

# Perceptual Wavelet Adaptive Denoising of Speech

Qiang Fu, Eric A. Wan

Center for Spoken Language Understanding  
OGI School of Science and Engineering, Oregon Health & Sciences University, USA  
q.fu@ieee.org, ericwan@ece.ogi.edu

## Abstract

This paper introduces a novel speech enhancement system based on a wavelet denoising framework. In this system, the noisy speech is first preprocessed using a generalized spectral subtraction method to initially lower the noise level with negligible speech distortion. A perceptual wavelet transform is then used to decompose the resulting speech signal into critical bands. Threshold estimation is implemented that is both time and frequency dependent, providing robustness to non-stationary and correlated noisy environments. Finally, to eliminate the “musical noise” artifact, we apply a modified Ephraim/Malah suppression rule to the thresholding operation - adaptive denoising. Both objective and subjective experiments prove that the new speech enhancement system is capable of significant noise reduction with little speech distortion.

## 1. Introduction

We assume the sampled noisy speech signal  $y_k$  is generated from

$$y_k = s_k + \sigma_k \cdot n_k, \quad k = 0, \dots, K-1 \quad (1)$$

where  $s_k$  is the clean speech signal,  $n_k$  represents an independent noise source with unit variance ( $\sigma_n^2 = 1$ ), and  $\sigma_k$  is the noise level. Wavelet denoising is a non-parametric estimation method that has been proposed in recent years for speech enhancement applications [1, 2]. The goal of wavelet denoising is to optimize the mean-squared error (MSE)

$$E \left[ \|\hat{\mathbf{s}} - \mathbf{s}\|^2 \right] = \sum_{k=0}^{K-1} E (\hat{s}_k - s_k)^2 \quad (2)$$

subject to the side condition that with high probability, the estimation  $\hat{\mathbf{s}}$  is at least as smooth as  $\mathbf{s}$  [3]. This constraint provides an optimal trade-off between the bias and variance of the estimate by keeping the two terms the same order of magnitude [4]. Wavelet denoising is also motivated by these observations: 1) The decorrelating property of a wavelet transform creates a sparse signal: most coefficients are close to zero; 2) Noise is spread out equally over all coefficients, while speech plus noise are concentrated in only a few coefficients. The implementation of wavelet denoising is a

three-step procedure involving wavelet decomposition, nonlinear thresholding and wavelet reconstructing [3].

In this paper, we introduce our perceptual wavelet adaptive denoising (PWAD) system of speech, which involves a spectral-subtraction preprocessing stage, a perceptual wavelet decomposition, a new adaptive threshold estimation technique using a quantile-based noise level estimator, followed by a novel application of the Ephraim/Malah suppression rule (EMSR) [5, 6] to adapt the thresholding techniques. The motivations and adopted techniques will be detailed in section 2. In section 3, the performance of the new system is evaluated using both artificially corrupted and real noisy speech. Conclusions will be given in section 4.

## 2. Description and analysis of algorithm

The system structure is shown in Figure 1. In order to initially reduce the noise level, the noisy speech is first preprocessed with a generalized spectral subtraction routine. A perceptual wavelet transform is then applied to decompose the noisy signal into critical subbands. To account for non-stationary and correlated noise, thresholds are independently estimated for each time frame and wavelet decomposition subband. This is further refined using an adaptive thresholding approach based on a modified EMSR. Finally, the inverse wavelet transform synthesizes the enhanced speech. Each of these stages is detailed further in the following sections.

### 2.1. Preprocessing with generalized spectral subtraction

Wavelet denoising has the best performance when the noise level is not *too* high [4]. Accordingly, the purpose of preprocessing is to initially lower the noise level of  $y_k$  while minimizing the distortion in  $\hat{s}_k$  (we denote  $y'_k$  as the output of this preprocessing stage). For this we implement a generalized spectral subtraction algorithm proposed by Bai and Wan [7]. This combines a state-of-the-art spectral subtraction routine based on a modified Ephraim/Malah suppression rule (EMSR) with a new quantile-based noise spectrum estimator to track the slowly varying non-stationary noise statistics. Experiments show that this generalized spectral subtraction algorithm can achieve moderate levels of noise suppression with very little distortion to the speech signal. Note, the performance advantages of using this preprocessing stage will be discussed in section 3.

