

Speech Enhancement Based on Wavelet Thresholding the Multitaper Spectrum

Yi Hu, *Student Member, IEEE*, and Philipos C. Loizou, *Member, IEEE*

Abstract—It is well known that the “musical noise” encountered in most frequency domain speech enhancement algorithms is partially due to the large variance estimates of the spectra. To address this issue, we propose in this paper the use of low-variance spectral estimators based on wavelet thresholding the multitaper spectra for speech enhancement. A short-time spectral amplitude estimator is derived which incorporates the wavelet-thresholded multitaper spectra. Listening tests showed that the use of multitaper spectrum estimation combined with wavelet thresholding suppressed the musical noise and yielded better quality than the subspace and MMSE algorithms.

Index Terms—Multitaper method, musical noise, power spectrum estimation, speech enhancement, wavelet thresholding.

I. INTRODUCTION

ALTHOUGH most speech enhancement algorithms improve speech quality, they suffer from an annoying artifact called “musical noise” [1], [2]. Musical noise is caused by randomly spaced spectral peaks that come and go in each frame, and occur at random frequencies. The randomly spaced peaks are due to the inaccurate and large-variance estimates of the spectra of the noise and noisy signals, typically computed using periodogram-type methods.

Several methods have been proposed to reduce musical noise [2]–[5]. Berouti *et al.* [2] suggested spectral flooring any negative spectral estimates and also over-subtracting the noise spectrum by a constant which depended on the segmental SNR. The over-subtraction approach did reduce the musical noise, however it was done at the expense of introducing speech distortion. Other methods focused on masking the musical noise using psychoacoustic models [3], [4]. The minimum mean square error (MMSE) short-time spectral amplitude (STSA) estimator proposed by Ephraim and Malah in [5] was shown by Cappé [6] to eliminate the musical noise via a different mechanism. The MMSE estimator applies a spectral gain $g(\omega_k)$ which is a function of two parameters: the *a priori* signal-to-noise ratio $\gamma_{prio}(\omega_k)$ and the *a posteriori* signal-to-noise ratio $\gamma_{post}(\omega_k)$. Cappé concluded that in the low SNR areas where musical noise frequently dominates, the estimate of the *a priori* SNR proposed in [5] corresponds to a highly smoothed version of the *a posteriori* SNR over successive short-time frames. As

a consequence, the variance of $\gamma_{prio}(\omega_k)$ was much smaller. The fundamental mechanism used in [5] for suppressing the musical noise was therefore the smoothness of $\gamma_{prio}(\omega_k)$. Similar conclusions were also reached by Vary who examined the theoretical limits of spectral-magnitude estimation [7]. Vary explained that in a stationary environment the estimate of the power of the noisy speech signal in a speech-absent frame is not equal to the estimate of the power of the noise signal obtained during silence frames, but fluctuates near the noise estimate. Consequently, the *a priori* SNR estimate fluctuates, and musical noise is produced.

Clearly an accurate estimate of the *a priori* SNR is critical for eliminating the musical noise. Rather than smoothing the *a priori* SNR estimate as done in [5], we take a different approach motivated by the fact that the variance of the *a priori* SNR estimate is greatly influenced by the variance of the spectral estimate of the noisy speech signal. Hence, we focused in this paper on finding spectrum estimators with low variance. In particular, we consider using the multitaper method proposed by Thomson [8] for power spectrum estimation. The multitaper method was shown in [8] to have good bias and variance properties. To further refine the spectral estimate, we wavelet threshold the log multitaper spectra. Unlike others who wavelet denoised the time-domain signal (e.g., [9], [10]), we wavelet denoise the speech spectrum. It should be pointed out that we do not use wavelet denoising techniques to remove the noise, but rather to get better (lower variance) spectral estimates.

This paper is organized as follows. Section II provides background information on low-variance spectrum estimators, and Section III presents the proposed approach which utilizes these spectral estimators. The implementation details are presented in Section IV, the experimental results are given in Section V, and the conclusions are given in Section VI.

II. LOW-VARIANCE SPECTRUM ESTIMATION: BACKGROUND

This section provides background information for the low-variance spectrum estimators used in the proposed approach.

A. Multitaper Spectrum Estimator

Direct spectrum estimation based on Hamming windowing is the most often used power spectrum estimator for speech enhancement. Although windowing reduces the bias, it does not reduce the variance of the spectral estimate [11]. The idea behind the multitaper spectrum estimator [8], [12] is to reduce this variance by computing a small number (L) of direct spectrum estimators each with a different taper (window), and then average the L spectral estimates. The underlying philosophy is similar to Welch’s method of modified periodogram [11]. If the

Manuscript received November 22, 2002; revised August 26, 2003. This work was supported in part by the National Institute of Deafness and Other Communication Disorders/National Institutes of Health under Grant R01 DC03421. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. C.-C. Jay Kuo.

The authors are with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: yihuyxy@utdallas.edu; loizou@utdallas.edu).

Digital Object Identifier 10.1109/TSA.2003.819949

L tapers are chosen to be pairwise orthogonal and properly designed to prevent leakage, then the resulting multitaper spectral estimator will be superior to the periodogram in terms of reduced bias and variance. At best, the variance of the multitaper estimate will be smaller than the variance of each spectral estimate by a factor of $1/L$.

The multitaper spectrum estimator is given by

$$\hat{S}^{mt}(\omega) = \frac{1}{L} \sum_{k=0}^{L-1} \hat{S}_k^{mt}(\omega) \quad (1)$$

with

$$\hat{S}_k^{mt}(\omega) = \left| \sum_{m=0}^{N-1} a_k(m)x(m)e^{-j\omega m} \right|^2 \quad (2)$$

where N is the data length and a_k is the k th data taper used for the spectral estimate $\hat{S}_k^{mt}(\cdot)$, which is also called the k th eigenspectrum. These tapers are chosen to be orthonormal, i.e., $\sum_m a_k(m)a_j(m) = 0$ for $j \neq k$ and equal to 1 for $j = k$. A good set of L orthogonal data tapers with good leakage properties are given by the Slepian or discrete prolate spheroidal sequences (dpss) which are a function of a prescribed mainlobe width W . The number of tapers L is chosen to be less than $2NW$, where W is expressed in units of normalized frequency, i.e., $0 < W < 1/2$ [8]. The Slepian sequences are the unique orthogonal sequences which maximize the spectral concentration of the window mainlobe within $[-W, W]$.

