ADAPTIVE ESTIMATION OF PERIODIC NOISE ENERGY DISTRIBUTIONS FOR SPEECH ENHANCEMENT

Thomas Weickert, **Uwe Kiencke**

Institute of Industrial Information Technology
Universität Karlsruhe (TH)
Hertzstr. 16, 76187 Karlsruhe, Germany
email: weickert@iiit.uni-karlsruhe.de

Abstract: Speech Enhancement aims at denoising a noisy speech signal and to extract the clean speech. In this work, nonstationary periodic noise is regarded. The coefficient energy distribution of a period of noise is analyzed by a Period Wavelet Packet transform. This transform can analyze signals of arbitrary (non-dyadic) length. The estimated coefficient energy distribution can be used as a threshold for wavelet thresholding techniques. The efficiency of the algorithm and its sensitivity to period variations is evaluated.

Keywords: signal processing, filtering, filter banks, non-stationary signals, speech

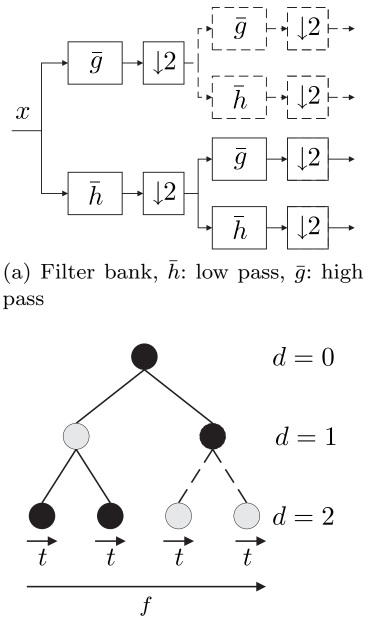
1. INTRODUCTION

Speech enhancement is an important task for applications like communication in noisy cockpits, speech recognition systems and hands-free car sets. In this work, the class of periodic noises will be regarded, because they occur in many environments, especially in presence of machines (engines, compressors, factory noise etc.). The approach in this paper is to estimate the time-frequency energy distribution of a period of the noise when only noise is present (speech pause). This knowledge is then used to remove the noise when speech is present. However, noise can have an arbitrary period while the common filter bank algorithm can only analyze signals of dyadic length. Therefore, a new representation is required.

This paper is structured as follows. A review of the signal adaptive wavelet packet transform and wavelet-based denoising is given in section 2. A wavelet representation of a periodic signal of non-dyadic length and appropriate filter algorithms are presented in section 3. Examples in section 4 will show the effectiveness and flaw of the proposed method. The conclusions are drawn in section 5.

2. REVIEW: DENOISING BY WAVELET PACKETS

A discrete wavelet transform (DWT) can be computed efficiently by a Conjugate Mirror Filter bank (Mallat, 1999). A signal x of dyadic length $N = 2^l, l \in \mathbb{N}$, is analyzed by cyclic convolutions with high pass and low pass filters and subsequent downsampling. For wavelet bases, only the low pass signals are decomposed into their high pass and low pass component (see figure 1, solid lines only). A generalized approach is to apply the decomposition stage to high pass and low pass signals alike. This yields a tree-like redundant decomposition as shown in figure 1b (solid and dashed lines), a so-called wavelet packet (WP). Each node represents a certain frequency band. The coefficients of a certain node correspond to



(a) Filter bank, $\bar{h}$: low pass, $\bar{g}$: high pass

(b) Subspace Tree

Fig. 1. Wavelet Packet Transform, max. depth $D = 2$, solid line: wavelet transform (subset of a wavelet packet), dashed: additional decomposition for wavelet packets

the development over time in this frequency band. Each coefficient of a WP can be addressed by

$$WP(d,b,n) \quad ,0 \leq d \leq D \quad ,0 \leq b < 2^d - 1$$
$$,0 \leq n < \frac{N}{2^d} - 1$$

where d denotes the depth in the tree, b is the node number in this depth and n is the coefficient number in the certain node.

Fast algorithms have been developed to search the "best" orthonormal basis in a wavelet packet (Coifman and Wickerhauser, 1992). The "best" basis is the sparsest representation of the signal within the WP, i.e. the signal is represented by few but large coefficients. This could be e.g. the wavelet basis or the basis indicated by the gray nodes in figure 1b.

