A Signal Subspace Approach for Speech Enhancement

Yariv Ephraim, *Fellow, IEEE,* and Harry L. Van Trees, *Life Fellow, IEEE*

***Abstract*— A comprehensive approach for nonparametric speech enhancement is developed. The underlying principle is to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace. Enhancement is performed by removing the noise subspace and estimating the clean signal from the remaining signal subspace. The decomposition can theoretically be performed by applying the Karhunen–Loève transform (KLT) to the noisy signal. Linear estimation of the clean signal is performed using two perceptually meaningful estimation criteria. First, signal distortion is minimized while the residual noise energy is maintained below some given threshold. This criterion results in a Wiener filter with adjustable input noise level. Second, signal distortion is minimized for a fixed spectrum of the residual noise. This criterion enables masking of the residual noise by the speech signal. It results in a filter whose structure is similar to that obtained in the first case, except that now the gain function which modifies the KLT coefficients is solely dependent on the desired spectrum of the residual noise. The popular spectral subtraction speech enhancement approach is shown to be a particular case of the proposed approach. It is proven to be a signal subspace approach which is optimal in an *asymptotic* (large sample) linear minimum mean square error sense, assuming the signal and noise are *stationary*. Our listening tests indicate that 14 out of 16 listeners strongly preferred the proposed approach over the spectral subtraction approach.**

I. Introduction

The GOAL of speech enhancement is to improve the performance of speech communication systems in noisy environments. The problem has been discussed in detail in [1] and in a series of tutorial papers [2]–[6]. We focus here on improving perceptual aspects of a given sample function of the noisy speech. Since, so far, no enhancement system was capable of improving both the quality and intelligibility of the noisy signal, we attempt to improve the quality of the noisy signal while minimizing any loss in its intelligibility.

We focus on linear estimation of the clean signal from the noisy signal. Estimation is performed on a frame-by-frame basis under the assumption that the noise is additive and uncorrelated with the clean signal. To motivate our estimation criterion, consider the residual signal, which is defined as the difference between the enhanced and the clean signals. For the given situation, it represents signal distortion and residual noise. Since both can not be simultaneously minimized, our estimation criterion will be to control the more perceptually harmful component of the residual signal while minimizing the other. Since speech signals are not strictly stationary, a time varying estimator must be used. Such an estimator provides nonstationary residual noise that is intolerable by the human auditory system. Hence, we have chosen to keep the level of the residual noise below some threshold while minimizing the signal distortion. This criterion is consistent with our goal stated above. The optimal linear estimator in this sense is shown to be a Wiener filter with adjustable input noise level.

A second, more perceptually meaningful, criterion is to minimize the signal distortion while keeping the energy of the residual noise in *each spectral component* below some given threshold. This strategy allows shaping of the spectrum of the residual noise so that its perception can be minimized. This can be accomplished by making the spectrum of the residual noise similar to that of the speech signal in the given frame. Thus, more noise is permitted to accompany high energy spectral components of the clean signal and vice-versa. This approach has been commonly used in minimizing the perceptual effects of quantization noise in speech coding systems [7]. It is shown that the optimal filter in this sense has a structure similar to that obtained using the first optimization criterion.

The two linear estimators depend on the covariance matrices of the signal and noise in each frame. The covariance matrix of the noise in each frame is assumed known and positive definite. It is usually estimated from adjacent frames of the noisy signal during which no speech activity is detected. The covariance matrices of vectors of the clean signal, however, need not be positive definite. Indeed, there is strong empirical evidence which indicates that the covariance matrices of most speech vectors have some eigenvalues which are practically zero. This can be observed by either examining the empirical covariance matrices of the speech signal or by studying the commonly used linear model for that signal [8]–[10]. This model assumes that each vector of speech is composed of a linear combination of some basis vectors, such as vectors of damped sinusoids. If the number of basis vectors is smaller than the dimension of the vector, as is usually the case, then the covariance matrix of that vector must have zero eigenvalues.