Other taper sequences were also proposed that minimize instead the local bias of the spectral window. In particular, Riedel and Sidorenko [13] proposed the sine tapers given by

$$a_k(m) = \sqrt{\frac{2}{N+1}} \sin \frac{\pi k(m+1)}{N+1}, \quad m = 0, \dots, N-1. \quad (3)$$

The sine tapers were shown in [13] to produce smaller local bias than the Slepian tapers, with roughly the same spectral concentration. For that reason, we adopted the sine tapers in this paper.

B. Refinement of the Spectrum Estimate by Wavelet Thresholding

Recent work has shown that wavelet thresholding techniques can be used to refine the spectral estimate and produce a smooth estimate of the logarithm of the spectrum [14]–[19]. Improved periodogram estimates were proposed in [15]–[17] and improved multitaper spectrum estimates were proposed in [18], [19]. The underlying idea behind these techniques is to represent the log periodogram as “signal” plus the “noise,” where the signal is the true spectrum and “noise” is the estimation error [20]. If the noise is Gaussian, then standard wavelet shrinkage techniques can be used to eliminate the “noise,” by employing for instance simple, level-independent “universal” thresholds [21] to obtain better spectral estimates.

It was shown in [12] that if the eigenspectra defined in (2) are assumed to be uncorrelated, the ratio of the estimated multitaper

spectrum $\hat{S}^{mt}(\omega)$ and the true power spectrum $S(\omega)$ conforms to a chi-square distribution with $2L$ degrees of freedom, i.e.,

$$v(\omega) = \frac{\hat{S}^{mt}(\omega)}{S(\omega)} \sim \frac{\chi_{2L}^2}{2L}, \quad 0 < \omega < \pi. \quad (4)$$

Taking the log of both sides, we get

$$\log \hat{S}^{mt}(\omega) = \log S(\omega) + \log v(\omega). \quad (5)$$

Hence, the log of the multitaper spectrum can be represented as the sum of the true log spectrum plus a noise term, which is $\log \chi^2$ distributed. If L is at least 5, the distribution of $\log v(\omega)$ will be very close to a normal distribution with mean $\phi(L) - \log L$ and variance $\phi'(L)$, where $\phi(L)$ and $\phi'(L)$ denote respectively the digamma and trigamma functions [22]. This means that for all ω (except near $\omega = 0$ and π) the random variable $\eta(\omega)$

$$\eta(\omega) = \log v(\omega) - \phi(L) + \log L \quad (6)$$

will be approximately Gaussian with zero mean and known variance $\sigma_\eta^2 = \phi'(L)$. If $Z(\omega)$ is defined as

$$Z(\omega) = \log \hat{S}^{mt}(\omega) - \phi(L) + \log L \quad (7)$$

then

$$Z(\omega) = \log S(\omega) + \eta(\omega). \quad (8)$$

So, the log multitaper power spectrum plus a constant ($\log L - \phi(L)$) can be written as the true log power spectrum plus a nearly Gaussian noise $\eta(\omega)$ with zero mean and known variance σ_η^2 [18].

The model in (8) is well suited for wavelet denoising techniques [21], [23]–[25] for eliminating the noise $\eta(\omega)$ and obtaining a better estimate of the log spectrum. The idea behind refining the multitaper spectrum by wavelet thresholding can be summarized in four steps.

- 1) Obtain the multitaper spectrum using (1) and (3), and calculate $Z(\omega)$ using (7).
- 2) Apply a standard, periodic Discrete Wavelet Transform (DWT) out to level q_0 to $Z(\omega)$ to get the empirical DWT coefficients $z_{j,k}$ at each level j , where q_0 is specified in advance [26].
- 3) Apply a thresholding procedure to $z_{j,k}$ (the scaling coefficients are kept intact).
- 4) Apply the inverse DWT to the thresholded wavelet coefficients to obtain the refined log spectrum.

One of the key steps in the above process is the thresholding procedure, which is critical to the performance of the proposed algorithm. More details about the various thresholding techniques examined in this paper will be given in Section III-B.

The wavelet thresholding technique is clearly not the only technique that can be used to remove the noise in (8). The subspace method in [27], for instance, could also be used in the frequency domain. The input to the subspace method would be the log of the multi-taper spectrum, and the estimate of the log of the spectrum would be obtained as $\log \hat{S}(\omega) = H \cdot Z(\omega)$, where H

