bit better. There are some noise artifacts which are better audible in the mln signals.

As a result from this experiment is proved that mln and sln algorithm for scaling achieve very similar results with a slight human preference to sln.

C. Initial SNR

Another experiment is done to track the behavior of the investigated wavelet denoising procedure when processing speech signals contaminated more or less with noise. Processed are similar records contaminated with noise in different initial SNRs .The results of the wavelet denoising are shown on Fig. 7 and Fig. 8. The horizontal axis denotes the initial SNRs and the vertical – SNR and MOS results from denoising respectively. Given are several graphics corresponding to several common wavelet function used in practice (haar, db3, db5, …). "Base" denotes the results for the initial contaminated speech signal (no processing).

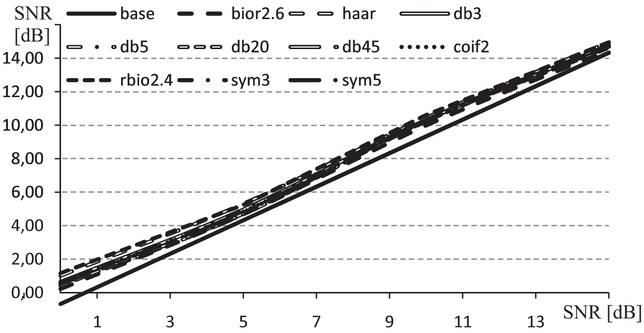

Fig. 7. –Different Levels of Contaminated Speech (SNR)

The graphical results in both metrics SNR and MOS from denoising with different wavelet functions prove to be quite similar. Seen is the improvement in comparison with the base result.

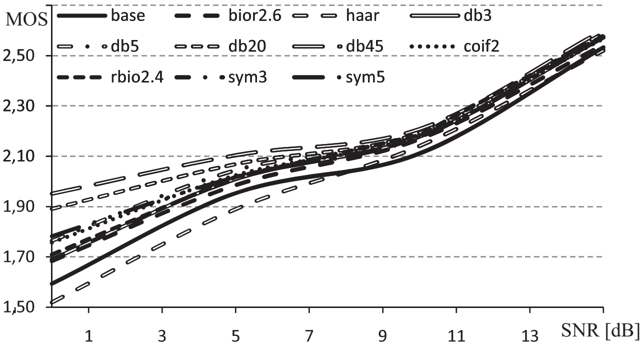

Fig. 8. – Different Levels of Contaminated Speech (MOS)

Since the different results from the wavelet functions used can barely be distinguished, for practical reasons it is wise to use wavelet transform with function having low computational complexity. The practical experiments gave the execution time of the experimented combinations of wavelet parameters. Due to its computational complexity, wavelet function sym3, sym4, … proved to be practically inapplicable for real time denoising. The rest of the functions have good execution times (eg. db3, db5). For best results in practical speech signal denoising, one should select the parameters of the wavelet transform according to the previous experiments presented too.

IV. Comment

With regard to wavelet denoising, the theoretical justifications and arguments in its favor remain highly compelling. The procedure does not require any assumptions about the nature of the signal, permits discontinuities and spatial variation in the signal, and exploits the spatially adaptive multiresolution features essential to the wavelet transform. Furthermore, the procedure exploits the fact that the wavelet transform maps white noise in the signal domain to white noise in the transform domain. Thus, while signal energy becomes more concentrated into fewer coefficients in the transform domain, noise energy does not. It is this important principle that enables the separation of signal from noise.

V. Conclusion

It is unlikely that one particular wavelet shrinkage denoising procedure will be suitable, no less optimal, for all practical problems. However, it is likely that there will be many practical problems, for which after appropriate experimentation, wavelet-based denoising with either hard or soft thresholding proves to be the most e effective procedure. Estimation of the power spectrum by wavelet-based denoising of the log-periodogram may prove to be one such important application with great promise for further development in speech signal enhancement.

Acknowledgement

This research was supported by the NIS TUS funded project, No 091pd007-03/2009.

References

[1] S. Mihov, D. Doychev, R. Ivanov. "*Practical investigation of specific types of noise signals for the purpose of their suppression in hearing-aid devices"*, Proceedings of ICEST-2009, vol. 2, pp. 399-402, Veliko Tarnovo 2009.

[2] C. Gargour, M. Gabrea, V. Ramachandran, J. Lina. "*A Short Introduction to Wavelets and Their Applications"*, IEEE Circuits and Systems Magazine, ISSN: 1531-636X, vol. 2, pp. 57-67, 2009.

[3] C. Taswell. "*The What, How and Why of Wavelet Shrinkage Denoising"*, Computing in Science and Engineering, ISSN: 1521-9615, vol. 2, no. 3, pp. 12-19, June 2000.

[4] S. Tsai, *"Wavelet Transform and Denoising"*, Master's Thesis, URN: etd-12062002-152858, Chapter 4, pp. 35-42.

[5] http://taco.poly.edu/WaveletSoftware/denoise2.html

[6] P. C. Loizou, *Speech Enhancement Theory and Practice*, Taylor & Francis Ltd, ISBN-13: 978-0849350320, 1st edition (7 June 2007).

[7] IEEE Subcommittee (1969)."IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. Audio and Electroacoustics," AU-17(3), 225-246.

[8] Kawahara, H., Masuda-Katsuse, I., and de Cheveigner, A. (1999). ''Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,'' Speech Commun. 27, 187–207.

[9] S. Mihov, T. Gleghorn, I. Tashev. "*Enhanced Sound Capture System for Small Devices"*, Proceedings of ICEST-2008, pp. 57-67, Nis, Serbia, 2008.

[10] I. Tashev, S. Mihov, T. Gleghorn, A. Acero. "*Sound Capture System and Spatial Filter for Small Devices"*, Proceedings of Interspeech-2008, pp. 57-67, Brisbane, Australia, October 2008.

[11] J. Qian. *"Denoising by Wavelet Transform"* Tech. Rep. Rice University. Department of Electrical Engineeringsg. 2001.