Denoising in the wavelet domain was introduced by Donoho for additive, white noise (Donoho, 1995) and by Johnstone and Silverman for stationary, colored noise (Johnstone and Silverman, 1997). The results can be translated easily to WPs. An additive, white noise in the time domain corresponds to an additive white noise in every orthonormal basis. Thus, a threshold

$$t = \sigma\sqrt{2\log_{\mathrm{e}} N} \qquad \begin{aligned} \sigma^2 &: \text{ noise covariance} \\ N &: \text{ signal length} \end{aligned}$$

can be found for which the probability is high, that the absolute values of the noise coefficients are below this threshold. Setting all coefficients with an absolute value smaller than the threshold to zero is a simple but efficient denoising method:

$$WP_{filt}(d,b,n) = \begin{cases} WP(d,b,n) & ,|WP(d,b,n)| \geq t \\ 0 & ,|WP(d,b,n)| < t \end{cases} \tag{1}$$

This method becomes much more effective, if the best basis of the WP is selected for the thresholding. Besides this "hard" thresholding (1), several other thresholding functions have been proposed (Ayat *et al.*, 2004), but the decision which one to use is not relevant for this paper.

For colored noise, a node-dependent threshold can be calculated for each node. After thresholding all coefficients according to the thresholding function and the threshold t, the WP synthesis filter bank yields the filtered signal.

3. THE PERIOD WAVELET PACKET TRANSFORM

In recent years, wavelet packets have been used for denoising distorted speech, e.g. see (Sheikhzadeh and Abutalebi, 2001), (Shao and Chang, 2005). They usually regard stationary noise. However, an advantage of time-frequency representations over Fourier transforms is the processing of nonstationary signals. Thus, the approach in this paper is to use wavelet packets to estimate the noise energy distribution in the time-frequency plane for the class of periodic noise. Periodic noise does not denote a periodic signal, but a disturbance with a power density which changes periodically over time. The basic idea is to analyze the coefficient energy distribution of a single period in the time-frequency plane and subsequently use this estimation to shrink each WP coefficient of the noisy signal individually. The noise coefficient energy distribution of multiple periods can be averaged in order to get a more robust estimation. An adaptive averaging is given by

$$F_1(d,b,n) = P_1(d,b,n)$$
$$F_k(d,b,n) = \frac{\sum_{i=1}^{k} \lambda^{k-i} P_i(d,b,n)}{\sum_{i=0}^{k-1} \lambda^i}$$

where $F_k(d,b,n)$ is the averaged noise energy distribution after k periods, $P_i(d,b,n)$ is the analyzed noise energy distribution of period i and λ is a forgetting factor. This results to the recursive equation:

$$F_{k+1}(d,b,n) = \frac{F_k(d,b,n) \cdot \sum_{i=1}^{k} \lambda^i + P_k(d,b,n)}{\sum_{i=0}^{k} \lambda^i} . \tag{2}$$

The basic problem of this approach is, how to calculate the noise WP coefficients of a single period, because the WP filter bank needs an input signal of dyadic length $N = 2^l, l \in \mathbb{N}$, whereas in

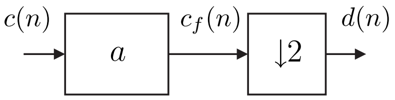


Fig. 2. WP Filter bank stage

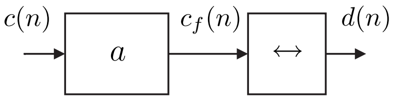


Fig. 3. Period WP filter bank stage

general a noise period can be of arbitrary length. Therefore, a new representation is required.

3.1 Analysis

A single filter bank decomposition stage is regarded in figure 2. The input coefficient array $c(n)$ is convoluted with the impulse response a (high pass or low pass respectively). The result $c_f(n)$ is downsampled which finally yields the new coefficient array $d(n)$. If the coefficient array $c(n)$ has the period T, then the convoluted coefficient array $c_f(n)$ has the same period T, due to the cyclic convolution used. The downsampling operator only keeps the even-indexed samples.

If the period T is even, the downsampled array $d(n)$ consists of the even-indexed samples of the period. An example is shown in figure 4, double-lines marking a period. The signal $d(n)$ has the period $T/2$, but all multiples of $T/2$ are also valid periods of $d(n)$.

If the period T is odd, then the even-indexed samples of the first period are kept followed by the odd-indexed samples of the second period, see figure 5. The period of $c_f(n)$ for this case is still T. From this follows, that the period T is not changed by a filter bank stage.

