A generalized subspace approach for enhancing speech corrupted by colored noise

Article *in* IEEE Transactions on Speech and Audio Processing · August 2003

DOI: 10.1109/TSA.2003.814458 · Source: IEEE Xplore

Yi Hu
University of Wisconsin - Milwaukee

A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise

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

***Abstract*—A generalized subspace approach is proposed for enhancement of speech corrupted by colored noise. A nonunitary transform, based on the simultaneous diagonalization of the clean speech and noise covariance matrices, is used to project the noisy signal onto a signal-plus-noise subspace and a noise subspace. The clean signal is estimated by nulling the signal components in the noise subspace and retaining the components in the signal subspace. The applied transform has built-in prewhitening and can therefore be used in general for colored noise. The proposed approach is shown to be a generalization of the approach proposed by Ephraim and Van Trees for white noise. Two estimators were derived based on the nonunitary transform, one based on time-domain constraints and one based on spectral domain constraints. Objective and subjective measures demonstrated improvements over other subspace-based methods when tested with TIMIT sentences corrupted with speech-shaped noise and multi-talker babble.**

***Index Terms*—Colored noise, KLT, noise reduction, speech enhancement, subspace-based method.**

I. Introduction

DUE TO its low complexity and high efficiency, the spectral subtraction algorithm is perhaps the most popular speech enhancement algorithm used today. The major drawback of spectral subtraction, however, is the incurrence of speech distortion, and particularly "musical noise." Although several methods (e.g., [1], [2]) were proposed to reduce speech distortion, it was done at the expense of introducing residual noise. Understanding that a compromise needs to be made between residual noise and speech distortion, several researchers proposed the use of subspace-based methods which attempt to minimize the speech distortion while keeping the residual noise below a preset threshold [3]–[6].

The idea behind subspace methods is to project the noisy signal onto two subspaces: the signal-plus-noise subspace, or simply signal subspace (since the signal dominates this subspace), and the noise subspace. The noise subspace contains signals from the noise process only, hence an estimate of the clean signal can be made by removing or nulling the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace. The decomposition of the space into two subspaces can be done using either the singular value decomposition (SVD) [7], [8] or the eigenvalue decomposition (EVD) [3]–[6].

The SVD-based method proposed by Dendrinos *et al.* [7], was based on the idea that the eigenvectors corresponding to the largest singular values contained signal information, while the eigenvectors corresponding to the smallest singular values contained noise information. The enhanced signal was reconstructed using only the information associated with the largest singular values. Large SNR gains were found for speech corrupted in white noise. Jensen *et al.* [8] extended the approach by Dendrinos *et al.* [7] to colored noise using the quotient SVD (QSVD). Rather than arranging the signal data in a Toeplitz matrix, they arranged the data in a Hankel matrix and computed the least squares estimate of the signal-only Hankel matrix. The QSVD was found to be computationally expensive and provided no method for shaping or controlling the residual noise.

A different formulation to subspace-based methods for speech enhancement was provided by Ephraim and Van Trees (EV) in [3]. EV seeked for an optimal estimator that would minimize the speech distortion subject to the constraint that the residual noise fell below a preset threshold. Using the eigenvalue decomposition of the covariance matrix, EV showed that the decomposition of the vector space of the noisy signal into a signal and noise subspace can be obtained by applying the Karhunen–Loéve transform (KLT) to the noisy signal. The KLT components representing the signal subspace were modified by a gain function determined by the estimator, while the remaining KLT components representing the noise subspace were nulled. The enhanced signal was obtained from the inverse KLT of the modified components. Subjective tests indicated that the subspace approach yielded better speech quality than the traditional spectral subtraction approach. EV's formulation of the subspace approach was based on the assumption that the input noise was white.

Mittal and Phamdo [4] and Rezayee and Gazor [5] later extended EV's work to colored noise. In [4], they focused on providing proper noise shaping for colored noise without prewhitening. To do that, they first classified the noisy speech frames into speech-dominated and noise-dominated frames and used a different KLT matrix for these frames to construct the estimator. Objective results showed that their approach provided better noise shaping than EV's approach. Rezayee and Gazor [5] extended EV's approach to deal with colored noise by approximating the covariance matrix of the KLT-transformed noise vectors with a diagonal matrix. This approximation led to a suboptimal estimator, however, subjective results showed that the majority of the listeners preferred the quality of speech

Manuscript received January 30, 2002; revised February 23, 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. Li Deng.

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.2003.814458

produced by their estimator over the quality of speech produced by EV's estimator.

In this paper, we derive a generalized subspace approach with built-in prewhitening for enhancing speech corrupted with colored noise. Unlike [5], we make no assumptions about the covariance matrix of the KLT-transformed noise vectors, hence our estimator is optimal. We show that the EV estimator is a special case of our estimator when the noise is white. Following EV, we derive two different estimators, one based on time-domain constraints and one based on spectrum domain constraints.

This paper is organized as follows. In Section II, the proposed approach using time-domain constraints is presented. Section II also proves that the spectral subtraction method by [1] can be reformulated as a constrained minimization problem. In Section III, the proposed approach using spectrum domain constraints is described. Implementation details and experimental results are given in Sections IV and V respectively, and the conclusions are given in Section VI.

II. Subspace Approach Based on Time Domain Constraints (TDC)

In this section, we first derive the linear optimal estimator which minimizes the speech distortion using time-domain constraints, and then show that the well-known power spectrum over-subtraction method proposed in [1] can also be reformulated in a similar way.

A. Principles

A linear signal model is assumed for the clean signal $\mathbf{x}$:

$$\mathbf{x} = \Psi \cdot \mathbf{s} \tag{1}$$

where Ψ is a $K \times M$ matrix whose rank is M ($M < K$), and $\mathbf{s}$ is an $M \times 1$ vector. The covariance matrix of $\mathbf{x}$ is given by

$$R_{\mathbf{x}} \triangleq E\{\mathbf{x} \cdot \mathbf{x}^T\} = \Psi \cdot R_{\mathbf{s}} \cdot \Psi^T \tag{2}$$

where $R_{\mathbf{s}}$ is the covariance matrix of the vector $\mathbf{s}$, assumed to be positive definite. The rank of $R_{\mathbf{x}}$ is M, hence it has $K - M$ zero eigenvalues.