The fact that some of the eigenvalues of the clean signal may be zero indicates that the energy of the clean signal vector is distributed among a subset of its coordinates, and the signal is confined to a subspace of the noisy Euclidean space. Since all noise eigenvalues are strictly positive, the noise fills in

Manuscript received October 2, 1992; revised November 27, 1994. The associate editor coordinating the review of this paper and approving it for publication was Prof. John H. L. Hansen.

Y. Ephraim was with AT&T Bell Laboratories, Murray Hill, NJ 07974 USA. He is now with the Department of Electrical and Computer Engineering and the Center of Excellence in Command, Control, Communications, and Intelligence, George Mason University, Fairfax, VA 22030 USA.

H. L. Van Trees is with the Department of Electrical and Computer Engineering and the Center of Excellence in Command, Control, Communications, and Intelligence, George Mason University, Fairfax, VA 22030 USA.

IEEE Log Number 9411832.

the entire vector space of the noisy signal. Hence, the vector space of the noisy signal is composed of a signal-plus-noise subspace and a complementary noise subspace. The signal-plus-noise subspace, or simply the signal subspace, comprises vectors of the clean signal as well as of the noise process. The noise subspace contains vectors of the noise process only. Hence, in estimating the clean signal, the noise subspace can be removed and estimation can be performed from the signal subspace only.

The decomposition of the vector space of the noisy signal into a signal subspace and a noise subspace is performed by applying the Karhunen–Loève transform (KLT) to the noisy signal [11]–[13]. The linear estimation is performed by modifying the KLT components which represent the signal subspace by a gain function determined by the estimation criterion. The remaining KLT components are nulled. The enhanced signal is obtained from inverse KLT of the altered components.

It is interesting to note that a similar estimation procedure has been commonly used by the popular spectral subtraction speech enhancement approach [1]–[6], [14]–[20]. In this approach, the discrete Fourier transform (DFT) rather than the KLT is used. The DFT components which represent the noise subspace are characterized by nonpositive signal spectral variance estimates. Since the KLT and the DFT are related [21], the spectral subtraction must be an approximate signal subspace approach. Furthermore, the approach must be approximately optimal in a sense determined by the specific gain function used to modify the DFT components. Indeed, it is shown here that if the Wiener gain function [2] is used, then the spectral subtraction is optimal in an asymptotic linear minimum mean square error sense when the frame length goes to infinity and the signal and noise are assumed stationary. To the best of our knowledge, this is the first time optimality conditions are attributed to the spectral subtraction approach.

The spectral subtraction approach has become almost standard in speech enhancement. It has gained popularity because it is relatively easy to understand and implement, i.e., it only requires DFT of noisy signal, application of a gain function, and inverse DFT. Furthermore, the spectral subtraction makes minimal assumptions about the signal and noise, and when carefully implemented, it results in reasonably clear enhanced signals. This approach has been heuristically developed and experimentally optimized. It turns out that many of its aspects can be interpreted by the linear estimation signal subspace approach developed here. For example, in implementing the Wiener gain function, an estimate of the variance of each spectral component of the clean signal is needed. This estimate is obtained by subtracting an estimate of the variance of the noise spectral component from an estimate of the variance of the spectral component of the noisy signal. The estimate obtained in this way is often found to be negative. This difficulty was never resolved, and it was always attributed to a flaw of the variance estimator. However, if the existence of the signal subspace is taken into account, then negative variance estimates are hardly surprising, since zero quantities are being estimated by a nonperfect estimator. Such an estimator provides negative variance estimates 50% of the time if it is unbiased.

The major drawback of the spectral subtraction approach is that it provides residual noise with annoying noticeable tonal characteristics referred to as "musical noise." It represents nonstationary residual noise due to the time varying filter applied to the noisy signal. Since, otherwise, the quality of the enhanced signal is satisfactory, the goal of speech enhancement for many years has been to minimize the perception of the musical noise without hurting the clarity of the enhanced signal (see, e.g., [18]). It will be demonstrated here that the proposed signal subspace approach outperforms the spectral subtraction in the sense that it almost eliminates the musical noise without elevating signal distortion.