This means, if a single period is to be analyzed, the coefficient array lengths must stay constant. Otherwise, the coefficient arrays of later decomposition stages get shorter while the period remains constant. Important information would get lost. Thus, a sorting operator "↔" is applied instead of a downsampling operator. It translates a coefficient array $c_f(n)$ of length $N = T$ to an array $d(n)$ of the same length. The sorting depends on T. If T is even, the array of the even-indexed samples is repeated once, in order to form an array of length T (compare the first period of $c_f(n)$ and $d(n)$ in figure 4). If T is odd, the array of the even-indexed samples is concatenated with the array of the odd-indexed samples, yielding again an array of length T (compare the first period of $c_f(n)$ and $d(n)$ in figure 5).

These explanations can be summarized by the mathematical representation

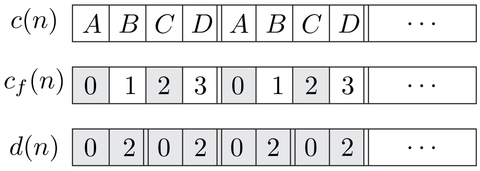


Fig. 4. Example: Period is even, gray: even-indexed samples

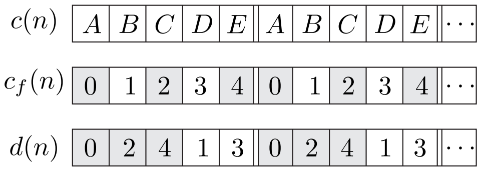


Fig. 5. Example: Period is odd, gray: even-indexed samples

$$c_f(n) = c(n) * a(n) \tag{3}$$

$$d(n) = \underbrace{[d(0), \ldots, d(T-1)]}_{T \text{ samples}}$$
$$= \begin{cases} [\, \overbrace{c_{f,even}(m)}^{\frac{T}{2}\text{ samples}}, \overbrace{c_{f,even}(m)}^{\frac{T}{2}\text{ samples}} \,] & ,T \text{ even} \\ [\, \underbrace{c_{f,even}(m)}_{\lceil \frac{T}{2} \rceil \text{ samples}}, \underbrace{c_{f,odd}(l)}_{\lfloor \frac{T}{2} \rfloor \text{ samples}} \,] & ,T \text{ odd} \end{cases} \tag{4}$$

where $c_{f,even}(m)$ are the even-indexed samples of $c_f(n)$ and $c_{f,odd}(l)$ are the odd-indexed samples of $c_f(n)$. The FIR impulse response $a(n)$ is the high pass or low pass filter impulse response respectively.

This decomposition results in a subspace tree similar to the ordinary WP tree. However, the number of coefficients N_c of each node is constant $N_c = T$ now. With the Period Wavelet Packet (PWP) transform (3)–(4), the coefficient energy of a single period of arbitrary length can be analyzed. The absolute value of the PWP coefficients

$$P_k(d,k,n) = |\,\mathrm{PWP}\,\{\,[e(0 + k\cdot T), \ldots$$
$$\ldots, e(T - 1 + k \cdot T)]\,\}\,| \quad ,$$

where $e(n)$ is the periodic noise and the operation $|\cdot|$ returns the absolute values, represents the time-frequency coefficient energy distribution for the k-th period of noise. It can be used for the adaptive estimation equation (2) of $F_k(d,b,n)$. The PWP $F_k(d,b,n)$ is then the adaptively estimated time-frequency coefficient energy distribution after k periods. It is called "filter packet" because it will be used to threshold the WP coefficients of the noisy speech.

3.2 Denoising

If a noisy signal $y(n)$ (i.e. a corrupted speech frame) of dyadic length $N = 2^l, l \in \mathbb{N}$

$$y(n) = x(n) + e(n) \quad ,$$

$c_f(n)$	0	1	2	3	4	0	1	2	0	1	$\cdots$

(a) Cyclic convolution, repetition error

$c_f(n)$	0	1	2	3	4	0	1	2	3	4	$\cdots$

(b) Corrected convolution

Fig. 6. Example: Signal of length $N = 8$, period $T = 5$, $\Delta T = 3$, gray: repetition

where $x(n)$ is the clean signal (speech) and $e(n)$ the periodic noise, is to be denoised, the noisy signal $y(n)$ can be decomposed into a regular WP. The filter packet $F_k(d, b, n)$ (PWP representation) can be translated into a corresponding WP $F_k^{WP}(d, b, n)$ as follows. Each node of the filter packet has T coefficients. The number of coefficients of a node of WP$\{y(n)\}$ in depth d is $N_c(d) = N/2^d$. If $N_c(d) > T$ for a given node of the tree, the coefficients of the filter packet can be repeated to create an array of length $N_c(d)$. If $N_c(d) < T$ for a given node, the coefficient array of the filter packet must be cut to length $N_c(d)$. This filter packet $F_k^{WP}(d, b, n)$ in the WP domain can be used to extract thresholds for each coefficient of the basis selected by the best basis algorithm.