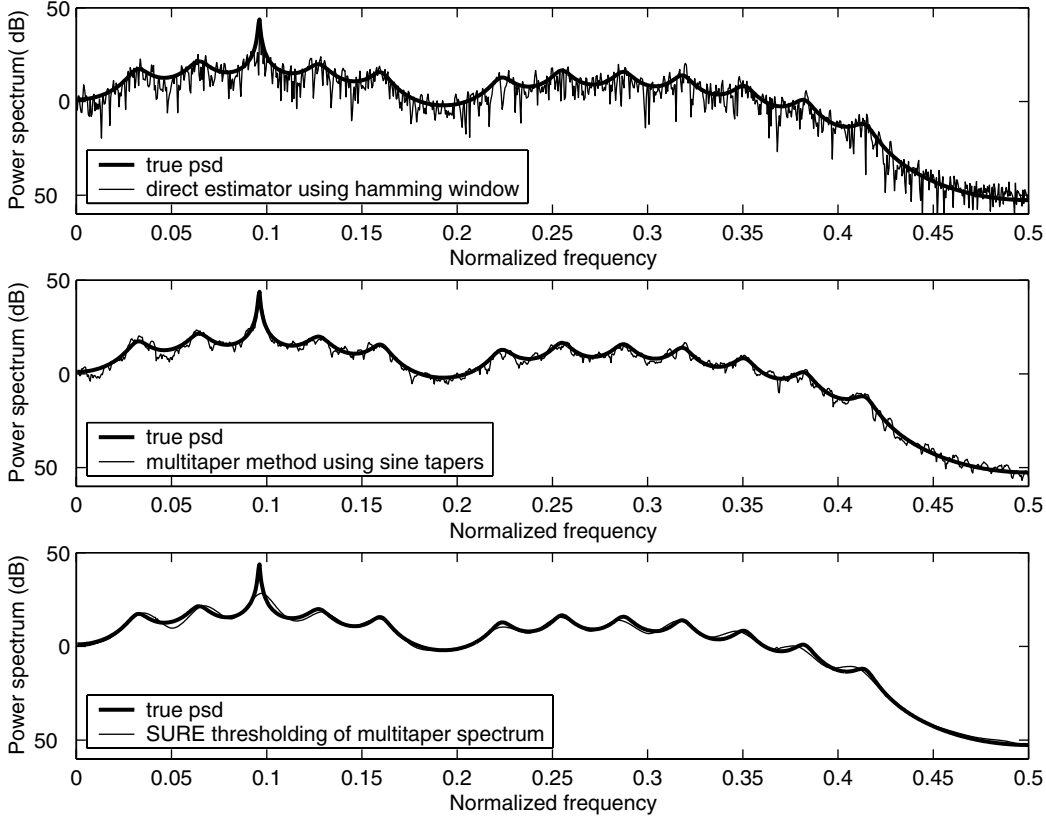


Fig. 1. Comparison of the power spectrum of an AR(24) process [18] estimated by the direct spectrum method using a Hamming window (top panel), the multitaper method using sine tapers (middle panel, with $N = 2048$, $L = 5$) and the SURE wavelet thresholding method (bottom panel, with $q_0 = 5$ and 16-tap symlets wavelets). The thick line shows the true power spectrum of the AR(24) process.

is the linear estimation matrix, which is a function of the signal and noise covariance matrices. The noise covariance matrix is known, in this case, because the autocorrelation of $\eta(\omega)$ can be estimated as in [18].

Fig. 1 shows example spectra of an AR(24) process [18] estimated using the conventional direct spectrum estimator with Hamming window, the multitaper method (with sine tapers) and a wavelet thresholding method (SURE). Clearly, the wavelet thresholding method produced an estimate of the spectrum that was closer to the true spectrum. Also, the resulting spectrum had less variance than the multitaper spectrum.

III. SPEECH ENHANCEMENT BY WAVELET THRESHOLDING THE MULTITAPER SPECTRUM

In this section, we derive the STSA estimator which uses the above mentioned low-variance spectral estimators. The STSA estimator is based on extension of the subspace approach, proposed in [28], to the frequency domain.

A. Proposed STSA Estimator

We assume that the noise signal is additive and uncorrelated with the speech signal, i.e., $\mathbf{y} = \mathbf{x} + \mathbf{n}$, where \mathbf{y} , \mathbf{x} and \mathbf{n} are the N -dimensional noisy speech, clean speech and noise vectors respectively. By denoting the N -point discrete Fourier Transform matrix by F , the Fourier transform of the noisy speech vector \mathbf{y} can be written as $\mathbf{Y}(\omega) = F^H \cdot \mathbf{y} = F^H \cdot \mathbf{x} + F^H \cdot \mathbf{n} = \mathbf{X}(\omega) + \mathbf{N}(\omega)$, where $\mathbf{X}(\omega)$ and $\mathbf{N}(\omega)$ are the $N \times 1$ vectors

containing the spectral components of the clean speech vector \mathbf{x} and the noise vector \mathbf{n} , respectively.

Let $\hat{\mathbf{X}}(\omega) = G \cdot \mathbf{Y}(\omega)$ be the linear estimator of $\mathbf{X}(\omega)$, where G is a $N \times N$ matrix. The error signal obtained in this estimation is given by $\boldsymbol{\varepsilon}(\omega) = \hat{\mathbf{X}}(\omega) - \mathbf{X}(\omega) = \boldsymbol{\varepsilon}_x(\omega) + \boldsymbol{\varepsilon}_n(\omega)$, where $\boldsymbol{\varepsilon}_x(\omega) = (G - I) \cdot \mathbf{X}(\omega)$ represents the speech distortion in the frequency domain and $\boldsymbol{\varepsilon}_n(\omega) = G \cdot \mathbf{N}(\omega)$ represents the residual noise in the frequency domain. After defining the energy of the frequency domain speech distortion as $\varepsilon_x^2(\omega) = E(\boldsymbol{\varepsilon}_x^H(\omega) \cdot \boldsymbol{\varepsilon}_x(\omega))$ and the energy of the frequency domain residual noise as $\varepsilon_n^2(\omega) = E(\boldsymbol{\varepsilon}_n^H(\omega) \cdot \boldsymbol{\varepsilon}_n(\omega))$, we can obtain the optimal linear estimator by solving the following constrained optimization problem:

$$\begin{aligned} & \min_G \varepsilon_x^2(\omega) \\ & \text{subject to: } \frac{1}{N} \varepsilon_n^2(\omega) \leq c \end{aligned} \quad (9)$$

where c is a positive number. It can be shown [29] that the optimal G satisfies the following equation:

$$G(F^H \cdot R_x \cdot F + \mu \cdot F^H \cdot R_n \cdot F) = F^H \cdot R_x \cdot F \quad (10)$$

where μ is the Lagrange multiplier. The above equation can be simplified if we assume that each frequency component is modified individually by a gain, that is, if we assume that G is a diagonal matrix. The matrices $F^H \cdot R_x \cdot F$ and $F^H \cdot R_n \cdot F$ are asymptotically diagonal [30] (assuming that R_x and R_n are Toeplitz) and the diagonal elements of $F^H \cdot R_x \cdot F$ and $F^H \cdot R_n \cdot F$ are

the power spectrum components $S_x(\omega)$ and $S_n(\omega)$ of the clean speech vector \mathbf{x} and noise vector \mathbf{n} , respectively. Denoting the k th diagonal element of G by $g(k)$, then for large N , (10) can be rewritten as

$$g(k) = \frac{S_x(k)}{S_x(k) + \mu \cdot S_n(k)} = \frac{\gamma_{prio}(k)}{\gamma_{prio}(k) + \mu} \quad (11)$$

where $\gamma_{prio}(k) = S_x(k)/S_n(k)$ is the *a priori SNR* at frequency ω_k .