### 2.2. Perceptual wavelet transform

In this paper, speech is enhanced for improving the subjective auditory quality. Classical enhancements systems are based

---

This work was supported in part by the NSF under grant IRI-9712346

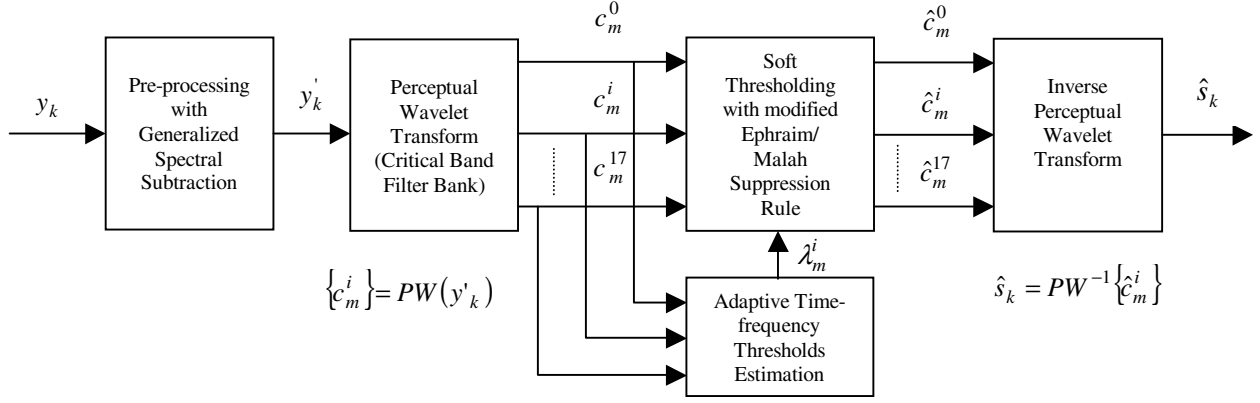


Figure 1: Algorithm structure of speech enhancement based on perpetual wavelet denoising

on uniformly spaced frequency resolution. However, the human ear is often compared to analysis filter-banks with a non-uniform frequency resolution [8]. Thus we use a perceptual-motivated wavelet transform to obtain an auditory relevant frequency resolution as critical band. Specifically, we utilize a wavelet packet (WP) decomposition designed to mimic the auditory critical bands. The implementation, first proposed by Black and Zeytinoglu for coding [9], is based on an efficient 6-stage tree structure decomposition using 16-tap FIR filters derived from the Daubechies wavelet and provides for an exact invertible decomposition.

This perceptual wavelet (PW) transform is used to decompose  $y_k$  into subbands,

$$\{c_m^i\} = PW(y'_k), \quad i = 0, \dots, 17 \quad (3)$$

where  $\{c_m^i\}$  are the decomposition coefficients with index  $i$  corresponding to the subband, and  $m$  corresponding to the 'time' location. For 8kHz speech, the decomposition results in 18 critical bands. Note that the downsampling operation in the multi-level wavelet packet transform results in a multi-rate signal representation (i.e., the resulting number of samples corresponding to index  $m$  differ for each subband  $i$ ). Figure 2a illustrates a time-frequency representative of a noisy speech signal.

### 2.3. Time-frequency dependent adaptive threshold estimation

Wavelet denoising involves *thresholding* in which coefficients below a specified value (i.e., threshold) are set to zero. This is called *hard*-thresholding. Alternatively, *soft*-thresholding simply shrinks or scales coefficients below the threshold value [3]. Donoho and Johnstone derived a general optimal *universal* threshold for the Gaussian white noise under a MSE criterion described in equation (2) and its side condition [10]. However, in practice this threshold is not ideal for speech signals due the poor correlation between MSE and subjective quality and the more realistic presence of correlated noise.

Here we use a new adaptive time-frequency dependent thresholds estimation method. This involves first estimating the standard deviation of the noise,  $\sigma$ , for every subband and time frame. For this we adapt a quantile-based noise tracking approach [7]. Given  $\sigma$ , we calculate the threshold,  $\lambda$ , again for each subband and time frame.

We start by segmenting each  $i$ -th subband of decomposed coefficients,  $c_m^i$ , into frames of length  $L_{frm}^i$ . Denote  $\hat{\sigma}^{i,p}$  as the corresponding estimated noise level of the  $p$ -th frame in the  $i$ -th subband. These are estimated using the segment of previous data  $\{c_m^{i,p}, m = 0, \dots, L_{seg}^i - 1\}$ , where  $L_{seg}^i > L_{frm}^i$  (see Figure 3). The quantile approach requires sorting the data, i.e.,  $c_0^{i,p} < c_1^{i,p} < \dots < c_{L_{seg}^i - 1}^{i,p}$ . Given a  $q$  value ( $0 < q < 1$ ), the quantile for this segment is  $c_{\text{int}(q \cdot L_{seg}^i)}^{i,p}$ , where  $\text{int}(\cdot)$  rounds to the nearest integer value. The noise estimate is then given as