However, the cyclic convolution of the regular WP transform has to be changed. The signal length N of $y(n)$ can be written as a multiple of the period plus a fraction ΔT of the period:

$$N = k \cdot T + \Delta T \quad , 0 \leq \Delta T < T$$

with usually $k > 0$ and $\Delta T > 0$. The incomplete period ΔT must be regarded by the cyclic convolution. This is best shown by the example in figure 6. If the signal is repeated cyclically (figure 6a), the unfinished period is not completed correctly. Instead of a cyclic repetition, the signal should be repeated from the $\Delta T + 1$ sample on, thus completing the unfinished period correctly (figure 6b). Unfortunately, this can only be done down to a critical depth

$$D_{crit} = \lfloor \log_2 N - \log_2 T + 1 \rfloor$$

because the coefficient array length of nodes in depth D_{crit} or deeper is shorter than the period T:

$$N_c(d \geq D_{crit}) < T \quad .$$

To solve this problem, a structure can be build composed of a WP above depth D_{crit} and a PWP below. This composite packet is shown in figure 7. All black nodes are of length

$$N_c(d < D_{crit}) = N/2^d > T$$

and are processed by WP filter stages (filtering & downsampling). The convolution can be corrected. The gray nodes are of length

$$N_c(d \geq D_{crit}) = T$$

and are processed by PWP filter stages (filtering & sorting). After the calculation, this WP/PWP composition can be translated into a regular WP the same way as the filter packet.

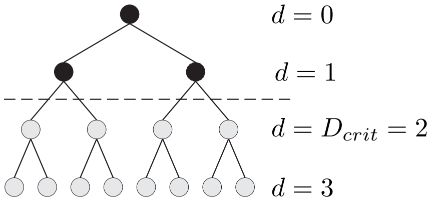


Fig. 7. Example for a composition of WP / PWP, $D_{crit} = 2$, black: WP nodes, gray: PWP nodes

The thresholding paradigm (section 2) can be applied now. The individual threshold for each coefficient is extracted from the filter packet $F_k^{WP}(d, b, n)$ in the WP domain. This is represented by

$$WP_{filt}(d, b, n) = \begin{cases} WP(d, b, n) & , |WP(d, b, n)| \geq F_k^{WP}(d, b, n) \\ 0 & , |WP(d, b, n)| < F_k^{WP}(d, b, n) \end{cases} \quad (5)$$

for hard thresholding but every other thresholding function could be used. The inverse wavelet packet transform for the "best basis" finally yields the filtered signal.

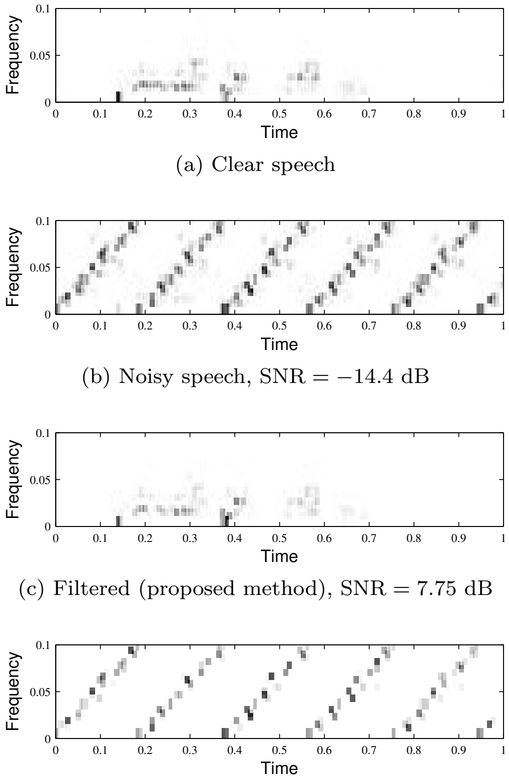
4. VALIDATION

The effectiveness of the proposed method compared to the classical (stationary) wavelet thresholding scheme is shown for a nonstationary chirp signal. Afterwards, the deterioration of the performance for period variations is examined. Finally, a real engine noise is used to evaluate the algorithm performance compared to classical wavelet thresholding.