We note that the signal subspace approach has its roots in classical detection theory [11]. In an M-ary detection problem, the signal subspace is spanned by the M signals, and detection is performed using the M KLT components which represent the signal subspace. The signal subspace has also been utilized in spectral estimation and array processing problems. The multiple signal characterization (MUSIC) algorithm for frequency and bearing estimation (see, e.g., [22] and [13]) uses orthogonality of vectors from the signal and noise subspaces. Another interesting application is the nonlinear prediction of speech signals proposed and studied in [23], [24]

The organization of this paper is as follows. In Section II, the principles of the signal subspace approach are discussed. In Section III, the proposed linear estimators are derived. In Section IV, the relations between the proposed linear estimator and other estimators, such as the linear minimum mean square error (LMMSE), the least squares (LS), and the spectral subtraction estimators are studied. In Section V, we discuss implementation issues resulting from lack of explicit knowledge of the second-order statistics of the signal and noise. Furthermore, we study the performance of the proposed speech enhancement approach. Comments are given in Section VI.

II. Signal Subspace Principles

In this section, the principles of the signal subspace approach are described. This can be best done by assuming a *linear model* for the clean signal. Alternatively, the fact that the covariance matrices of speech vectors are positive semidefinite can be used.

A. Signal and Noise Models

The *linear model* for the clean signal assumes that each K-dimensional vector y of that signal can be represented as

$$y = \sum_{m=1}^{M} s_m V_m, \quad M \leq K \tag{1}$$

where in general, $\{s_1, \cdots, s_M\}$ are zero mean complex random variables, and $V_1, \cdots, V_M$ are K-dimensional complex basis vectors, which are assumed linearly independent. Clearly, if $M = K$, such representation is always possible. For speech signals, however, such representation is also possible when $M < K$. The resulting model (1) is consistent with reliable, commonly used models for speech signals. The *damped* *complex sinusoid* model whose mth basis vector is given by [25]

$$V_m = \left(1, \rho_m^1 e^{j\omega_m 1}, \cdots, \rho_m^{K-1} e^{j\omega_m (K-1)}\right)^T \quad (2)$$

is the best known model [8]–[10]. In (2), $0 \leq \omega_m \leq 2\pi$, $0 \leq \rho_m \leq 1$ and $(\cdot)^T$ denotes vector transpose. Other related models such as the *pure complex sinusoid* model, the *damped real sinusoid* model, and the *pure real sinusoid* model, are particular cases of (2) [25]. Note that the model (1), (2) is implicitly used in the successful adaptive transform coder [7]. At rates of 9.6–16.0 kb/s, many spectral components are assigned zero bits, i.e., $s_m \equiv 0$, and the encoded signal is reconstructed using $M \ll K$ components. Nevertheless, high quality and intelligibility encoded speech signals are obtained.

The model (1) can be written as

$$y = Vs \quad (3)$$

where $V \triangleq [V_1, \cdots, V_M]$ is a $K \times M$ matrix whose rank is M, and $s \triangleq (s_1, \cdots, s_M)^T$. When $M < K$, the set of all possible signal vectors $\{y\}$ lie in a subspace of the Euclidean space R^K spanned by the columns of V. This subspace is referred to as the "signal subspace." When $M = K$, the signal subspace coincides with the entire R^K space. The latter situation will not be considered further since it is less relevant to speech signals. The covariance matrix of the vector y is given by

$$R_y \triangleq E\{yy^\#\} = VR_sV^\# \quad (4)$$

where $(\cdot)^\#$ denotes vector conjugate transpose, and R_s denotes the covariance matrix of the vector s, which is assumed positive definite. Hence, the rank of R_y is M, and this matrix has $K - M$ zero eigenvalues, as expected from a covariance matrix of a speech vector.

