

Incorporating a Psychoacoustical Model in Frequency Domain Speech Enhancement

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

Abstract—A frequency domain optimal linear estimator is proposed which incorporates the masking properties of the human auditory system to make the residual noise distortion inaudible. The use of wavelet-thresholded multitaper spectra is also proposed for frequency-domain speech enhancement methods as an alternative to the traditional fast Fourier transform (FFT)-based magnitude spectra. Experiments with multitalker babble noise indicated that the proposed estimator outperformed the minimum mean-square error log-spectral amplitude estimator (MMSE-LSA), particularly when wavelet-thresholded multitaper spectra were used in place of the FFT spectra.

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

I. INTRODUCTION

THE KNOWN phenomenon of auditory masking has been successfully applied and used in wideband audio coding. In an effort to make the residual noise perceptually inaudible, more speech enhancement methods today are exploiting the auditory masking properties [1]–[4]. In the subtractive-type approach proposed by Virag [1], for instance, a psychoacoustical model was used to guide the derivation of the spectral subtractive parameters. Heuristic rules were used in [2] to derive spectral subtractive equations that incorporated masking thresholds. A simplified constrained minimization approach was used in [3] to derive a spectral weighting rule which was a function of the masking thresholds.

In most of the above speech enhancement methods, the incorporation of auditory masking was done heuristically. This letter formulates speech enhancement in the frequency domain as a constrained minimization problem and includes the masking thresholds as the constraints. The psychoacoustical model is thereby integrated in the derived spectral weighting function. We further investigate the importance of using good (low variance) spectrum estimators in speech enhancement.

This letter is organized as follows. In Section II, the proposed approach is described. Implementation details are presented in Section III, experimental results are given in Section IV, and the conclusions are given in Section V.

Manuscript received March 31, 2002; revised June 4, 2003. This work was supported in part by the National Institute of Deafness and Other Communication Disorders/National Institutes of Health (NIDCD/NIH) under Grant R01 DC03421. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. See-May Phoong.

The authors are with 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/LSP.2003.821714

II. PROPOSED FREQUENCY DOMAIN SPEECH ENHANCEMENT METHOD

A. Principles of Proposed Method

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 then be written as $\mathbf{Y} = F^H \cdot \mathbf{y} = F^H \cdot \mathbf{x} + F^H \cdot \mathbf{n} = \mathbf{X} + \mathbf{N}$, where \mathbf{X} and \mathbf{N} 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 $\hat{\mathbf{X}} = G \cdot \mathbf{Y}$ be the linear estimator of \mathbf{X} , where G is a $N \times N$ matrix. The error signal obtained in this estimation is given by $\boldsymbol{\varepsilon} = \hat{\mathbf{X}} - \mathbf{X} = \boldsymbol{\varepsilon}_x + \boldsymbol{\varepsilon}_n$, where $\boldsymbol{\varepsilon}_x = (G - I) \cdot \mathbf{X}$ represents the spectrum of the speech distortion and $\boldsymbol{\varepsilon}_n = G \cdot \mathbf{N}$ represents the spectrum of the residual noise. Next, we define the energy of the frequency domain speech distortion as $\varepsilon_x^2 = E(\boldsymbol{\varepsilon}_x^H \cdot \boldsymbol{\varepsilon}_x)$ and the k th spectral component of the residual noise as $\varepsilon_{n,k} = \mathbf{e}_k^T \cdot \boldsymbol{\varepsilon}_n$, where \mathbf{e}_k^T is a selector choosing the k th component of $\boldsymbol{\varepsilon}_n$ and is defined as

$$\mathbf{e}_k^T = [0 \cdots 0 \cdots \underbrace{1}_{k} \cdots 0 \cdots 0]$$

The proposed linear estimator minimizes the frequency domain speech distortion subject to constraints on components of the spectrum of the residual noise. More specifically, we require that the spectral energy of $\varepsilon_{n,k}$ be smaller than or equal to some preset threshold α_k ($\alpha_k \geq 0$), for $k = 1, \dots, N$. As we will show later, these thresholds can be set equal to the masking thresholds. The estimator is obtained by solving the following constrained optimization problem:

$$\begin{aligned} & \min_G \varepsilon_x^2 \\ & \text{subject to } \varepsilon_{n,k}^2 \leq \alpha_k, \quad k = 1, \dots, N \end{aligned} \quad (1)$$

where $\varepsilon_{n,k}^2 = E\{|\varepsilon_{n,k}|^2\}$.