The Lagrange multiplier μ ($\mu \geq 0$) controls the tradeoff between speech distortion and residual noise [28]. This point can be illustrated by noting that μ is related to $\varepsilon_x^2(k)$ and $\varepsilon_n^2(k)$ by

$$\varepsilon_x^2(k) = \sum_k (1 - g(k))^2 S_x(k) = \sum_k \left(\frac{\mu}{\gamma_{prio}(k) + \mu} \right)^2 S_x(k)$$

$$\varepsilon_n^2(k) = \sum_k g^2(k) S_n(k) = \sum_k \left(\frac{\gamma_{prio}(k)}{\gamma_{prio}(k) + \mu} \right)^2 S_n(k).$$

A large μ , for instance, would produce more speech distortion with less residual noise. Conversely, a small μ would produce smaller amount of speech distortion with more residual noise. Ideally, we would like to minimize the speech distortion in speech-dominated frames since the speech signal will mask the noise in those frames. Similarly, we would like to reduce the residual noise in noise-dominated frames. To accomplish that, we can make the value of μ dependent on the estimated *a priori SNR*. We therefore chose the following equation for estimating μ :

$$\mu = \mu_0 - \frac{(\text{SNR}_{dB})}{s} \quad (12)$$

where μ_0 and s are constants chosen experimentally, and $\text{SNR}_{dB} = 10 \log_{10} \text{SNR}$. It is interesting to note that the over-subtraction factor α in [2] plays the same role as the Lagrange multiplier μ in this paper [29].

The power spectrum $S_x(\omega)$ in (11) of the clean speech signal is not available, but in practice can be estimated as:

$$\hat{S}_x(\omega) = S_y(\omega) - \hat{S}_n(\omega) \quad (13)$$

where $\hat{S}_n(\omega)$ denotes the estimate of the noise spectrum obtained during speech-absent frames. As discussed in the introduction, the estimate of the $\gamma_{prio}(k)$ is crucial for eliminating musical noise. In this paper, we considered two different methods for obtaining a good estimate of $\gamma_{prio}(k)$.

In the first method, we form the ratio of the multitaper spectra $\hat{S}_x^{mt}(\omega)/\hat{S}_n^{mt}(\omega)$, and wavelet threshold the log of the ratio of the two spectra to get an estimate of $\gamma_{prio}(k)$. It can be proven (see Appendix) that the log *a priori SNR* estimate, based on multitaper spectra [denoted as $\gamma_{prio}^{mt}(k)$], can be modeled as the true log *a priori SNR* plus a Gaussian distributed noise $\xi(k)$, i.e.,

$$\log \gamma_{prio}^{mt}(k) = \log \gamma_{prio}(k) + \xi(k) \quad (14)$$

where $\xi(k)$ is approximately Gaussian distributed with zero mean and known variance $2\sigma_\eta^2$. Because of the nature of $\xi(k)$, wavelet denoising techniques can be used to eliminate $\xi(k)$.

The second method is based on the assumption that a good estimate of the *a priori SNR*, can be obtained using a good low-

variance spectral estimate of $\hat{S}_x(\omega)$ and $\hat{S}_n(\omega)$. We considered first obtaining the multitaper spectral estimates of $S_y(\omega)$ and $\hat{S}_n(\omega)$ and then wavelet thresholding the log of those estimates individually to obtain $\hat{S}_x(\omega)$. The refined spectrum of $\hat{S}_x(\omega)$, along with the wavelet thresholded estimate of $\hat{S}_n(\omega)$ are then used to obtain a better estimate of the *a priori SNR*.

The above Wiener-type STSA estimator given in (11) is not new, although it was derived differently. What is new, however, is the finding that the log *a priori SNR*, if estimated using multitaper spectra, can be modeled as the true log *a priori SNR* plus a Gaussian distributed noise. No such model was proposed in [5] for the MMSE estimator which was based on STFT analysis of speech. Instead, the authors in [5] proposed a ‘‘decision-directed’’ approach for estimating the *a priori SNR*, in which the *a priori SNR* was updated based on the previous amplitude estimate and the current *a posteriori SNR* estimate. The finding in this paper that the log *a priori SNR* estimate based on multitaper spectra can be modeled as the true log *a priori SNR* plus a Gaussian distributed noise $\xi(k)$ is important for two reasons. First, as indicated earlier, wavelet denoising or other techniques can be used to remove the Gaussian noise and therefore obtain a good estimate of the *a priori SNR*. Second, as demonstrated in [6], having a good estimate of the *a priori SNR* will help eliminate musical noise.

B. Wavelet Thresholding Techniques

Critical to the performance of wavelet denoising techniques is the choice of threshold levels. Several methods have been proposed in the literature for thresholding the wavelet coefficients [21], [23]–[25], [31], [32]. In what follows, we describe the thresholding techniques examined in this paper.

1) *Universal Thresholding Method*: Let $\{z_{j,k}\}$ be the wavelet coefficients of $Z(\omega)$ in (8), let $\{s_{j,k}\}$ be the wavelet coefficients of $\log S(\omega)$ and let $\{n_{j,k}\}$ be the wavelet coefficients of $\eta(\omega)$. Then by the linearity of the discrete wavelet transform, (8) is transformed to

$$z_{j,k} = s_{j,k} + n_{j,k} \quad (15)$$

where the subscript j indicates the j th scale, and the subscript k indicates the k th wavelet coefficient. Since the noise $\eta(\omega)$ is nearly Gaussian, the hard and soft thresholding techniques proposed in [23] can be used for noise reduction. The hard thresholding function is defined for threshold T as

$$\delta_h(z, T) = \begin{cases} z, & \text{if } |z| \geq T \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

and the soft thresholding function is defined as

$$\delta_s(z, T) = \begin{cases} z - T, & \text{if } z \geq T \\ 0, & \text{if } |z| < T \\ z + T, & \text{if } z \leq -T. \end{cases} \quad (17)$$

The main issue in wavelet denoising is determining an appropriate threshold level T . For i.i.d. zero-mean, normally distributed noise with variance σ , Donoho and Johnstone [23] proposed the so-called universal threshold given by $T = \sigma\sqrt{2\log N}$. This threshold is attractively simple since it does not depend on the input data, but on the noise variance,

and works well for uncorrelated noise. As pointed out in [25], if the noise is stationary and colored (as it is in our case), the variance of the noise wavelet coefficients, $n_{j,k}$, will be different for each scale in the wavelet decomposition. Consequently, scale-dependent thresholding was proposed to account for the different variances of the noise wavelet coefficients in each scale. Walden *et al.* [18] extended that idea and derived the variances of $n_{j,k}$ of the nearly Gaussian noise $\eta(\omega)$ in (8). The level-dependent variances of the noise wavelet coefficients, $n_{j,k}$, were estimated according to [18]

$$\text{var}(n_{j,k}) = \sigma_j^2 \equiv \frac{1}{N} \sum_{k=0}^{N-1} S(k) |H_j(k)|^2 \quad (18)$$

where $H_j(k)$ is the frequency response of the length N periodized wavelet filter of level j , and $S(k)$ is the Fourier transform of the autocorrelation function $r_{\eta\eta}$ of the noise $\eta(\omega)$ which is approximated by [18]

$$r_{\eta\eta}(i) = \begin{cases} \sigma_\eta^2 (1 - \frac{|i|}{L+1}), & \text{if } |i| \leq L+1 \\ 0, & \text{otherwise} \end{cases} \quad (19)$$

where L is the number of tapers, and σ_η^2 is the variance of $\eta(\omega)$.