It turns out that the exact type of the linear model is of no importance for signal enhancement. The relationship of $M < K$, however, plays a major role in the enhancement process, since it implies that the covariance matrices of the signal have zero eigenvalues.

Let w denote a K-dimensional vector of the noise process. The noise process is assumed zero mean, additive, and uncorrelated with the clean signal. The covariance matrix R_w of the noise vector w is assumed known and positive definite. Hence, it can be assumed, without loss of generality, that the noise is white, since the noise can always be whitened by applying the linear transformation $R_w^{-1/2}$ to w [12]. Thus, we proceed with the assumption that

$$R_w = E\{ww^\#\} = \sigma_w^2 I. \quad (5)$$

In this case, the rank of the covariance matrix of the noise equals the dimension of the vector space K, and the noise vectors fill in the entire Euclidean space R^K. Thus, the noise exists in the signal subspace as well as in the complement of that subspace. The latter is referred to as the "noise subspace."

This discussion indicates that the Euclidean space of the noisy signal is composed of a signal subspace and a complementary noise subspace. We now show that decomposition of the noisy Euclidean space into these two subspaces can be performed by applying the KLT to the noisy signal [11]–[13].

B. Signal and Noise Subspaces

Let z denote a K-dimensional vector of the noisy signal. From (3) and (4) we have that

$$z = Vs + w \quad (6)$$

and the covariance matrix of z is given by

$$R_z \triangleq E\{zz^\#\} = VR_sV^\# + R_w. \quad (7)$$

Let $R_z = U\Lambda_z U^\#$ be the eigendecomposition of R_z. Here, $U \triangleq [u_1, \cdots, u_K]$ denotes an orthonormal matrix of eigenvectors $\{u_k \in R^K\}$ of R_z, and $\Lambda_z \triangleq \text{diag}(\lambda_z(1), \cdots, \lambda_z(K))$ denotes a diagonal matrix of eigenvalues of R_z. Since the noise is assumed white, the eigenvectors of R_z are also the eigenvectors of both R_y and R_w. Furthermore, all eigenvalues of R_w equal σ_w^2. Since $\text{rank}(R_y) = M$, the matrix R_y has M positive eigenvalues and $K - M$ zero eigenvalues. Assume, without loss of generality, that the M positive eigenvalues of R_y are $\{\lambda_y(1), \cdots, \lambda_y(M)\}$ and the corresponding M eigenvectors are $\{u_1, \cdots, u_M\}$. For convenience, assume that $\{\lambda_y(1), \cdots, \lambda_y(M)\}$ are given in a descending order. Multiplying (7) by u_k, which is the common kth eigenvector of the three covariance matrices, we obtain

$$\lambda_z(k) = \begin{cases} \lambda_y(k) + \sigma_w^2 & \text{if } k = 1, \ldots, M \\ \sigma_w^2 & \text{if } k = M+1, \cdots, K. \end{cases} \quad (8)$$

Thus, the eigendecomposition of R_z is given by

$$R_z = U\Lambda_z U^\# \quad (9)$$

$$\Lambda_z = \text{diag}[\Lambda_{z,1}, \sigma_w^2 I] \quad (10)$$

$$\Lambda_{z,1} \triangleq \text{diag}(\lambda_z(1), \cdots, \lambda_z(M)) \quad (11)$$

and the eigendecomposition of R_y is given by

$$R_y = U\Lambda_y U^\# \quad (12)$$

$$\Lambda_y = \text{diag}[\Lambda_{y,1}, 0I] \quad (13)$$

$$\begin{aligned} \Lambda_{y,1} &= \text{diag}(\lambda_y(1), \cdots, \lambda_y(M)) \\ &= \Lambda_{z,1} - \sigma_w^2 I. \end{aligned} \quad (14)$$

The eigenvalues in $\Lambda_{z,1}$ and their corresponding eigenvectors are referred to as the principal eigenvalues and eigenvectors of R_z, respectively.