Problem (1) is a convex programming problem, and its solution can be found using the method of Lagrangian multipliers. Specifically, G is a stationary feasible point if it satisfies the gradient equation of the objective function

$$J(G, \mu_1, \mu_2, \dots, \mu_N) = \varepsilon_x^2 + \sum_{k=1}^N \mu_k (\varepsilon_{n,k}^2 - \alpha_k) \quad (2)$$

and

$$\mu_k (\varepsilon_{n,k}^2 - \alpha_k) = 0, \quad \text{for } k = 1, \dots, N \quad (3)$$

where $\mu_k \geq 0$ is the k th Lagrangian multiplier for the constraint on the k th component of $\boldsymbol{\varepsilon}_{\mathbf{n}}$. From $\nabla_G J(G, \mu_1, \mu_2, \dots, \mu_N) = 0$ we have

$$GF^H R_{\mathbf{x}} F + \left(\sum_{k=1}^N \mu_k \mathbf{e}_k \mathbf{e}_k^T \right) GF^H R_{\mathbf{n}} F = F^H R_{\mathbf{x}} F \quad (4)$$

Let Λ_{μ} be a diagonal matrix defined as $\Lambda_{\mu} = \sum_{k=1}^N \mu_k \mathbf{e}_k \mathbf{e}_k^T = \text{diag}(\mu_1, \dots, \mu_N)$, then the above equation can be rewritten as

$$GF^H R_{\mathbf{x}} F + \Lambda_{\mu} GF^H R_{\mathbf{n}} F = F^H R_{\mathbf{x}} F \quad (5)$$

To simplify matters, we assume that G is a diagonal matrix, i.e., we assume that the gain is applied to each frequency component individually. The matrices $F^H \cdot R_{\mathbf{x}} \cdot F$ and $F^H \cdot R_{\mathbf{n}} \cdot F$ are asymptotically diagonal [5] (assuming that $R_{\mathbf{x}}$ and $R_{\mathbf{n}}$ are Toeplitz) and the diagonal elements of $F^H \cdot R_{\mathbf{x}} \cdot F$ and $F^H \cdot R_{\mathbf{n}} \cdot F$ are the power spectrum components $S_{\mathbf{x}}(\omega)$ and $S_{\mathbf{n}}(\omega)$ of the clean speech vector \mathbf{x} and noise vector \mathbf{n} , respectively. Denoting the k th diagonal element of G by $g(k)$, (5) can be rewritten as

$$g(k) \cdot (S_{\mathbf{x}}(k) + \mu_k \cdot S_{\mathbf{n}}(k)) = S_{\mathbf{x}}(k), \text{ for } k = 1, 2, \dots, N$$

The gain function $g(k)$ for the k th frequency component is therefore obtained by

$$g(k) = \frac{S_{\mathbf{x}}(k)}{S_{\mathbf{x}}(k) + \mu_k \cdot S_{\mathbf{n}}(k)} = \frac{\gamma_{\text{prio}}(k)}{\gamma_{\text{prio}}(k) + \mu_k} \quad (6)$$

where $\gamma_{\text{prio}}(k) = S_{\mathbf{x}}(k)/S_{\mathbf{n}}(k)$ is defined as the *a priori* SNR at frequency ω_k .

The above equation reduces to the Wiener filter when $\mu_k = 1$ for all k . The μ_k values, in general, control the steepness of the suppression curves (spectral attenuation vs. SNR level) with large values of μ_k producing much attenuation at low SNR levels and small values of μ_k producing less attenuation. The μ_k values need therefore to be chosen carefully to avoid speech distortion. One possibility is to chose μ_k based on the *a posteriori* SNR of the k th spectral component. In the following section, we show how to chose μ_k based on a psychoacoustic model.