Given the above signal model, and assuming that the noise signal is additive and uncorrelated with the speech signal, we can write the corrupted signal as:

$$\mathbf{y} = \Psi \cdot \mathbf{s} + \mathbf{n} = \mathbf{x} + \mathbf{n} \tag{3}$$

where $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$ are the K-dimensional noisy speech, clean speech and noise vectors respectively. Let $\hat{\mathbf{x}} = H \cdot \mathbf{y}$ be a linear estimator of the clean speech $\mathbf{x}$, where H is a $K \times K$ matrix. The error signal ε obtained by this estimation is given by

$$\varepsilon = \hat{\mathbf{x}} - \mathbf{x} = (H - I) \cdot \mathbf{x} + H \cdot \mathbf{n} = \varepsilon_{\mathbf{x}} + \varepsilon_{\mathbf{n}} \tag{4}$$

where $\varepsilon_{\mathbf{x}}$ represents the speech distortion and $\varepsilon_{\mathbf{n}}$ represents the residual noise [3]. Defining the energies of the signal distortion $\overline{\varepsilon_{\mathbf{x}}^2}$ and the energies of the residual noise $\overline{\varepsilon_{\mathbf{n}}^2}$ as

$$\overline{\varepsilon_{\mathbf{x}}^2} = E\left[\varepsilon_{\mathbf{x}}^T \varepsilon_{\mathbf{x}}\right] = \operatorname{tr}\left(E\left[\varepsilon_{\mathbf{x}} \varepsilon_{\mathbf{x}}^T\right]\right) \tag{5}$$

$$\overline{\varepsilon_{\mathbf{n}}^2} = E\left[\varepsilon_{\mathbf{n}}^T \varepsilon_{\mathbf{n}}\right] = \operatorname{tr}\left(E\left[\varepsilon_{\mathbf{n}} \varepsilon_{\mathbf{n}}^T\right]\right) \tag{6}$$

we can obtain the optimum linear estimator by solving the following time-domain constrained optimization problem [3], [5]

$$\begin{aligned} &\min_H \overline{\varepsilon_{\mathbf{x}}^2} \\ &\text{subject to: } \frac{1}{K}\overline{\varepsilon_{\mathbf{n}}^2} \le \sigma^2 \end{aligned} \tag{7}$$

where σ^2 is a positive constant. The solution to (7) is given by [3]

$$H_{opt} = R_{\mathbf{x}} \left(R_{\mathbf{x}} + \mu R_{\mathbf{n}}\right)^{-1} \tag{8}$$

where $R_{\mathbf{x}}$ and $R_{\mathbf{n}}$ are the covariance matrices of the clean speech and noise respectively, and μ is the Lagrange multiplier.

Equation (8) can be simplified using the eigen-decomposition of $R_{\mathbf{x}} = U \Delta_{\mathbf{x}} U^T$ to:

$$H_{opt} = U \Delta_{\mathbf{x}} \left(\Delta_{\mathbf{x}} + \mu U^T R_{\mathbf{n}} U\right)^{-1} U^T \tag{9}$$

where U is the (unitary) eigenvector matrix and $\Delta_{\mathbf{x}}$ is the diagonal eigenvalue matrix of $R_{\mathbf{x}}$. Note that for white noise with variance $\sigma_{\mathbf{n}}^2$, $R_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I$ and the above estimator reduces to the Ephraim and Van Trees' estimator [3]. In [5], the matrix $U^T R_{\mathbf{n}} U$ was approximated by the diagonal matrix $\Delta_{\mathbf{n}}$

$$\Delta_{\mathbf{n}} = \operatorname{diag}\Big(E\Big(\left|\mathbf{u}_1^T \mathbf{n}\right|^2\Big), E\Big(\left|\mathbf{u}_2^T \mathbf{n}\right|^2\Big), \ldots, E\Big(\left|\mathbf{u}_K^T \mathbf{n}\right|^2\Big)\Big) \tag{10}$$

where $\mathbf{u}_k$ is the kth eigenvector of $R_{\mathbf{x}}$, and $\mathbf{n}$ is the noise vector estimated from the speech-absent segments of speech. The above approximation yielded the following estimator [5]:

$$H_{opt} \approx U \Delta_{\mathbf{x}} (\Delta_{\mathbf{x}} + \mu \Delta_{\mathbf{n}})^{-1} U^T. \tag{11}$$

Because of the approximation in (10), the estimator used in [5] was suboptimal. Next, we present an optimal [in the sense of (7)] estimator suited for colored noise.

Computer simulations indicated that the matrix $U^T R_{\mathbf{n}} U$ in (9) was not diagonal, although in some cases it was nearly diagonal. This was not surprising, since U, being the eigenvector matrix of the symmetric matrix $R_{\mathbf{x}}$, diagonalizes $R_{\mathbf{x}}$ and not $R_{\mathbf{n}}$. Rather than trying to approximate $U^T R_{\mathbf{n}} U$, we looked for a matrix that would simultaneously diagonalize $R_{\mathbf{x}}$ and $R_{\mathbf{n}}$. It can be shown [9] that such a matrix exists and can simultaneously diagonalize the two matrices in the following way:

$$\begin{aligned} V^T R_{\mathbf{x}} V &= \Lambda_{\mathbf{x}} \\ V^T R_{\mathbf{n}} V &= I \end{aligned} \tag{12}$$

where $\Lambda_{\mathbf{x}}$ and V are the eigenvalue matrix and eigenvector matrix respectively of $\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{x}}$, i.e.,

$$\Sigma V = V \Lambda_{\mathbf{x}}. \tag{13}$$

It can be shown that $\Lambda_{\mathbf{x}}$ is a real matrix [10], assuming that $R_{\mathbf{n}}$ is positive definite. Note that the eigenvector matrix V is not orthogonal, and that the rank of the matrix Σ is M since rank $(R_{\mathbf{x}}) = M$. Applying the above eigen-decomposition of Σ to (8), and using (12), we can rewrite the optimal linear estimator as

$$\begin{aligned} H_{opt} &= R_{\mathbf{n}} V \Lambda_{\mathbf{x}} (\Lambda_{\mathbf{x}} + \mu I)^{-1} V^T \\ &= V^{-T} \Lambda_{\mathbf{x}} (\Lambda_{\mathbf{x}} + \mu I)^{-1} V^T. \end{aligned} \tag{14}$$