All time and frequency axes in this section are normalized to observation time and Nyquist frequency respectively. "Symmlet 4" filters were used in the filter bank and the garrote function (Ayat *et al.*, 2004) was chosen as thresholding function.

4.1 Chirp noise without variations

The adaptive PWP noise estimation is applied to a nonstationary noise problem. The disturbance in this example is a series of chirps. The period does not vary for this example. The coefficient energy of the clear speech is shown in the phase plane in figure 8a. The phase plane of the noisy speech is shown below in figure 8b. The Signal-to-Noise-Ratio (SNR) of the noisy speech in time domain is −14.4 dB. A PWP filter packet is calculated from 20 periods of noise. Thresholding the WP coefficients according to the filter packet yields



(a) Clear speech

(b) Noisy speech, SNR = −14.4 dB

(c) Filtered (proposed method), SNR = 7.75 dB

(d) Classical WP thresholding, SNR = −9.87 dB

Fig. 8. Phase planes for chirp disturbances without period variations

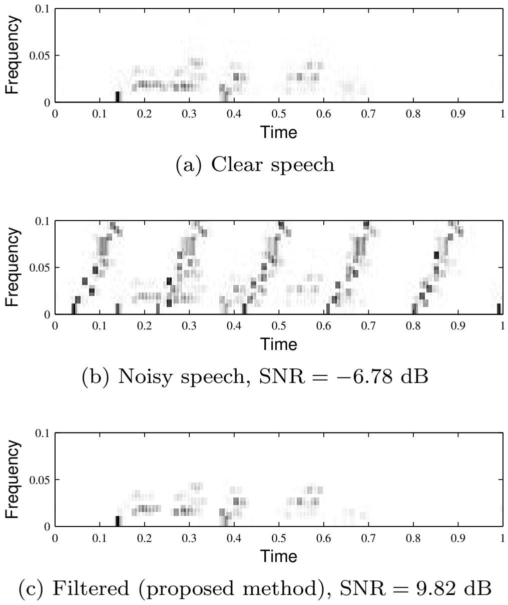
the filtered coefficients represented by figure 8c. The SNR was improved to 7.75 dB.

It is clear that the stationary thresholding method gives poor results for this nonstationary noise. Each frequency subband is shrinked by a fixed threshold over time. The SNR of the stationary filter method for this example is −9.87 dB. The corresponding phase plane is shown in figure 8d.

4.2 Chirp noise with variations

The disturbance in this example is again a series of chirps. Figure 9 shows the phase planes for no period variations. The SNR is increased from −6.78 dB to 9.82 dB. The starting point of each chirp period is now varied at random with a variance of 1% of the period length. The phase plane of the noisy speech is shown in figure 10a. Thresholding the WP coefficients yields the filtered coefficients represented by figure 10b. The SNR was improved from −6.78 dB to 9.18 dB. This is still a very good result even though not as good as without period variations.

Using a period variation of 5% of the period length yields the noisy and filtered phase plane in figure 11. Residues of some chirps can be seen in the phase plane 11b. The SNR deteriorates from 9.18 dB for the 1% case to 4.38 dB. This was expected because the assumption of periodicity was not satisfied as good as before. An additional problem is, that the discrete wavelet transform is not shift-invariant.