Let $U = [U_1, U_2]$ where U_1 denotes the $K \times M$ matrix of principal eigenvectors of R_z, i.e.

$$U_1 = \{u_k: \lambda_z(k) > \sigma_w^2\}. \quad (15)$$

Since U is orthonormal

$$I = U_1U_1^\# + U_2U_2^\#. \quad (16)$$

The matrix $U_1U_1^\#$ is idempotent and Hermitian. Hence, it is the orthogonal projector onto the subspace spanned by the columns of U_1, but span U_1 = span V [13, p. 454]. Hence, $U_1U_1^\#$ is the orthogonal projector onto the signal subspace. The complementary orthogonal subspace is spanned by columns of U_2 and it constitutes the noise subspace. The matrix $U_2U_2^\#$ is the orthogonal projector on that subspace.

Thus, from (16), a vector z of the noisy signal can be decomposed as

$$z = U_1 U_1^{\#} z + U_2 U_2^{\#} z \tag{17}$$

where $U_1 U_1^{\#} z$ is the projection of z onto the signal subspace and $U_2 U_2^{\#} z$ is the projection of z onto the noise subspace. The coefficient vectors of the two projections $U_1^{\#} z$ and $U_2^{\#} z$, respectively, are obtained from $U^{\#} z$ which is the KLT of z. Note that since

$$\begin{aligned} E\{U^{\#} z\} &= 0 \\ \operatorname{cov}(U^{\#} z) &= \operatorname{diag}[\Lambda_{y,1} + \sigma_w^2 I, \sigma_w^2 I] \end{aligned} \tag{18}$$

$\operatorname{cov}(U_2^{\#} z) = \sigma_w^2 I$, and the signal energy in the vector $U_2^{\#} z$ is zero. Hence, with probability one (w.p. 1), this vector does not contain signal information and can be nulled when estimating the clean signal.

III. Linear Signal Estimators

In this section, we derive the linear estimators which minimize the signal distortion while constraining the energy and spectrum of the residual noise. We first derive the estimator which maintains the energy of the residual noise in the entire frame below a given threshold. Then, we derive the estimator which guarantees that the energy of the residual noise in each spectral component is kept below a given threshold. The first estimator allows time domain constraint (TDC) on the residual noise, while the second is designed for noise shaping using spectral domain constraints (SDC). Both approaches result in interesting and intuitive estimators.

A. Time Domain Constrained Estimator

Let $\hat{y} = Hz$ be a linear estimator of y where H is a $K \times K$ matrix. The residual signal obtained in this estimation is given by

$$\begin{aligned} r &= \hat{y} - y \\ &= (H - I)y + Hw \\ &\triangleq r_y + r_w \end{aligned} \tag{19}$$

where $r_y \triangleq (H - I)y$ represents signal distortion and $r_w \triangleq Hw$ represents the residual noise. Let

$$\bar{\epsilon}_y^2 \triangleq \operatorname{tr} E\{r_y r_y^{\#}\} = \operatorname{tr}\{(H - I) R_y (H - I)^{\#}\} \tag{20}$$

be the energy of the signal distortion vector r_y. Similarly, let

$$\bar{\epsilon}_w^2 \triangleq \operatorname{tr} E\{r_w r_w^{\#}\} = \sigma_w^2 \operatorname{tr}\{HH^{\#}\} \tag{21}$$

denote the energy of the residual noise vector r_w. The linear estimator with TDC on the residual noise is obtained from

$$\begin{gathered} \min_H \bar{\epsilon}_y^2 \\ \text{subject to: } \frac{1}{K} \bar{\epsilon}_w^2 \leq \alpha \sigma_w^2 \end{gathered} \tag{22}$$

where $0 \leq \alpha \leq 1$. The estimator derived in this way minimizes the signal distortion over all linear filters which result in the permissible residual noise level $\alpha \sigma_w^2$. The value of α is restricted to [0, 1], since clearly α cannot take negative values. In addition, for $\alpha \geq 1$ the optimal filter which satisfies the constraint and results in minimum (zero) signal distortion is $H = I$.