B. Incorporating a Psychoacoustical Model

One can optimally select the μ_k values by exploiting the masking properties of the human auditory system. The human listener will not perceive any noise distortion as long as the power spectrum density of the distortion lies below the masking threshold (the masking thresholds can be obtained by performing critical band analysis of the speech signal [6]).

If we constrain the k th spectral component of the residual noise to be lower than the masking threshold, denoted as T_k , in frequency bin k we can compute the μ_k values that meet this constraint. Assuming that the constraints α_k in (1) are set equal to the masking thresholds T_k , and the equality in (1) is satisfied, then $\varepsilon_{\mathbf{n},k}^2 = \alpha_k$ implies that

$$g^2(k) \cdot S_{\mathbf{n}}(k) = T_k, \text{ for } k = 1, 2, \dots, N$$

Plugging (6) into the above equation, with the condition that $\mu_k \geq 0$, μ_k can be obtained by

$$\mu_k = \max \left(\sqrt{\frac{S_{\mathbf{n}}(k)}{T_k}} - 1, 0 \right) \cdot \frac{S_{\mathbf{x}}(k)}{S_{\mathbf{n}}(k)}$$

In terms of the *a priori* SNR, μ_k can also be expressed as

$$\mu_k = \max \left(\sqrt{\frac{S_{\mathbf{x}}(k)}{T_k}} \cdot \sqrt{\frac{1}{\gamma_{\text{prio}}(k)}} - 1, 0 \right) \cdot \gamma_{\text{prio}}(k). \quad (7)$$

Now, plugging the above equation into (6), $g(k)$ can be rewritten as

$$g(k) = \frac{1}{1 + \max \left(\sqrt{\frac{S_{\mathbf{n}}(k)}{T_k}} - 1, 0 \right)}. \quad (8)$$

It is clear from the above equation, that if the spectrum of the residual noise falls below the masking threshold, the gain $g(k)$ is set to 1, i.e., no attenuation is performed since the k th residual noise spectral component is masked. It should be noted that a similar gain function was derived in [3] using a simplified constrained minimization approach. Unlike our method, their approach was heuristic and was not based on the minimization of an error criterion.

III. IMPLEMENTATION

The computation of the spectral weighting function $g(k)$ in (8) is critical for the performance of the proposed algorithm. It depends largely on accurate estimation of the clean speech spectrum and the noise spectrum. In the following sections, we discuss the computation of the clean and noise spectra.

A. Spectrum Estimation

Pilot informal listening tests indicated that the computation of $g(k)$ is sensitive to the type of spectrum estimator used. In this letter, we focused on finding spectrum estimators that have low variance. More specifically, we considered using the multitaper method proposed by Thomson [7] for spectrum estimation. To further refine the spectrum estimate, the log multitaper spectrum is wavelet thresholded as in [8].

The multitaper spectrum estimator of a signal vector \mathbf{x} is given by

$$\hat{S}^{mt}(\omega) = \frac{1}{L} \sum_{k=0}^{L-1} \hat{S}_k^{mt}(\omega) \quad (9)$$

with

$$\hat{S}_k^{mt}(\omega) = \left| \sum_{m=0}^{N-1} h_k(m) x(m) e^{-j\omega m} \right|^2 \quad (10)$$

where L is the number of tapers, and h_k is the k th data taper used for the spectral estimate $\hat{S}_k^{mt}(\cdot)$. These tapers are chosen

to be orthonormal, and in this letter, we chose the sine tapers proposed by Riedel and Sidorenko [9]

$$h_k(m) = \sqrt{\frac{2}{N+1}} \sin \frac{\pi km}{N+1}, \quad m = 1, \dots, N. \quad (11)$$

The sine tapers were shown in [9] to produce smaller local bias than the Slepian tapers [7] with roughly the same spectral concentration. It was further shown in [8] that if L is chosen to be at least 5, for all ω (except near $\omega = 0$ and π) the logarithm of the multitaper power spectrum in (9) 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 [8], where $\phi(L)$ denotes the digamma function. More specifically, if $Z(\omega)$ is defined as

$$Z(\omega) = \log \hat{S}^{mt}(\omega) - \phi(L) + \log L \quad (12)$$

then

$$Z(\omega) = \log S(\omega) + \eta(\omega) \quad (13)$$

The model in (13) is well suited for wavelet denoising techniques for eliminating the “noise” term $\eta(\omega)$ and obtaining a better estimate of the log spectrum. The idea behind refining the multitaper spectrum by wavelet thresholding can be summarized in the following four steps.