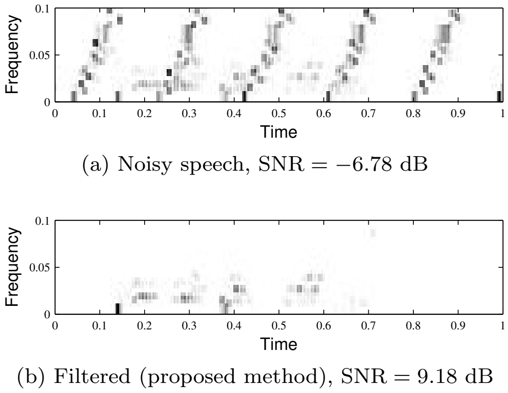


(a) Clear speech

(b) Noisy speech, SNR = −6.78 dB

(c) Filtered (proposed method), SNR = 9.82 dB

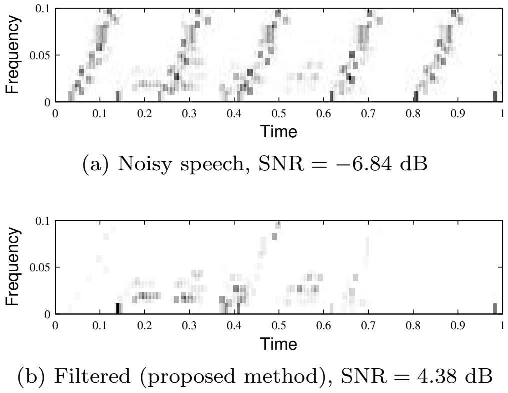
Fig. 9. Phase planes for chirp disturbances, no period variations



(a) Noisy speech, SNR = −6.78 dB

(b) Filtered (proposed method), SNR = 9.18 dB

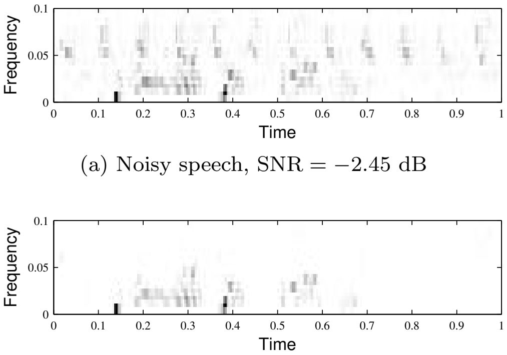
Fig. 10. Phase planes for chirp disturbances, small period variations



(a) Noisy speech, SNR = −6.84 dB

(b) Filtered (proposed method), SNR = 4.38 dB

Fig. 11. Phase planes for chirp disturbances, medium period variations



(a) Noisy speech, SNR = −2.45 dB

(b) Filtered (proposed method), SNR = 5.59 dB

Fig. 12. Phase planes for engine disturbance

4.3 Comparison for engine noise

The advantage of the proposed methods over stationary wavelet thresholding can also be shown for real engine noise. A sample of engine noise was taken from a noise database and used to calculate filter packets and frequency band thresholds. The subsequent filtering gave a SNR of 5.59 dB for the proposed method and a SNR of 3.6 dB for the stationary WP filtering. The corresponding phase planes are shown in figure 12.

5. CONCLUSIONS

In this paper, the problem of nonstationary noise was regarded for the class of periodic noise. The noise was reduced by wavelet thresholding. However, the coefficient energy distribution of a noise period was adaptively estimated and applied to the filtering, instead of just applying a constant threshold for each subband. For these purposes, a method was proposed to calculate the WP transform of a periodic signal of arbitrary period length. The thresholding results based on the proposed Period Wavelet Packet showed superior performance over common WP thresholding in a nonstationary noise environment. The proposed method was not yet compared to adaptive filters, which will be part of future work. However, adaptive filters should not yield better results for the given examples because the signal spectra change too quickly.

Further improvements to this method could be achieved by applying a complex wavelet transform to the analysis of the noisy speech. Complex wavelet transforms are (nearly) shift invariant with an redundancy factor of only 2 (Selesnick *et al.*, 2005), promising good results by only moderate computation costs. A shift-invariant transform should give more robustness against period variations of real signals.

REFERENCES

Ayat, Saeed, Mohammad T. Manzuri and Roohollah Dianat (2004). Wavelet based speech enhancement using a new thresholding algorithm. In: *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing.*

Coifman, Ronald R. and Mladen Victor Wickerhauser (1992). Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory* **38**(2), 713–718.

Donoho, David L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory* **41**(3), 613–627.

Johnstone, I. M. and B. W. Silverman (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society B* **59**(2), 319–351.

Mallat, Stéphane (1999). *A wavelet tour of signal processing.* 2. ed. ed.. Academic Press.

Selesnick, I., R. Baraniuk and N. Kingsbury (2005). The Dual-Tree Complex Wavelet Transform. *IEEE Signal Processing Magazine* **22**(6), 123–151.

Shao, Yu and Chip-Hong Chang (2005). A versatile speech enhancement system based on perceptual wavelet denoising. In: *IEEE International Symposium on Circuits and Systems.* Vol. 2. pp. 864 – 867.

Sheikhzadeh, Hamid and Hamid Reza Abutalebi (2001). An improved wavelet-based speech enhancement system. In: *EUROSPEECH 2001.* pp. 1855–1858.