In scale-dependent universal thresholding, the threshold T at each scale j is selected as $T = \sigma_j \sqrt{2 \log N}$. The wavelet coefficients $z_{j,k}$ can be thresholded at each level j using either $\delta_h(\cdot, \cdot)$ or $\delta_s(\cdot, \cdot)$

$$\hat{z}_{j,k} = \begin{cases} \delta(z_{j,k}, T), & \text{if } 1 \leq j \leq q_0 \\ z_{j,k}, & \text{if } j > q_0 \end{cases} \quad (20)$$

where q_0 is some specified coarse resolution level.

2) *SURE Method*: The universal thresholding method tends to use a high threshold level, and in many cases it oversmooths the noisy signal [24]. Better performance, in terms of the mean squared error (MSE), was obtained with smaller thresholds. In [24], Donoho and Johnstone showed that the Stein's unbiased risk estimator (SURE) could be used as the unbiased estimate of the MSE for the soft-thresholding scheme. Johnstone and Silverman [25] later generalized this idea to the case of colored noise, and showed that the SURE method can also be used in the presence of correlated noise.

The Stein's unbiased estimate of the risk for a specific threshold T and input signal $\mathbf{x} = \{x_i\}_{i=1}^N$ using the soft thresholding function is given by [25]

$$\hat{U}(\mathbf{x}, T) = \hat{\sigma}^2 N + \sum_{i=1}^N \{ \min(x_i^2, T^2) - 2\hat{\sigma}^2 I(|x_i| \leq T) \} \quad (21)$$

where $\hat{\sigma}^2$ is the variance of the noise and I is the indicator function ($I(\cdot) = 1$ if $|x_i| \leq T$ and $I(\cdot) = 0$ if $|x_i| > T$). The SURE threshold is obtained as [24], [25]:

$$T = \text{argmin}_{0 \leq T \leq \hat{\sigma} \sqrt{2 \log N}} \hat{U}(\mathbf{x}, T). \quad (22)$$

For level-dependent thresholding, the noise variance $\hat{\sigma}_j^2$ for level j can be obtained using the median absolute deviation (MAD) from zero estimate [25]:

$$\hat{\sigma}_j = \frac{\text{MAD}(z_{j,k})}{0.6745} \quad (23)$$

where 0.6745 is a normalization factor. The operator MAD picks out the median of the absolute values of all the wavelet coefficients $z_{j,k}$ at resolution level j .

The study in [23] showed that when noise dominates the observed data, the universal threshold method performs better, and when the underlying signal dominates the observed data, the SURE method performs better. This observation led to the heuristic SURE method, which selects either the universal threshold or the SURE threshold according to a test of significance presence of the signal. The decision on the choice of threshold is based on comparing $s_d^2 = N^{-1} \sum_{i=1}^N x_i^2 - \sigma^2$ to a threshold $T_d = \sigma (\log_2 N)^{1.5} / \sqrt{N}$ [23]. The threshold T used in the heuristic SURE method is therefore computed as

$$T = \begin{cases} \sigma \sqrt{2 \log N}, & s_d^2 \leq T_d \\ T_{SURE}, & s_d^2 > T_d \end{cases} \quad (24)$$

where T_{SURE} is the SURE threshold obtained according to (22).

IV. IMPLEMENTATION DETAILS

The proposed method can be implemented in four steps. For each speech frame:

Step 1) Compute the multitaper power spectrum $S_{\mathbf{y}}^{mt}$ of the noisy speech \mathbf{y} using (1), and estimate the multitaper power spectrum $S_{\mathbf{x}}^{mt}$ of the clean speech signal by: $S_{\mathbf{x}}^{mt}(\omega) = S_{\mathbf{y}}^{mt}(\omega) - S_{\mathbf{n}}^{mt}(\omega)$, where $S_{\mathbf{n}}^{mt}(\omega)$ is the multitaper power spectrum of the noise. $S_{\mathbf{n}}^{mt}(\omega)$ can be obtained using noise samples collected during speech-absent frames. Here L is set to 5. Any negative elements of $S_{\mathbf{x}}^{mt}(\omega)$ are floored as follows:

$$S_{\mathbf{x}}^{mt}(\omega) = \begin{cases} S_{\mathbf{y}}^{mt}(\omega) - S_{\mathbf{n}}^{mt}(\omega), & \text{if } S_{\mathbf{y}}^{mt}(\omega) > S_{\mathbf{n}}^{mt}(\omega) \\ \beta S_{\mathbf{n}}^{mt}(\omega), & \text{if } S_{\mathbf{y}}^{mt}(\omega) \leq S_{\mathbf{n}}^{mt}(\omega) \end{cases}$$

where β is the spectral floor set to $\beta = 0.002$.

Step 2) Estimate the *a priori* SNR using one of the two methods described in Section III-A. For the second method, for instance, first compute $Z(\omega) = \log S_{\mathbf{y}}^{mt}(\omega) - \phi(L) + \log L$ and then apply the Discrete Wavelet Transform (DWT) of $Z(\omega)$ out to level q_0 to obtain the empirical DWT coefficients $z_{j,k}$ for each level j [q_0 was set to 5, in this paper, based on listening tests]. Eighth-order Daubechie's wavelets with least asymmetry and highest number of vanishing moments, for a given support, were used. Threshold the wavelet coefficients $z_{j,k}$ using one of the two thresholding techniques described in Section III-B, and apply the inverse DWT to the thresholded wavelet coefficients to obtain the refined log spectrum, $\log S_{\mathbf{y}}^{wmt}(\omega)$, of the noisy signal. Repeat the above procedure to obtain the refined log spectrum, $\log S_{\mathbf{n}}^{wmt}(\omega)$, of the noise signal. The estimated power spectrum $S_{\mathbf{x}}^{wmt}$ of the clean speech signal can be estimated using (13). The *a priori* SNR $\gamma_{prio}(k)$ estimate for frequency ω_k is computed as $S_{\mathbf{x}}^{wmt}(\omega_k) / S_{\mathbf{n}}^{wmt}(\omega_k)$.