It can be shown that μ must satisfy

$$\sigma^2 = \frac{1}{K}\,\mathrm{tr}\left\{\left(V^T V\right)^{-1}\Lambda_{\mathbf{x}}^2(\Lambda_{\mathbf{x}}+\mu I)^{-2}\right\}. \tag{15}$$

The enhanced signal $\hat{\mathbf{x}}$ is obtained by applying the transform V^T to the noisy signal, appropriately modifying the components of $V^T\mathbf{y}$ by a gain function, and then taking the inverse transform (V^{-T}) of the modified components. The gain matrix $G = \Lambda_{\mathbf{x}}(\Lambda_{\mathbf{x}}+\mu I)^{-1}$ is diagonal and its kth diagonal element g_{kk} can be written as

$$g_{kk} = \begin{cases} \dfrac{\lambda_{\mathbf{x}}^{(k)}}{\lambda_{\mathbf{x}}^{(k)}+\mu}, & k = 1, 2, \ldots, M \\ 0, & k = M+1, \ldots, K \end{cases} \tag{16}$$

where $\lambda_{\mathbf{x}}^{(k)}$ is the kth diagonal element of the eigenvalue matrix $\Lambda_{\mathbf{x}}$ and M is the rank of the matrix Σ and the assumed dimension of the speech signal subspace. Note that in our case, $V^T\mathbf{y}$ is not the Karhunen–Loéve transform (KLT) of $\mathbf{y}$. However, as we show below, if the noise is white, $V^T\mathbf{y}$ becomes the KLT of $\mathbf{y}$.

Comparing the above estimator given in (14) with the corresponding linear estimator obtained for white noise in [3], we can see that both estimators have the same form. In fact, the Ephraim and Van Trees' estimator [3] is a special case of the proposed estimator in (14). For white noise $R_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I$, and V becomes the unitary eigenvector matrix (U) of $R_{\mathbf{x}}$, since $\Sigma = (1/\sigma_{\mathbf{n}}^2)R_{\mathbf{x}}$, and the diagonal matrix $\Lambda_{\mathbf{x}}$ becomes $(1/\sigma_{\mathbf{n}}^2)\Delta_{\mathbf{x}}$, where $\Delta_{\mathbf{x}}$ is the diagonal eigenvalue matrix of $R_{\mathbf{x}}$. Therefore, for white noise (14) reduces to

$$H_{opt} = U\Delta_{\mathbf{x}}\left(\Delta_{\mathbf{x}}+\mu\sigma_{\mathbf{n}}^2 I\right)^{-1}U^T \tag{17}$$

which is the Ephraim and Van Trees' estimator [3]. The proposed approach is therefore a generalization of the subspace approach developed in [3] and can be used for both white and colored noise. In fact, the proposed approach makes no assumptions about the spectral characteristics of the noise.

For the above proposed estimator, we need an estimate of the matrix Σ. Since we have no access to the covariance matrix of the clean speech signal, we can estimate Σ from the noisy speech signal as follows. Assuming that speech is uncorrelated with noise, we have

$$R_{\mathbf{y}} = R_{\mathbf{x}} + R_{\mathbf{n}} \tag{18}$$

and so

$$\Sigma = R_{\mathbf{n}}^{-1}R_{\mathbf{x}} = R_{\mathbf{n}}^{-1}(R_{\mathbf{y}} - R_{\mathbf{n}}) = R_{\mathbf{n}}^{-1}R_{\mathbf{y}} - I. \tag{19}$$

B. Estimating μ

The estimation of μ in the gain function (16) affects the quality of speech as it controls the tradeoff between residual noise and speech distortion. A large value of μ would eliminate much of the background noise at the expense of introducing speech distortion. Conversely, a small value of μ would minimize the speech distortion at the expense of introducing large residual noise. Hence, a compromise between residual noise and speech distortion needs to be made by an appropriate choice of μ.

Ideally, we would like to minimize the speech distortion in speech-dominated frames since the speech signal will have a masking effect on noise. 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 short-time SNR. We therefore chose the following equation for estimating μ

$$\mu = \mu_0 - (\mathrm{SNR}_{\mathrm{dB}})/s \tag{20}$$

where μ_0 and s are constants chosen experimentally, and $\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\mathrm{SNR}$. Note that a similar equation was used in [1] to estimate the over-subtraction factor in spectral subtraction. We believe that the proposed method which uses a variable μ obtains a better trade-off between speech distortion and residual noise than the approach in [3], [6] which used a fixed value of μ regardless of the frame SNR.

Since we know from (12) that the eigenvalue $\lambda_{\mathbf{x}}^{(k)}$ is equal to the signal energy along the corresponding eigenvector v_k [i.e., $\lambda_{\mathbf{x}}^{(k)} = E(|v_k^T\mathbf{x}|^2)$], we can derive the estimate of the SNR value directly in the transform domain using the following equation:

$$\mathrm{SNR} = \frac{\mathrm{tr}\left(V^T R_{\mathbf{x}} V\right)}{\mathrm{tr}\left(V^T R_{\mathbf{n}} V\right)} = \frac{\sum_{k=1}^{M}\lambda_{\mathbf{x}}^{(k)}}{K}. \tag{21}$$

Note that the above SNR definition reduces to the traditional SNR definition of $\mathrm{tr}(R_{\mathbf{x}})/\mathrm{tr}(R_{\mathbf{n}})$ for an orthonormal matrix V.