$$\hat{\sigma}^{i,p} = \beta \cdot \sum_{j=0}^{\text{int}(q \cdot L_{seg}^i)} c_j^{i,p} / L_{seg}^i \quad (4)$$

where the constant  $\beta$  is an appropriate scale factor. Nominal values:  $q = 0.2$ ,  $\beta = 0.38$ . The corresponding time lengths of  $L_{seg}$  and  $L_{frm}$  are 512ms and 64ms respectively, and the frame shift is 32ms.

Finally, the threshold for each subband at the  $p$ -th frame,  $\lambda^{i,p}$ , is estimated as in [3, 4, 10],

$$\lambda^{i,p} = \sqrt{2 \log(L_{seg}^i \log_2(L_{seg}^i))} \cdot \hat{\sigma}^{i,p} \quad (5)$$

The solid curve in Figure 2(b) illustrates an estimated threshold sequence for one of the subbands.

### 2.4. Adaptive thresholding with modified Ephraim/Malah suppression rule

Our experiments show that the synthesized speech using a classical hard or soft thresholding operation can result in an unnatural quality similar to the "musical noise" artifacts common to spectral subtraction techniques. To overcome this we implement a new wavelet thresholding technique by modifying an approach by Ephraim and Malah developed specifically to suppress the musical noise in spectral subtraction (note, in practice we use a simpler approximation to the Ephraim and Malah algorithm as proposed in [6].)

In section 2.3, we determined an initial threshold  $\lambda^{i,p}$  for the  $p$ -th frame in  $i$ -th level. For simplicity, let  $\lambda_m^i = \lambda^{i,p}$  if  $m$  falls into  $p$ -th frame. Define the *a posteriori* Coefficient to Threshold Ratio (CTR) (this is analogous to SNR used in the Ephraim and Malah rule),

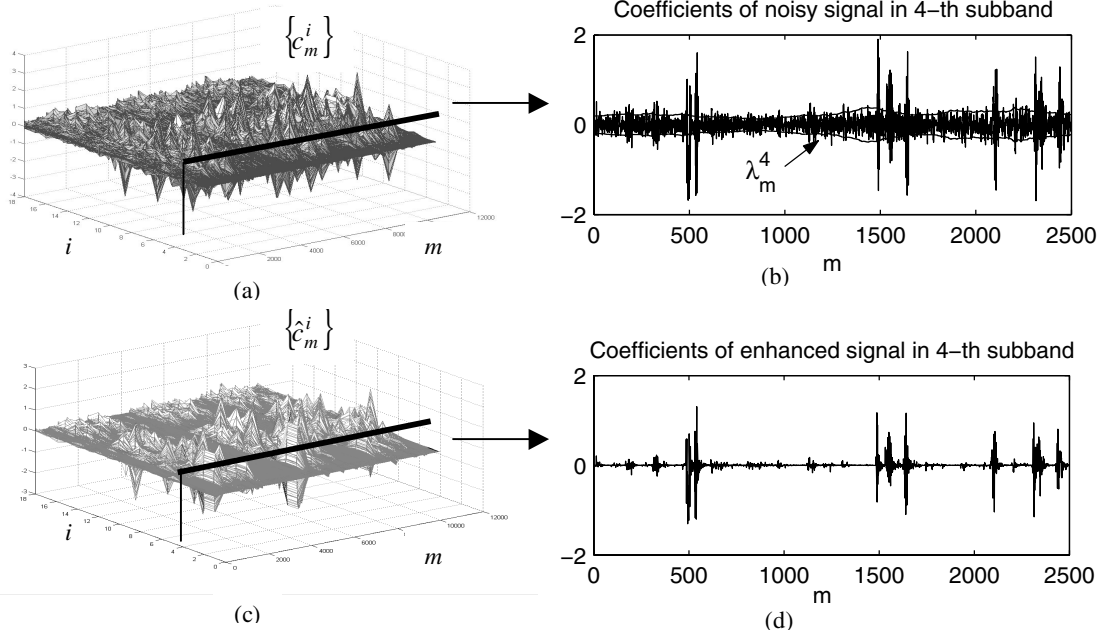


Figure 2: Time-frequency representative of noisy and enhanced speech signal. (a) time-frequency representative of noisy speech (b)  $c_m^4$ -coefficients of noisy speech in 4-th subband (c) time-frequency representative of enhanced speech and (d)  $\hat{c}_m^4$  - coefficients of enhanced speech in 4-th subband.