Step 3) Compute the μ value in (11) according to the segmental SNR:

$$\mu = \begin{cases} \mu_0 - \frac{(\text{SNR}_{dB})}{s} & -5 < \text{SNR}_{dB} < 20 \\ 1 & \text{SNR}_{dB} \geq 20 \\ \mu_{\max} & \text{SNR}_{dB} \leq -5 \end{cases} \quad (25)$$

where μ_{\max} is the maximum allowable value of μ , which was set to 10, $\mu_0 = (1 + 4\mu_{\max})/5$, $s = 25/(\mu_{\max} - 1)$, $\text{SNR}_{dB} = 10\log_{10} \text{SNR}$ and the SNR is computed as:

$$\text{SNR} = \frac{\sum_{i=0}^{N-1} S_{\mathbf{x}}^{wmt}(i)}{\sum_{i=0}^{N-1} S_{\mathbf{n}}^{wmt}(i)}. \quad (26)$$

Step 4) Estimate the gain function $g(k)$ for frequency component ω_k using (11). Obtain the enhanced spectrum $\hat{X}(\omega_k)$ by $\hat{X}(\omega_k) = g(k) \cdot Y(\omega_k)$. Apply the inverse FFT of $\hat{X}(\omega)$ to obtain the enhanced speech signal.

The above estimator was applied to 32-ms duration frames of the noisy signal with 50% overlap between frames. The enhanced speech signal was combined using the overlap and add method.

V. EXPERIMENTAL RESULTS

Objective and subjective listening tests were conducted to evaluate the quality of the proposed STSA estimator. Ten sentences from the HINT database [33] produced by a male speaker and ten sentences from the TIMIT database produced by ten female speakers were used. Speech-shaped noise and Volvo car interior noise were added to the clean speech files at 5 and 0 dB SNR respectively.

Each noisy sentence was enhanced by five methods: the proposed estimator in (11) using the multitaper spectrum (MT), the proposed estimator in (11) using wavelet-thresholded multitaper spectra with heuristic SURE thresholding (MT_SURE), the proposed estimator in (11) using wavelet-thresholded multitaper spectra with universal level-dependent soft thresholding (MT_UNIV), an improved version of the signal subspace approach (SigSub) proposed in [27] which utilizes the multiwindow covariance matrix estimator as in [34], and the Minimum Mean Square Error Log-Spectral Amplitude Estimator (MMSE-LSA) proposed by Ephraim and Malah in [35]. The second method, discussed in Section III-A, was used to obtain $\gamma_{prio}(k)$ in both MT_SURE and MT_UNIV methods, since no significant differences in quality were noted between the two methods proposed for the estimation of $\gamma_{prio}(k)$.

A. Objective Measure Evaluation

The modified bark spectral distortion (MBSD) measure [36], the overall SNR and the segmental SNR measures were used for evaluation of the proposed approach. The MBSD measure is an improved version of the Bark Spectral Distortion (BSD) [37], which was found to be highly correlated with speech quality [36].

The MBSD measure was first used to assess the effect of the wavelet decomposition level q_0 on algorithm performance.

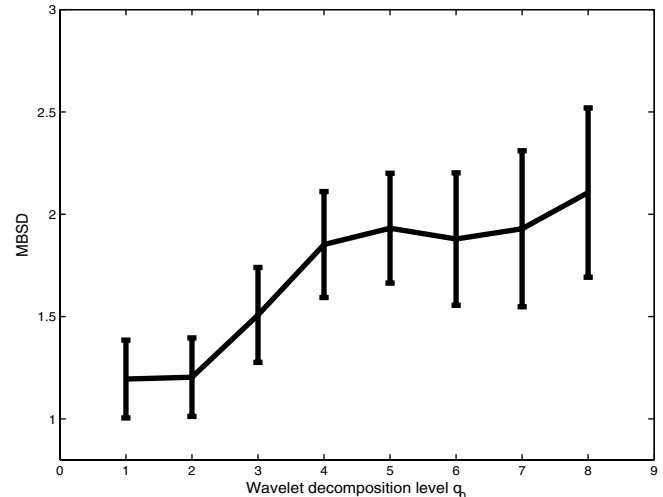


Fig. 2. Performance of the MT_SURE method, in terms of mean MBSD measure for 10 sentences, as a function of the wavelet decomposition level q_0 . Error bars indicate standard deviations.

TABLE I
COMPARATIVE PERFORMANCE, IN TERMS OF MEAN MBSD, OVERALL SNR, AND SEGMENTAL SNR MEASURES, FOR 10 HINT SENTENCES CORRUPTED BY SPEECH-SHAPED NOISE AT 5 dB SNR

	MBSD	SNR	SegSNR
Noisy speech	3.31	5.00	-0.75
MT	1.22	7.43	3.62
MMSE-LSA	2.30	6.85	0.64
MT_UNIV	1.93	6.06	2.94
SigSub [27]	1.58	6.45	2.99
MT_SURE	1.93	5.98	2.91

TABLE II
RESULTS OF LISTENING TESTS IN TERMS OF PERCENTAGE OF TIME THAT LISTENERS PREFERRED THE MT_SURE METHOD OVER THE OTHER METHODS

Noise type	Comparison with MT	Comparison with MT_UNIV	Comparison with SigSub	Comparison with MMSE-LSA
Speech shaped	75%	46%	91%	61%
Car noise	53%	56%	98%	97%

Fig. 2 gives the mean performance of the proposed MT_SURE algorithm, in terms of the MBSD measure, as a function of q_0 , with q_0 varying from 1 to 8. The decomposition level q_0 clearly had an effect in performance. Best performance was obtained for $q_0 < 4$, and slightly worse performance was obtained for $q_0 \geq 4$. Listening tests revealed, however, that the enhanced speech obtained with $q_0 < 4$, had a significant amount of residual noise. For that reason, we set $q_0 = 5$ in all subsequent objective and subjective listening tests.