C. Reformulation of the Power Spectrum Subtraction Method

To reduce the musical noise incurred by the spectral subtraction algorithm, Berouti *et al.* [1] proposed the use of an over-subtraction factor and the use of a spectral floor. The basic idea is to subtract more from the noisy speech spectrum when the noise dominates the current frame, i.e., in low-SNR frames, and subtract less in speech-dominant frames, i.e., in high-SNR frames. Next, we prove that this power spectrum subtraction approach can be formulated as the solution to a constrained optimization problem that minimizes the energy of the speech distortion in the frequency domain while maintaining the energy of the residual noise below a preset threshold.

We denote the N-point discrete Fourier transform matrix by F

$$F = \frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{j\omega_0} & \cdots & e^{j(N-1)\omega_0} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{j(N-1)\omega_0} & \cdots & e^{j(N-1)(N-1)\omega_0} \end{bmatrix} \tag{22}$$

where $\omega_0 = 2\pi/N$. The Fourier transform of the noisy speech vector $\mathbf{y}$ can then be written as

$$\mathbf{Y}(\omega) = F^H\cdot\mathbf{y} = F^H\cdot\mathbf{x} + F^H\cdot\mathbf{n} = \mathbf{X}(\omega) + \mathbf{N}(\omega) \tag{23}$$

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

$$\begin{aligned}\varepsilon(\omega) &= \hat{\mathbf{X}}(\omega) - \mathbf{X}(\omega)\\ &= (G - I) \cdot \mathbf{X}(\omega) + G \cdot \mathbf{N}(\omega)\\ &= \varepsilon_{\mathbf{x}}(\omega) + \varepsilon_{\mathbf{n}}(\omega)\end{aligned} \tag{24}$$

where $\varepsilon_{\mathbf{x}}(\omega) = (G - I) \cdot \mathbf{X}(\omega)$ represents the speech distortion in the frequency domain and $\varepsilon_{\mathbf{n}}(\omega) = G \cdot \mathbf{N}(\omega)$ represents the residual noise in the frequency domain. After defining the energy of the frequency domain speech distortion, $\overline{\varepsilon_{\mathbf{x}}^2(\omega)}$, and the energy of the frequency domain residual noise, $\overline{\varepsilon_{\mathbf{n}}^2(\omega)}$, as

$$\begin{aligned}\overline{\varepsilon_{\mathbf{x}}^2(\omega)} &= E\left(\varepsilon_{\mathbf{x}}^H(\omega) \cdot \varepsilon_{\mathbf{x}}(\omega)\right) = \mathrm{tr}\big(E\left(\varepsilon_{\mathbf{x}}(\omega) \cdot \varepsilon_{\mathbf{x}}^H(\omega)\right)\big)\\ &= \mathrm{tr}\big((G - I) \cdot F^H \cdot R_{\mathbf{x}} \cdot F \cdot (G - I)^H\big)\end{aligned} \tag{25}$$

$$\begin{aligned}\overline{\varepsilon_{\mathbf{n}}^2(\omega)} &= E\left(\varepsilon_{\mathbf{n}}^H(\omega) \cdot \varepsilon_{\mathbf{n}}(\omega)\right) = \mathrm{tr}\big(E\left(\varepsilon_{\mathbf{n}}(\omega) \cdot \varepsilon_{\mathbf{n}}^H(\omega)\right)\big)\\ &= \mathrm{tr}\big(G \cdot F^H \cdot R_{\mathbf{n}} \cdot F \cdot G^H\big)\end{aligned} \tag{26}$$

we can obtain the optimal linear estimator by solving the following constrained optimization problem:

$$\begin{aligned}&\min_G \overline{\varepsilon_{\mathbf{x}}^2(\omega)}\\ &\text{subject to: } \frac{1}{N}\overline{\varepsilon_{\mathbf{n}}^2(\omega)} \le \sigma^2\end{aligned} \tag{27}$$

where σ^2 is a positive number. The estimator derived this way minimizes the energy of the frequency domain speech distortion while maintaining the energy of the frequency domain residual noise below the preset threshold σ^2.

Using the Lagrange method to solve the above problem, it can be shown that G satisfies the following equation:

$$G\left(F^H \cdot R_{\mathbf{x}} \cdot F + \mu \cdot F^H \cdot R_{\mathbf{n}} \cdot F\right) = F^H \cdot R_{\mathbf{x}} \cdot F. \tag{28}$$

Note that in the Berouti *et al.* [1] method, the gain is applied to each frequency component individually, hence G should be a diagonal matrix. The matrices $F^H \cdot R_{\mathbf{x}} \cdot F$ and $F^H \cdot R_{\mathbf{n}} \cdot F$ are asymptotically diagonal [11] (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 $P_{\mathbf{x}}(\omega_i)$ and $P_{\mathbf{n}}(\omega_i)$ of the clean speech vector $\mathbf{x}$ and noise vector $\mathbf{n}$, respectively. Denoting the diagonal elements of G by $g(\omega_i)$, (28) can be written as

$$g(\omega_i) \cdot (P_{\mathbf{x}}(\omega_i) + \mu \cdot P_{\mathbf{n}}(\omega_i)) = P_{\mathbf{x}}(\omega_i). \tag{29}$$

The gain function $g(\omega_i)$ for the frequency component ω_i is therefore given by

$$g(\omega_i) = \frac{P_{\mathbf{x}}(\omega_i)}{P_{\mathbf{x}}(\omega_i) + \mu \cdot P_{\mathbf{n}}(\omega_i)}. \tag{30}$$

In [1], the gain function $g'(\omega_i)$ for the frequency component ω_i was given by

$$g'(\omega_i) = \frac{D(\omega_i)}{P_s(\omega_i)} = \frac{D(\omega_i)}{D(\omega_i) + \alpha \cdot P_{\mathbf{n}}(\omega_i)} \tag{31}$$

where $P_s(\omega_i)$ is the power spectrum of the noisy signal, and $D(\omega_i) = P_s(\omega_i) - \alpha \cdot P_{\mathbf{n}}(\omega_i)$ is the estimate of the power spectrum of the clean signal. As can be seen, the gain functions in (30) and (31) have the same form. It is interesting to note that the over-subtraction factor α in [1] plays the same role as the Lagrangian multiplier μ in this paper, in that it controls the trade-off between speech distortion and residual noise.