$$(R_m^i)^{post} = \frac{|c_m^i|}{\lambda_m^i} \quad (6)$$

The corresponding *a priori* CTR is then given by a ‘decision – directed’ estimation [5]:

$$(R_m^i)^{prio} = \alpha \frac{|\hat{c}_{m-L_{fm}}^i|}{\lambda_{m-L_{fm}}^i} + (1-\alpha) \max[0, (R_m^i)^{post} - 1] \quad (7)$$

where  $\alpha$  ( $0 \leq \alpha \leq 1$ ) is used to control the degree of suppression (nominal setting:  $\alpha = 0.5$ ). With the *a priori* and *posteriori* CTRs, a suppression filter can be written as

$$H_m^i = \frac{(R_m^i)^{prio}}{1 + (R_m^i)^{prio}} \left( \frac{1}{(R_m^i)^{post}} + \frac{(R_m^i)^{prio}}{1 + (R_m^i)^{prio}} \right) \quad (8)$$

We then apply above the suppression filter to the decomposed noisy coefficients,  $c_m^i$ , that is

$$\hat{c}_m^i = H_m^i \cdot c_m^i \quad i = 0, \dots, 17 \quad (9)$$

The filter in (8) is not a function of estimated threshold as in classic hard or soft thresholding. The suppression ratio is also adaptively adjusted according to *priori* and *posteriori* CTRs of current frame. Therefore a smoothing effect between

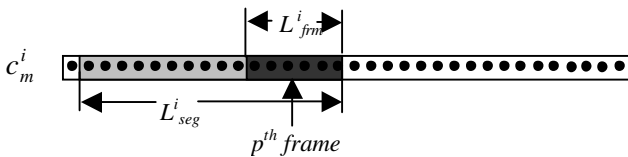


Figure 3: Quantile-based threshold estimation for  $p$ th frame of  $i$ -th subband

successive frames and protection from local overtaking is achieved [5]. Figure 2d again illustrates this process

### 2.5. Inverse perceptual wavelet transform

The last stage simply involves re-synthesizing the enhanced speech using the inverse perceptual wavelet transform,

$$\hat{s}_k = PW^{-1}\{\hat{c}_m^i\} \quad i = 0, \dots, 17 \quad (10)$$

## 3. Performance evaluation and discussions

Figure 4 and 5 give illustrative waveforms resulting from using the PWAD approach on both artificially corrupted and real world noisy speech. Subjective evaluations over a wide range of speech samples recorded in real environments indicate that background noise is greatly reduced with little distortion to speech information (processed audio samples are available on-line at <http://cslu.cse.ogi.edu/research/nse1.htm>).

To further access the performance of the new method and the relative importance of the different processing stages, a clean speech dataset (10 sentences from the TIMIT database - 5 males and 5 females) is corrupted with pink noise for SNR levels ranging from -10 to 10dB. These corrupted sentences are processed with the following algorithm variations:

**Algorithm-1:** The full implementation of the PWAD;

**Algorithm-2:** The quantile-based time-frequency dependent threshold is replaced by a frequency dependent universal threshold;

**Algorithm-3:** The soft thresholding with Ephraim/ Malah suppression rule is replaced by traditional soft thresholding;

**Algorithm-4:** PWAD without generalized spectral subtraction preprocessing;

**Algorithm-5:** Baseline generalized spectral subtraction with quantile-based noise spectrum estimation.

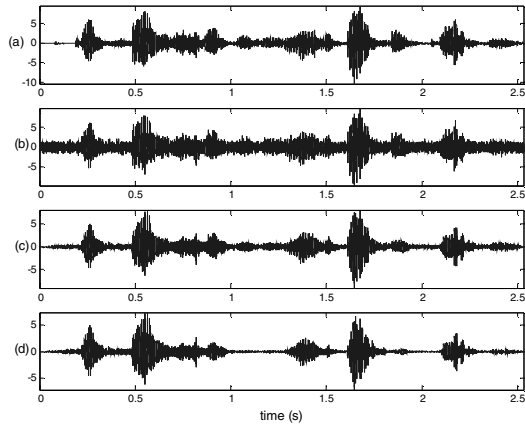


Figure 4: Speech enhancement results: (a) clean speech (b) noisy speech (SNR = 5dB) (c) enhanced speech with normal spectral subtraction and (d) enhanced speech with PWAD