Table I presents the mean results (using $q_0 = 5$) in terms of the MBSD, overall SNR and segmental SNR measures for 10 HINT sentences corrupted by the speech shaped noise at 5 dB SNR. Best performance was obtained in all three measures by the estimator based on multitaper spectra (MT). There was only a small difference in performance between the two wavelet

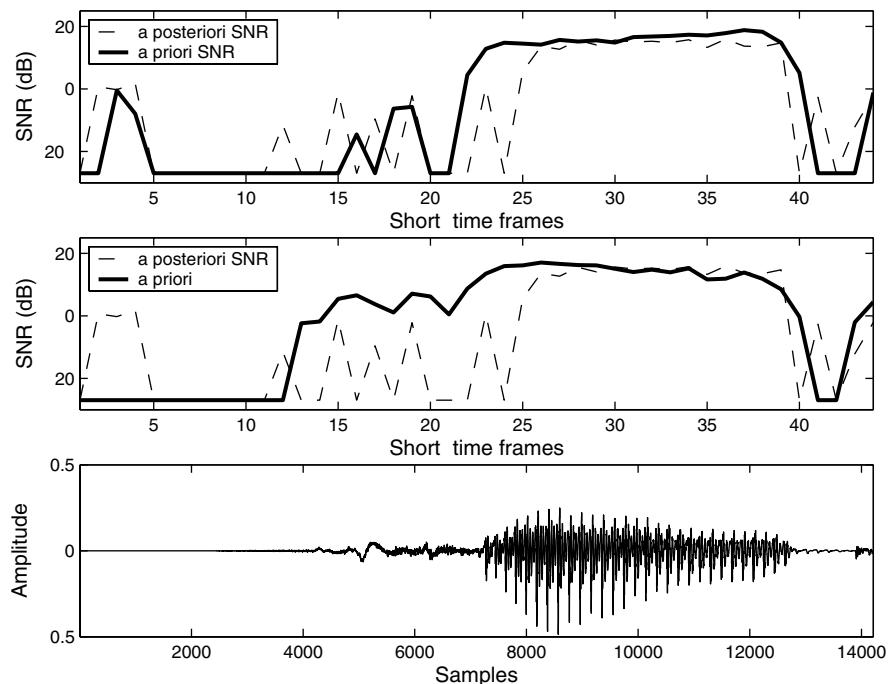


Fig. 3. Estimation of the a priori SNR $\gamma_{prio}(k)$ at frequency 312.5 Hz for the first 44 frames of the word “hayed” corrupted by speech-shaped noise at 15 dB. The top panel shows $\gamma_{prio}(k)$ obtained by the multitaper method using sine tapers. The middle panel shows the $\gamma_{prio}(k)$ obtained by wavelet thresholding the multitaper spectra, and the bottom panel shows the clean speech file. The a posteriori SNR, defined as the ratio of the squared magnitude spectrum of the noisy speech to the noise spectrum [5], is shown in dashed lines for comparison.

thresholding methods (MT_SURE and MT_UNIV). The performance of the MMSE-LSA method was slightly worse in terms of the MBSD measure than the proposed methods, but was significantly worse in terms of the segmental SNR. The performance of the subspace method was comparable to the performance of the proposed methods.

B. Subjective Listening Tests

Given that most objective measures are only partly correlated with speech quality, we conducted listening tests to assess and compare the quality of the proposed methods. Fifteen native speakers of American English participated in the subjective tests [10 listeners participated in the comparison between the MT_SURE method and all other methods except MMSE-LSA, and five listeners participated in the comparison between the MT_SURE method and the MMSE-LSA method] and asked to compare the speech quality of pairs of sentences processed with the above methods. Table II summarizes the subjective evaluation tests for 20 sentences (produced by 10 male and 10 female speakers) in terms of preference percentage. The numbers in Table II indicate the mean percentage of time that the listeners preferred the speech quality of the MT_SURE method over the other methods.

From Table II we can see that listeners preferred the quality of the MT_SURE method over the quality of the MT method when speech was corrupted with speech-shaped noise. Speech enhanced with the MT method had some musical noise. In contrast, speech enhanced with either the MT_SURE or MT_UNIV methods had *no* musical noise. The wavelet thresholding techniques applied to the multitaper spectra eliminated the musical noise. The speech quality of the MT_SURE method was found

to be superior to the quality of the signal subspace method for both types of noise. The speech quality of the MT_SURE method was also found to be markedly better than the quality of the MMSE-LSA method for speech corrupted by car noise. A small advantage was also found for speech-shaped noise. As indicated by the results in Table II, there was a small, but nonsignificant difference, in quality between the two wavelet thresholding techniques. This was consistent with the objective evaluation tests (Table I).

C. Eliminating Musical Noise

Listening tests showed that the musical noise was eliminated when the multitaper spectra were wavelet thresholded. We believe that the mechanisms responsible for that are similar to those in the MMSE method [5] and discussed in [7]. More specifically, we believe it is due to the better estimate of the a priori SNR $\gamma_{prio}(k)$. Fig. 3 shows example plots of the a priori SNR estimation for a single frequency component (312.5 Hz). As can be seen, the estimate of $\gamma_{prio}(k)$ obtained by wavelet thresholding the multitaper spectra was smooth. In contrast, the estimate of $\gamma_{prio}(k)$ obtained with no wavelet thresholding was more erratic. As pointed out by Cappé [6] and Vary [7], it is this erratic behavior that produces the musical noise. Hence, the method which wavelet thresholds the multitaper spectra (MT_SURE or MT_UNIV) eliminated the musical noise not by smoothing directly the $\gamma_{prio}(k)$ as done in [5], but by obtaining better, lower-variance, spectral estimates. Listening tests showed that the reduction in spectral variance produced by the use of multitaper spectral estimators was not sufficient to eliminate the musical noise. Musical noise was eliminated only after wavelet thresholding the multitaper spectra.

VI. SUMMARY AND CONCLUSIONS

A new speech enhancement method was proposed in this paper which, unlike most frequency domain methods, uses low-variance spectrum estimators. The low-variance spectrum estimators were based on wavelet thresholding the multitaper spectrum of speech. It was shown in this paper that the log *a priori* SNR estimate based on multitaper spectra can be modeled as the true log *a priori* SNR plus a Gaussian distributed noise. This is important because wavelet denoising techniques can be used to remove the Gaussian noise and therefore get a better estimate of the *a priori* SNR. Wavelet denoising techniques were used in this paper not to remove the noise from the signal, but rather to produce better, lower-variance, spectral estimates and better *a priori* SNR estimates. Listening tests revealed that the enhanced speech had no musical noise and we attribute that to the use of low-variance spectrum estimators and the better estimate of the *a priori* SNR. Listening tests also showed that the proposed method had superior speech quality to the signal subspace and MMSE-LSA methods.

APPENDIX

In this appendix we show how the *a priori* SNR estimate obtained using the log multitaper spectra can be modeled as the true log *a priori* SNR plus a Gaussian distributed noise.