III. Subspace Approach Based on Spectrum Domain Constraints (SDC)

The TDC estimator proposed in the previous section did not provide for any noise spectral shaping for possible masking by the speech signal. In this section, we derive a linear estimator based on spectral constraints. The idea behind the spectrum domain constrained (SDC) linear optimal estimator is to minimize the signal distortion subject to constraints on the shape of the spectrum of the residual noise [3].

Specifically, suppose that the kth spectral component of the residual noise is given by $v_k^T \varepsilon_{\mathbf{n}}$, where v_k is the kth column vector of the eigenvector matrix of $\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{x}}$. For $k = 1, \ldots, M$, we require the energy in $v_k^T \varepsilon_{\mathbf{n}}$ to be smaller than or equal to α_k ($0 < \alpha_k < 1$), while for $k = M + 1, \ldots, K$, we require the energy in $v_k^T \varepsilon_{\mathbf{n}}$ to be zero, since the signal energy in the noise subspace is zero [3]. Therefore, the filter H is designed by

$$\begin{aligned}&\min_H \overline{\varepsilon_{\mathbf{x}}^2}\\ &\text{subject to: } \begin{cases} E\left\{\left|v_k^T \varepsilon_{\mathbf{n}}\right|^2\right\} \le \alpha_k, & k = 1, 2, \ldots, M\\ E\left\{\left|v_k^T \varepsilon_{\mathbf{n}}\right|^2\right\} = 0, & k = M + 1, \ldots, K.\end{cases}\end{aligned} \tag{32}$$

The optimal estimator, in the sense of (32), can be found using the method of Lagrange multipliers. It can be shown that the optimal H must satisfy the following matrix equation:

$$HR_{\mathbf{x}} + LHR_{\mathbf{n}} - R_{\mathbf{x}} = 0 \tag{33}$$

where $L = V\Lambda_\mu V^T$, and $\Lambda_\mu = \mathrm{diag}(\mu_1, \ldots, \mu_K)$ is a diagonal matrix of Lagrangian multipliers. Using (12), we can rewrite (33) as

$$V^T H V^{-T} \Lambda_{\mathbf{x}} + V^T V \Lambda_\mu V^T H V^{-T} - \Lambda_{\mathbf{x}} = 0 \tag{34}$$

which can be further reduced to the following equation:

$$Q\Lambda_{\mathbf{x}} + V^T V \Lambda_\mu Q = \Lambda_{\mathbf{x}} \tag{35}$$

where $Q = V^T H V^{-T}$. Note that for white noise, (35) reduces to the same equation given in [3, p. 255] for the spectral domain estimator.

The above equation is the well known Lyaponov equation encountered frequently in control theory. The Lyapunov equation can be solved numerically using the algorithm proposed in [12]. Explicit solutions can be found in [13, p. 414]. After solving for Q in (35), we can compute the optimal H by

$$H_{opt} = V^{-T} Q V^T. \tag{36}$$

Depending on the assumptions made about the structure of the matrix Q, we derived two different estimators. In the first method, which we refer to as spectral domain constrain method 1 (SDC1), we make no assumptions about the structure of the matrix Q. In that method, the Lyapunov equation (35) is solved in each frame for Q, and from that the optimal estimator is solved according to (36).

In the second method, which we refer to as spectral domain constrain method 2 (SDC2), we assume that the matrix Q is diagonal. We further assume that the matrix $V^T V$ is nearly diagonal. Making those two assumptions simplifies the solution of the Lyapunov equation (35) a great deal. Let $\lambda_{\mathbf{x}}^{(k)}$ be the kth diagonal element of the matrix $\Lambda_{\mathbf{x}}$ and q_{kk} be the kth diagonal element of Q, (35) can then be rewritten as

$$\left(\lambda_{\mathbf{x}}^{(k)} + \|v_k\|^2 \cdot \mu_k\right) \cdot q_{kk} = \lambda_{\mathbf{x}}^{(k)} \tag{37}$$

and q_{kk} can be solved as

$$q_{kk} = \frac{\lambda_{\mathbf{x}}^{(k)}}{\lambda_{\mathbf{x}}^{(k)} + \|v_k\|^2 \cdot \mu_k} \qquad k = 1, 2, \ldots, K. \tag{38}$$

Without loss of generality, we can make the norm of v_k equal to 1 and rewrite the above solution as

$$q_{kk} = \begin{cases} \dfrac{\lambda_{\mathbf{x}}^{(k)}}{\lambda_{\mathbf{x}}^{(k)} + \mu_k}, & k = 1, 2, \ldots, M \\ 0, & k = M+1, \ldots, K. \end{cases} \tag{39}$$

Comparing (39) and (16) we can see that (39) can be interpreted as a multiband version of (16) in that it uses a different value of μ for each spectral component. Note that (38) is similar to the corresponding equation in [3], with the $\|v_k\|^2$ in place of the noise variance σ_ω^2.

Using the above Q and the assumption that $V^T V$ is nearly diagonal, we have

$$E\left\{\left|v_k^T \boldsymbol{\varepsilon}_{\mathbf{n}}\right|^2\right\} = \begin{cases} q_{kk}^2, & k = 1, \ldots, M \\ 0, & k = M+1, \ldots, K. \end{cases} \tag{40}$$

If the nonzero constraints in (32) are satisfied with equality, then $q_{kk}^2 = \alpha_k$, suggesting that

$$q_{kk} = (\alpha_k)^{1/2}, \qquad k = 1, \ldots, M \tag{41}$$

and

$$\mu_k = \begin{cases} \lambda_{\mathbf{x}}^{(k)} \left[(1/\alpha_k)^{1/2} - 1\right], & k = 1, \ldots, M \\ 0, & k = M+1, \ldots, K. \end{cases} \tag{42}$$