- 1) Obtain the multitaper spectrum using (9)–(11), and calculate $Z(\omega)$ using (12).
- 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 [10].
- 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.

We denote the wavelet denoised multitaper spectrum as $S_y^{\text{wt}}(\omega)$. It should be pointed out that the wavelet denoising is not done to remove the additive noise, but rather to obtain a better (lower variance) estimate of the spectrum.

B. Noise Spectrum Estimation

For nonstationary environments (e.g., multitalker babble) it is imperative to update the estimate of the noise spectrum very often. One such noise estimation method, which was found to work well for nonstationary environments, is the minimum-statistics method originally proposed by Martin [11] and later modified by Cohen and Berdugo [12]. Because of its simplicity, the latter method was chosen in this letter for noise spectrum tracking. The minimum tracking is based on a recursively smoothed spectrum which is estimated using first-order recursive averaging

$$S_y(k, l) = \alpha_s S_y(k, l-1) + (1 - \alpha_s) S_y^{\text{wt}}(k, l)$$

where $S_y(k, l)$ is the k th component of the smoothed noisy speech spectrum at frame l , α_s ($0 < \alpha_s < 1$) is a smoothing

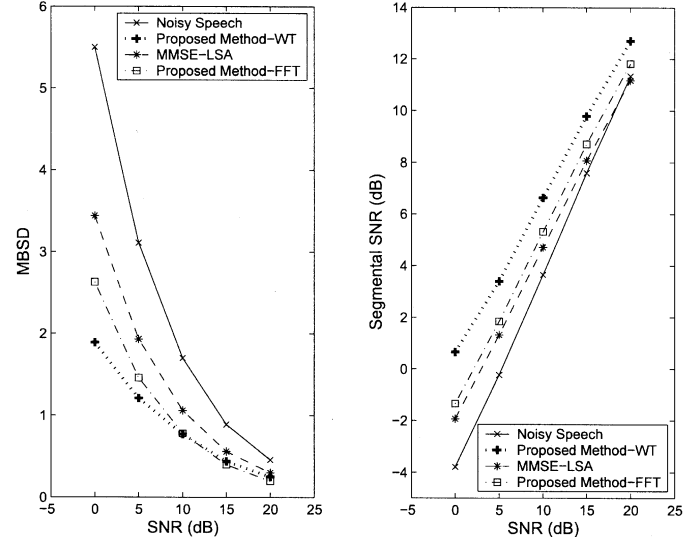


Fig. 1. Comparative performance, in terms of mean MBSD and segmental SNR measures, for 60 HINT sentences corrupted by multitalker babble at 0–20 dB SNR. The + symbols indicate performance obtained with proposed method using wavelet-thresholded multitaper spectra, and the open square symbols indicate performance obtained with the proposed method using FFT magnitude spectra.

factor, and S_y^{wt} is the wavelet-thresholded multitaper spectrum of the noisy speech [note that a different method was used in [12] to obtain a smoothed version of the noisy speech spectrum, in place of $S_y^{\text{wt}}(k, l)$]. The noise spectrum is obtained by tracking the minimum of $S_y(k, l)$ over P frames using a simplified version of the minimum statistics algorithm [12]. The estimated noise spectrum $\hat{S}_n(k, l)$ at frame $l + 1$ is updated according to

$$\hat{S}_n(k, l+1) = \hat{S}_n(k, l) + [1 - \tilde{\alpha}_d(k, l)] S_y^{\text{wt}}(k, l) \quad (14)$$

where $\tilde{\alpha}_d(k, l) = \alpha_d + (1 - \alpha_d)p'(k, l)$ is a time-varying smoothing factor, α_d is an averaging parameter, and $p'(k, l)$ is the conditional signal presence probability updated as in [12].