Denoting the *a priori* SNR estimated by the multitaper spectra as γ_{prio}^{mt} , it is clear that

$$\frac{\gamma_{prio}^{mt}(k)}{\gamma_{prio}(k)} = \frac{\left(\frac{S_x^{mt}(k)}{S_n^{mt}(k)}\right)}{\left(\frac{S_x(k)}{S_n(k)}\right)} = \frac{\left(\frac{S_x^{mt}(k)}{S_x(k)}\right)}{\left(\frac{S_n^{mt}(k)}{S_n(k)}\right)}.$$

Using the relationship in (4), we see that the right hand side is the ratio of two chi-square distributions each with $2L$ degrees of freedom. Taking the log of both sides, we get

$$\log \gamma_{prio}^{mt}(k) = \log \gamma_{prio}(k) + \log \chi_x^2(k) - \log \chi_n^2(k) \quad (27)$$

where $\chi_x^2(k)$ is the chi-square distribution of $S_x^{mt}(k)/S_x(k)$ and $\chi_n^2(k)$ is the chi-square distribution of $S_n^{mt}(k)/S_n(k)$ with $2L$ degrees of freedom. If $L \geq 5$, $\log \chi_x^2(k)$ and $\log \chi_n^2(k)$ are nearly Gaussian with mean $\phi(L) - \log L$ and variance $\phi'(L)$, where $\phi(L)$ and $\phi'(L)$ denote respectively the digamma and trigamma functions [22]. Hence, the above equation can be simplified to

$$\log \gamma_{prio}^{mt}(k) = \log \gamma_{prio}(k) + \xi(k) \quad (28)$$

where $\xi(k)$ is nearly Gaussian with zero mean and variance $2\phi'(L)$.

REFERENCES

- [1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, 1979.
- [2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 208–211, 1979.
- [3] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 479–514, Nov. 1997.
- [4] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 126–137, Mar. 1999.
- [5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, 1984.
- [6] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 345–349, 1994.
- [7] P. Vary, "Noise suppression by spectral magnitude estimation-mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, 1985.
- [8] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE*, vol. 70, no. 9, pp. 1055–1096, Sept. 1982.
- [9] J. W. Seok and K. S. Bae, "Speech enhancement with reduction of noise components in the wavelet domain," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1323–1326, 1997.
- [10] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Processing Lett.*, vol. 8, no. 1, pp. 10–12, Jan. 2001.
- [11] S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [12] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge, MA: Cambridge Univ. Press, 1993.
- [13] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Trans. Signal Processing*, vol. 43, pp. 188–195, 1995.
- [14] B. Lumeau, J. C. Pesquet, J. F. Bercher, and L. Louveau, "Optimization of bias-variance tradeoff in non parametric spectral analysis by decomposition into wavelet packets," in *Progress in Wavelet Analysis and Applications*, Y. Meyer and S. Roques, Eds. Frontières, 1993, pp. 285–290.
- [15] H.-Y. Gao, "Wavelet Estimation of Spectral Densities in Time Series Analysis," Ph.D., Dept. Statist., Univ. California, Berkeley, 1993.
- [16] —, "Choice of thresholds for wavelet shrinkage estimate of the spectrum," *J. Time Series Anal.*, vol. 18, pp. 231–251, 1997.
- [17] P. Moulin, "Wavelet thresholding techniques for power spectrum estimation," *IEEE Trans. Signal Processing*, vol. 42, pp. 3126–3136, Dec. 1994.
- [18] A. T. Walden, D. B. Percival, and E. J. McCoy, "Spectrum estimation by wavelet thresholding of multitaper estimators," *IEEE Trans. Signal Processing*, vol. 46, pp. 3153–3165, 1998.
- [19] A. C. Cristán and A. T. Walden, "Multitaper power spectrum estimation and thresholding: Wavelet packets versus wavelets," *IEEE Trans. Signal Processing*, vol. 50, pp. 2976–2986, Dec. 2002.
- [20] G. Wahba, "Automatic smoothing of the log periodogram," *J. Amer. Statist. Assoc.*, vol. 75, pp. 122–132, 1980.
- [21] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
- [22] M. S. Bartlett and D. G. Kendall, "The statistical analysis of variance heterogeneity and the logarithmic transformation," *Suppl. J. R. Statist. Soc.*, vol. 8, pp. 128–138, 1946.
- [23] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613–627, May 1995.
- [24] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, 1995.
- [25] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. R. Statist. Soc. B*, vol. 59, pp. 319–351, 1997.
- [26] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet presentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, July 1989.
- [27] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Processing Lett.*, vol. 9, pp. 204–206, July 2002.
- [28] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, 1995.
- [29] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 334–341, July 2003.
- [30] R. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, 1972.
- [31] R. R. Coifman and D. L. Donoho, "Translation-Invariant Denoising," Dept. Statistics, Stanford Univ., Stanford, CA, 1995.
- [32] X. Zhang and M. D. Desai, *Adaptive Denoising Based on Sure Risk*, vol. 5, pp. 265–267, Oct. 1998.

- [33] M. Nilsson, S. Soli, and J. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1085–1099, 1994.
- [34] Y. Hu and P. C. Loizou, "A perceptually motivated subspace approach for speech enhancement," in *Int. Conf. Spoken Language Processing*, Denver, CO, 2002, pp. 1797–1800.
- [35] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.
- [36] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 541–544.
- [37] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 819–829, 1992.



Yi Hu (S'01) received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China (USTC), Hefei, in 1997 and 2000, respectively. Currently he is pursuing the Ph.D. degree in electrical engineering at University of Texas at Dallas, Richardson.

His research interests are in the general area of signal processing, ASIC/FPGA design of DSP algorithms, and VLSI CAD algorithms.



Philipos C. Loizou (S'90–M'91) received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from Arizona State University, Tempe, in 1989, 1991, and 1995, respectively.

From 1995 to 1996, he was a Postdoctoral Fellow with the Department of Speech and Hearing Science at Arizona State University, working on research related to cochlear implants. He was an Assistant Professor at the University of Arkansas at Little Rock from 1996 to 1999. He is now an Associate Professor in the Department of Electrical Engineering at the University of Texas at Dallas. His research interests are in the areas of signal processing, speech processing, and cochlear implants.

Dr. Loizou is a member of the Industrial Technology Track Technical Committee of the IEEE Signal Processing Society and was also an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1999–2002).