Since $\mu_k \geq 0$, the Kuhn–Tucker necessary conditions [14] for the constrained minimization problem are satisfied by the solution in (39). From (36) and (39), we conclude that the desired H is given by

$$\begin{aligned} q_{kk} &= \begin{cases} \dfrac{\lambda_{\mathbf{x}}^{(k)}}{\lambda_{\mathbf{x}}^{(k)} + \mu_k}, & k = 1, 2, \ldots, M \\ 0, & k = M+1, \ldots, K \end{cases} \\ Q &= \operatorname{diag}(q_{11}, q_{22}, \ldots, q_{KK}) \\ H &= V^{-T} Q V^T. \end{aligned} \tag{43}$$

In the previous section, we pointed out that the value of μ controls the trade-off between the speech distortion and the residual noise, and showed that a better method to select the value of μ is to let it vary according to the segmental SNR. Since (39) can be interpreted as a multiband version of (16), it seems reasonable then to let μ_k vary with the estimated SNR value of each spectral component, i.e.,

$$\mu_k = \mu_0 - (\mathrm{SNR}_k)/s \tag{44}$$

where SNR_k is the SNR value of the kth spectral component. The value of SNR_k is simply $\lambda_{\mathbf{x}}^{(k)}$, since in the transform domain the noise energy along the corresponding eigenvector v_k is equal to one [see (12)]. The μ_k values are then computed as

$$\mu_k = \mu_0 - \lambda_{\mathbf{x}}^{(k)}/s \tag{45}$$

where μ_0 and s are experimentally chosen.

IV. Implementation

In this section, we give the implementation details of the three methods developed in this paper.

A. Implementation of the TDC Estimator

The proposed TDC approach can be formulated in the following six steps. For each speech frame:

Step 1: Compute the covariance matrix $R_{\mathbf{y}}$ of the noisy signal, and estimate the matrix $\Sigma = R_{\mathbf{n}}^{-1} R_y - I$ using (19). The noise covariance matrix $R_{\mathbf{n}}$ is computed using noise samples collected during speech-absent frames.

Step 2: Perform the eigen-decomposition of Σ

$$\Sigma V = V \Lambda_{\mathbf{x}}. \tag{46}$$

Step 3: Assuming that the eigenvalues of Σ are ordered as $\lambda_{\mathbf{x}}^{(1)} \geq \lambda_{\mathbf{x}}^{(2)} \geq \cdots \geq \lambda_{\mathbf{x}}^{(K)}$, estimate the dimension of the speech signal subspace as follows:

$$M = \arg \max_{1 \leq k \leq K} \left\{\lambda_{\mathbf{x}}^{(k)} > 0\right\}. \tag{47}$$

Step 4: Compute the μ value according to

$$\mu = \begin{cases} \mu_0 - (\mathrm{SNR}_{\mathrm{dB}})/s, & -5 < \mathrm{SNR}_{\mathrm{dB}} < 20 \\ 1 & \mathrm{SNR}_{\mathrm{dB}} \geq 20 \\ 5 & \mathrm{SNR}_{\mathrm{dB}} \leq -5 \end{cases} \tag{48}$$

where $\mu_0 = 4.2$, $s = 6.25$, $\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \mathrm{SNR}$ and SNR is computed as per (21).

Step 5: The optimal linear estimator is computed as follows:

$$\begin{aligned} g_{kk} &= \begin{cases} \dfrac{\lambda_{\mathbf{x}}^{(k)}}{\lambda_{\mathbf{x}}^{(k)} + \mu}, & k = 1, 2, \ldots, M \\ 0, & k = M+1, \ldots, K \end{cases} \\ G_1 &= \operatorname{diag}\{g_{11}, \ldots, g_{MM}\} \\ H_{opt} &= R_{\mathbf{n}} V \begin{bmatrix} G_1 & 0 \\ 0 & 0 \end{bmatrix} V^T \\ &= V^{-T} \begin{bmatrix} G_1 & 0 \\ 0 & 0 \end{bmatrix} V^T. \end{aligned} \tag{49}$$

Step 6: Estimate the enhanced speech signal by $\hat{\mathbf{x}} = H_{opt} \cdot \mathbf{y}$.