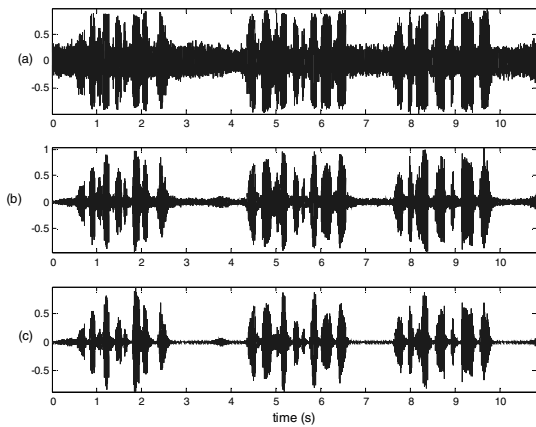


Figure 5: Speech enhancement results: (a) noisy speech (aircraft cockpit), (b) enhanced speech with normal spectral subtraction and (c) enhanced speech with PWAD

Figure 6 shows the improved average *segmental* SNR for each noise level. The wavelet denoising by itself (*i.e.*, w/o preprocessing) tends to work remarkably well on signals with moderate levels of noise. Even further reduction in noise can be achieved using a more aggressive threshold. However, this also leads to an increase in perceived speech distortion, especially when the noise level is high. Thus the current system with the combination of spectral subtraction to initially lower the noise level followed by wavelet denoising appears to produce the best overall performance both objectively and subjectively.

#### 4. Conclusions

Although wavelet denoising provides a theoretical framework to the estimation problem, attributes specific to speech must still be exploited to achieve good performance for the speech enhancement application. In this paper, the spectral-subtraction preprocessing is incorporated to initially reduce the noise level with almost no distortion to speech. The

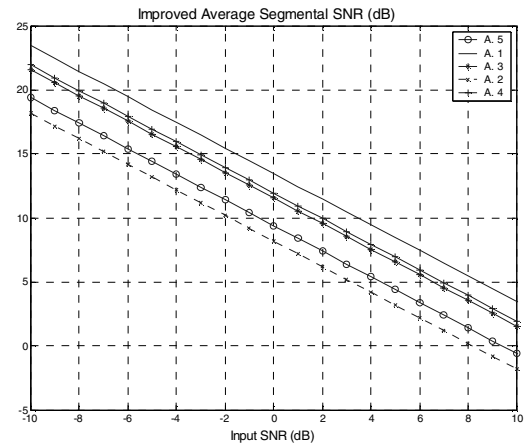


Figure 6: Speech enhancement performance evaluation (pink noise)

perceptual wavelet transform provides an efficient auditory representation. Thresholds are adaptively estimated, providing robustness to non-stationary and correlated noisy environments. Finally, an adaptive thresholding operation using a modified Ephraim/Malah suppression rule is implemented to eliminate the “musical noise” artifact.

#### 5. References

- [1] Gulzow, T. Engelsberg, A. and Heute, U., “Comparison of a discrete wavelet transformation and nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement,” *Signal Processing*, vol. 64, pp5-19, 1998.
- [2] Bahoura, M. and Rouat, J., “Wavelet speech enhancement based on the Teager energy operator”, *IEEE Signal Processing letter*, Vol. 8, No. 1, pp10-12, Jan. 2001.
- [3] Donoho, D. L., “Denoising via soft thresholding”, *IEEE Trans. Information Theory*, 41: 613-627, 1995.
- [4] Jansen, M., *Noise Reduction by Wavelet Thresholding*, Springer-Verlag, New York, 2001.
- [5] Cappe, O., “Elimination of the musical noise phenomenon with the Ephraim and Malah noise Suppressor”, *IEEE Trans. on speech and audio processing*, Vol. 2, No. 2, pp345-349, April 1994.
- [6] Wolfe, P. J., and Godsill, S. J., “Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement”, *IEEE Workshop on Statistical Signal Processing*, pp 496-499, Singapore, August 2001.
- [7] Bai, H. and Wan, E., “Two-pass quantile based noise spectrum estimation”. Submitted to EuroSpeech2003.
- [8] Fillon, T., and Prado, J. “Evaluation of an ERB frequency scale noise reduction for hearing aids: A comparative study”, *Speech Communication*, 39, pp23-32, 2003.
- [9] Black, M., and Zeytinoglu, M., “Computationally Efficient Wavelet Packet Coding of Wide-band Stereo Audio Signals”, *ICASSP 95*, pp. 3075-3078.
- [10] Donoho, D.L, and Johnstone. I. M., “Ideal spatial adaption via wavelet shrinkage”, *Biometrika*, 81:425-455, September 1994.