IV. EXPERIMENTAL RESULTS

The proposed 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 approach. The masking thresholds T_k were computed from the estimated clean signal spectrum $\hat{S}_x = S_y^{\text{wt}} - \hat{S}_n$ using the approach outlined in [6].

The following parameter values were used in the noise spectrum estimation algorithm: $\alpha_s = 0.8$, $\alpha_d = 0.95$, $\delta = 5$, $\alpha_p = 0.2$, and the duration of the search window (P) for minimum tracking was set to 1 s. Five tapers ($L = 5$) were used in multitaper spectrum estimation. Level-dependent soft thresholding was used in the wavelet thresholding procedure as described in [8], [13] with the wavelet decomposition level q_0 set to 5.

For evaluation purposes, we used 60 sentences from the Hearing in Noise Test (HINT) database [14]. For nonstationary noise, we used multitalker babble (two male and two female talkers) added to the clean speech files at 0–20 dB SNR. The modified bark spectral distortion (MBSD) measure [15] 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) [16], which was found to be highly correlated with speech quality [15]. For comparative purposes, we also implemented and evaluated the MMSE-LSA method proposed by Ephraim and Malah [17].¹ In order to make a fair comparison, the same noise spectrum estimation method was used in the MMSE-LSA estimator. Finally, in order to assess the individual contribution of the spectrum estimation method, we implemented the proposed approach using the fast Fourier transform (FFT) magnitude spectra in place of the wavelet-thresholded multitaper spectra.

Fig. 1 presents the mean results in terms of the MBSD and segmental SNR measures for 60 HINT sentences corrupted by the multitalker babble noise at 0–20 dB SNR. As can be seen, the proposed approach outperformed the MMSE-LSA estimator in terms of the MBSD and segmental SNR measures. The benefit in using wavelet thresholded multitaper spectra in the computation of the gain function is also evident from Fig. 1, particularly at low SNR levels (0–5 dB). Informal listening tests confirmed that the proposed method obtained better quality with significantly lower noise distortion than the MMSE-LSA speech enhancement method.

V. SUMMARY AND CONCLUSION

An optimal frequency domain estimator was derived based on the masking properties of the human auditory system. The use of wavelet-thresholded multitaper spectrum estimators was shown to yield better performance in low SNR levels compared to the traditional FFT-based spectrum estimators. This advantage was attributed to the lower variance associated with multitaper spectrum estimation. Experiments with multitalker babble demonstrated improved performance, in terms of objective measures, over the MMSE-LSA speech enhancement method.

¹Note that several new algorithms were proposed recently (e.g., [18] and [19]) providing improvements to the MMSE-LSA estimator, however, those algorithms incorporated signal-presence uncertainty in their spectral estimator, and our proposed estimator does not. For that reason, we only compare the performance of our estimator against the performance of the MMSE-LSA estimator.

REFERENCES

- [1] 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.
- [2] 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.
- [3] S. Gustafsson, P. Jax, and P. Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1998, pp. 397–400.
- [4] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 2002, pp. 569–572.
- [5] R. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, 1972.
- [6] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
- [7] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE*, vol. 70, pp. 1055–1096, Sept. 1982.
- [8] 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, Dec. 1998.
- [9] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE Trans. Signal Processing*, vol. 43, pp. 188–195, Jan. 1995.
- [10] 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.
- [11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 504–512, July 2001.
- [12] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Lett.*, vol. 9, pp. 12–15, Jan. 2002.
- [13] 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.
- [14] 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.
- [15] 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. Acoustics, Speech, Signal Processing*, 1998, pp. 541–544.
- [16] 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, June 1992.
- [17] 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.
- [18] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Lett.*, vol. 9, pp. 113–116, Apr. 2002.
- [19] N. Kim and J. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Lett.*, vol. 7, pp. 108–110, May 2000.