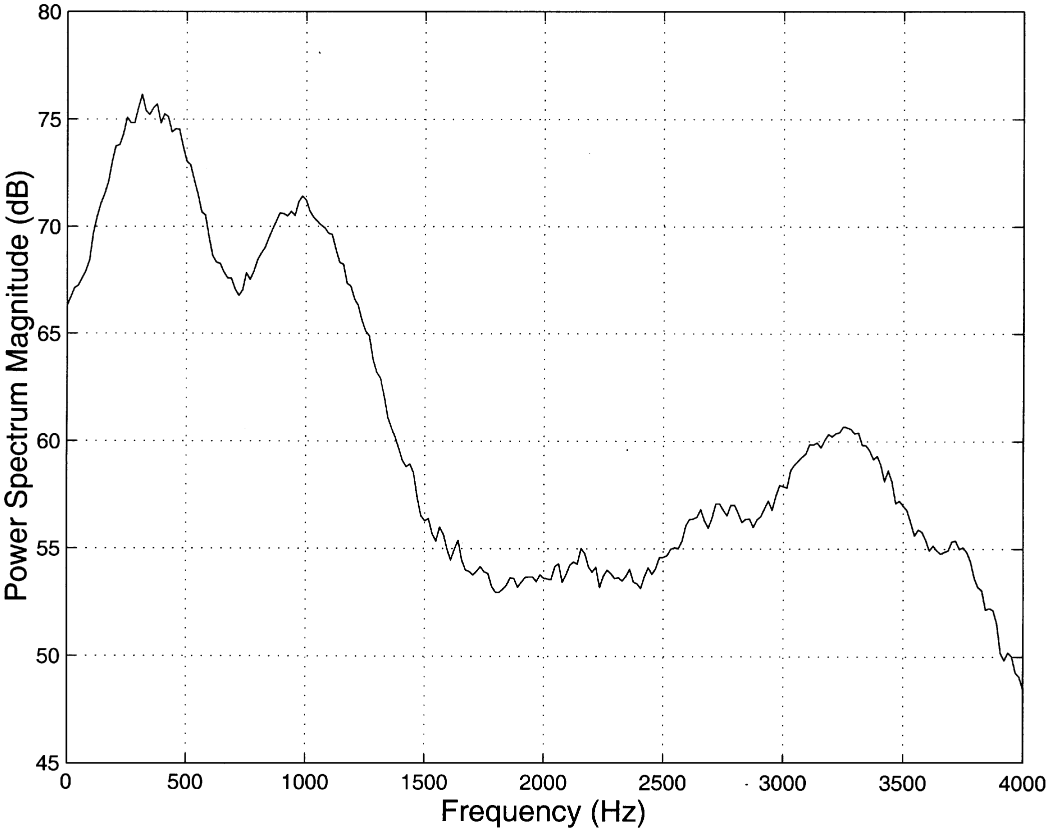


Fig. 1. Spectrum of speech-shaped noise used in this study.

B. Implementation of the SDC1 Estimator

The proposed SDC1 approach can be formulated in six steps, the first three of which are the same as the TDC approach.

Step 4: Initialize the diagonal matrix Λ_μ in (35) with μ_k $(1 \leq k \leq K)$ given by

$$\mu_k = \begin{cases} \mu_0 - \lambda'_k/s, & -5 < \lambda'_k < 20 \\ 1 & \lambda'_k \geq 20 \\ 20 & \lambda'_k \leq -5 \end{cases} \qquad (50)$$

where $\mu_0 = 16.2$, $s = 1.32$, and $\lambda'_k = 10\log_{10}\lambda_{\mathbf{x}}^{(k)}$.

Step 5: Solve for the Q matrix in the Lyapunov equation (35), and from that obtain H_{opt} as

$$H_{opt} = V^{-T}QV^T.$$

The numerical techniques given in [12] can be used to solve the Lyapunov equation.

Step 6: Estimate the enhanced speech signal by: $\hat{\mathbf{x}} = H_{opt} \cdot \mathbf{y}$.

C. Implementation of the SDC2 Estimator

The proposed SDC2 approach can also be formulated in six steps, the first three of which are the same as the TDC approach.

Step 4: Compute μ_k for each component according to (50), except that the maximum value of μ_k was set to 5 rather than 20, and $\mu_0 = 4.2$, $s = 6.25$.

Step 5: The optimal linear estimator is computed using (43).

Step 6: Estimate the enhanced speech signal by: $\hat{\mathbf{x}} = H_{opt} \cdot \mathbf{y}$.

D. Implementation Details

The proposed estimators require accurate estimates of the corrupted and noise covariance matrices, $R_{\mathbf{y}}$ and $R_{\mathbf{n}}$ [3]. In our study, we constructed a Toeplitz covariance matrix from the K samples of the unbiased autocorrelation sequence. Unlike [3], we did not use future or past frames to estimate the covariance matrix. In fact, we found that the objective performance of the estimators that used future and past frames to estimate the covariance matrix was not as good as when a single frame of speech was used, although the speech quality was comparable. To keep the computational complexity to a minimum, we chose $K = 40$ samples for speech sampled at 8 kHz. The noise covariance matrix $R_{\mathbf{n}}$ was estimated using the first few silence frames in the test sentences. No voice-activity detector (VAD) was used to update the noise covariance matrix $R_{\mathbf{n}}$.

The estimators were applied to frames of the noisy signal which overlapped each other by 50%. Rectangular windows were used to estimate the covariance matrices. The enhanced speech signal was Hamming windowed and combined using the overlap and add approach [15].

V. Results

For evaluation purposes, we used 20 sentences from the TIMIT database. The sentences were produced by 10 male and 10 female speakers. For colored noise, we used speech-shaped noise and multi-talker babble (seven talkers) added to the clean speech file at SNR $= 5$ dB. The speech-shaped noise, included in the HINT database [16], was computed by filtering white noise through an FIR filter with frequency response that matched the long-term spectrum of the sentences in the HINT database. The spectrum of the speech-shaped noise is shown in Fig. 1.

TABLE I
COMPARATIVE PERFORMANCE FOR SPEECH-SHAPED NOISE AT 5 dB, IN TERMS OF MEAN ITAKURA–SAITO DISTANCE MEASURE, FOR 20 TIMIT SENTENCES PRODUCED BY TEN MALE SPEAKERS AND TEN FEMALE SPEAKERS

	Male Speakers	**Female Speakers**
Noisy Speech	2.00	2.17
TDC Approach	1.26	1.36
SDC1 Approach	1.70	1.83
SDC2 Approach	1.28	1.38
Approach in [5]	1.34	1.47

TABLE II
COMPARATIVE PERFORMANCE FOR SPEECH-SHAPED NOISE AT 5 dB, IN TERMS OF MEAN OVERALL AND SEGMENTAL SNR VALUES (SNR/SNRSEG), FOR 20 TIMIT SENTENCES PRODUCED BY TEN MALE AND TEN FEMALE SPEAKERS

	Male Speakers	**Female Speakers**
Noisy Speech	5.0 / -2.20	5.0 / -2.10
TDC Approach	8.66 / 0.75	8.55 / 0.33
SDC1 Approach	-0.65 / -4.91	-0.69 / -5.04
SDC2 Approach	8.51 / 0.57	8.39 / 0.14
Approach in [5]	0.33 / -3.06	0.22 / -3.55

TABLE III
COMPARATIVE PERFORMANCE FOR MULTI-TALKER BABBLE AT 5 dB, IN TERMS OF MEAN ITAKURA–SAITO DISTANCE MEASURE, FOR 20 TIMIT SENTENCES PRODUCED BY TEN MALE AND TEN FEMALE SPEAKERS

	Male Speakers	**Female Speakers**
Noisy Speech	1.79	1.93
TDC Approach	1.80	2.11
SDC1 Approach	1.77	1.89
SDC2 Approach	1.80	2.07
Approach in [5]	1.49	1.63

TABLE IV
COMPARATIVE PERFORMANCE FOR MULTI-TALKER BABBLE AT 5 dB, IN TERMS OF MEAN OVERALL AND SEGMENTAL SNR VALUES (SNR/SNRSEG), FOR 20 TIMIT SENTENCES PRODUCED BY TEN MALE AND TEN FEMALE SPEAKERS

	Male Speakers	**Female Speakers**
Noisy Speech	5.0 / -1.99	5.0 / -1.96
TDC Approach	7.71 / -0.06	7.53 / -0.44
SDC1 Approach	-1.41 / -5.52	–1.39 / -5.57
SDC2 Approach	7.50 / -0.18	7.34 / -0.57
Approach in [5]	-0.43 / -4.12	-0.47 / -4.40

The Itakura–Saito (IS) distance measure, the overall (global) SNR and the segmental SNR [15] measures were adopted for evaluation of the proposed algorithms. For the IS distance measure, the largest 5% of the IS distance values were discarded, as suggested in [17], to exclude unrealistically large spectral distance values. For the segmental SNR, only frames with segmental SNR values greater than -10 dB and less than 35 dB were considered. For comparative purposes, we also implemented and evaluated a version of the approach in [5].

Tables I and II give the mean results in terms of the three objective measures for 20 TIMIT sentences corrupted by speech-shaped noise at 5 dB. The results are given separately for male and female speakers. As can be seen from Tables I and II, our proposed approach (TDC and SDC2) outperformed Rezayee and Gazor's approach [5] for both male and female speakers. Particularly large improvements were noted for the two SNR measures. Informal listening tests confirmed that the proposed subspace method preserved more speech information and had less speech distortion than the approach in [5]. The performance obtained with the SDC1 method, was not as good as the performance obtained with either the TDC or SDC2 methods. This was surprising given that no assumptions were made about the Q matrix used in the SDC1 approach. We suspect that the lower performance can be attributed to the choice of the diagonal matrix Λ_μ, i.e., the choice of the μ_k values.

Tables III and IV give the mean results in terms of the three objective measures for 20 TIMIT sentences corrupted by multi-talker babble at 5 dB. In terms of the SNR measures, our approach outperformed the adaptive KLT approach in [5]. In terms of the mean IS measures, the approach in [5], performed better than the proposed methods. We suspect that this might be due to the fact that the method in [5] was inherently adaptive and updated the noise statistics correspondingly. No VAD algorithm was used in our approach to update the noise covariance matrix. Hence, we expect further improvements in performance if we use a reliable VAD algorithm to update the noise covariance matrix $R_{\mathbf{n}}$.

Listening tests revealed that the speech quality of the TDC and SDC2 methods was sensitive to the choice of μ values, and in particular the maximum μ value allowed in (48) and (50). For evaluation purposes, we used $\mu = 5$ here as the maximum value for μ. Better speech quality with reduced residual noise can be achieved using larger values of μ ($\mu > 20$) as the maximum value. The SDC1 method, on the other hand, was not as sensitive to the choice of the μ values. Extremely high values of μ (>100) were needed before noticing any change in speech quality.

VI. CONCLUSIONS

A generalized subspace approach for enhancing speech degraded by colored additive noise was proposed and linear estimators using time domain and spectrum domain constraints were developed. The proposed approach is based on the simultaneous diagonalization of the covariance matrices of the speech signal and the noise signal. For white noise, the proposed approach reduces to the estimators derived by Ephraim and Van Trees [3] for enhancing speech corrupted by white noise. Our proposed approach makes no assumptions about the spectral characteristics of the noise. We showed that the spectral subtraction method proposed by [1] could also be formulated in a similar way as the solution to a constrained minimization problem.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments.

References

[1] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 208–211.

[2] P. Lockwood and J. Boudy, “Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars,” *Speech Commun.*, vol. 11, no. 2, pp. 215–228, 1992.

[3] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, 1995.

[4] U. Mittal and N. Phamdo, “Signal/noise KLT based approach for enhancing speech degraded by colored noise,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 159–167, Mar. 2000.

[5] A. Rezayee and S. Gazor, “An adaptive KLT approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.

[6] Y. Hu and P. C. Loizou, “A subspace approach for enhancing speech corrupted by colored noise,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Orlando, FL, May 2002, pp. 573–576.

[7] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: A regenerative approach,” *Speech Commun.*, vol. 10, pp. 45–57, 1991.

[8] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, “Reduction of broad-band noise in speech by truncated QSVD,” *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 439–448, Nov. 1995.

[9] S. B. Searle, *Matrix Algebra Useful for Statistics*. New York: Wiley, 1982.

[10] G. Strang, *Linear Algebra and Its Applications*, 3rd ed. New York: Harcourt Brace Jovanonich, 1988.

[11] R. Gray, “On the asymptotic eigenvalue distribution of Toeplitz matrices,” *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 725–730, 1972.

[12] R. H. Bartels and G. W. Stewart, “Solution of the matrix equation $AX + XB = C$,” *Commun. ACM*, vol. 15, no. 9, pp. 820–822, 1972.

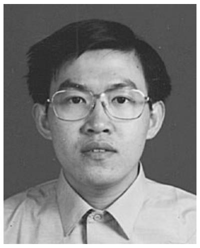
[13] P. Lancaster and M. Tismentetsky, *The Theory of Matrices*. New York: Academic, 1985.

[14] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley, 1984.

[15] J. R. Deller, J. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: IEEE Press, 2000.

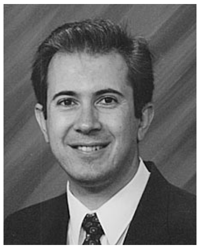
[16] M. Nilsson, S. Soli, and J. Sullivan, “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise,” *J. Acoust. Soc. Amer.*, vol. 95, pp. 1085–1099, 1994.

[17] J. Hansen and B. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *Int. Conf. Spoken Language Processing*, Sydney, Australia, Dec. 1998, pp. 2819–2822.



Yi Hu (S’01) received his B.S. and M.S. degrees in electrical engineering from University of Science and Technology of China (USTC) 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 and ASIC/FPGA design of DSP 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 in 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).