The optimal estimator in the sense of (22) can be found using the Kuhn–Tucker necessary conditions for constrained minimization [26]. Specifically, H is a stationary feasible point if it satisfies the gradient equation of the Lagrangian

$$L(H, \mu) = \bar{\epsilon}_y^2 + \mu(\bar{\epsilon}_w^2 - \alpha K \sigma_w^2) \tag{23}$$

and

$$\mu(\bar{\epsilon}_w^2 - \alpha K \sigma_w^2) = 0 \quad \text{for} \quad \mu \geq 0. \tag{24}$$

From $\nabla_H L(H, \mu) = 0$ we obtain

$$H_{\text{TDC}} = R_y (R_y + \mu \sigma_w^2 I)^{-1} \tag{25}$$

where μ is the Lagrange multiplier. From (24) we have that

$$\bar{\epsilon}_w^2 = \alpha K \sigma_w^2. \tag{26}$$

Substituting $H = H_{\text{TDC}}$ from (25) in (26) we find that μ must satisfy

$$\alpha = \frac{1}{K} \operatorname{tr}\{R_y^2 (R_y + \mu \sigma_w^2 I)^{-2}\}. \tag{27}$$

Hence, the optimal filter (25) is a Wiener filter with adjustable input noise level $\mu \sigma_w^2$.

Applying the eigendecomposition (12) of R_y to (25), we can rewrite the optimal linear estimator as

$$H_{\text{TDC}} = U \begin{bmatrix} G_\mu & 0 \\ 0 & 0 \end{bmatrix} U^{\#} \tag{28}$$

where

$$G_\mu \triangleq \Lambda_{y,1} (\Lambda_{y,1} + \mu \sigma_w^2 I)^{-1}. \tag{29}$$

Hence, the signal estimate $\hat{y}_{\text{TDC}} = H_{\text{TDC}} z$ is obtained by applying the KLT to the noisy signal, appropriately modifying the components of the KLT $U^{\#} z$ by a gain function, and by inverse KLT of the modified components. Note that the gain modification nulls noisy components which lie in the noise subspace. Hence

$$H_{\text{TDC}} = U_1 G_\mu U_1^{\#}. \tag{30}$$

This implies that, in the proposed estimation, the coefficients of the projection of the noisy signal onto the signal subspace $(U_1^{\#} z)$ are first calculated, then those coefficients are modified by G_μ and used to reconstruct the signal in the signal subspace. The linear estimator (30) can also be written more explicitly as

$$H_{\text{TDC}} = \sum_{m=1}^{M} g_\mu(m) u_m u_m^{\#} \tag{31}$$

where $g_\mu(m)$ denotes the mth diagonal element of G_μ given by

$$g_\mu(m) = \frac{\lambda_y(m)}{\lambda_y(m) + \mu \sigma_w^2}. \tag{32}$$

A block diagram of this estimator is shown in Fig. 1.

The Lagrange multiplier constraint (27) can also be simplified if the eigendecomposition (12) of R_y is used. In this case,

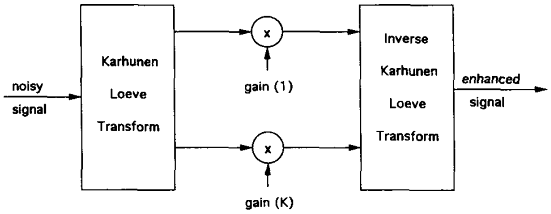


Fig. 1. Signal subspace linear estimator.

we find that μ must satisfy

$$\alpha = \frac{1}{K}\mathrm{tr}\{\Lambda_{y,1}^2(\Lambda_{y,1} + \mu\sigma_w^2 I)^{-2}\}. \tag{33}$$

Note that, since rank $(R_y) = M < K$, the maximum value that α can take is $\alpha_{\max} = M/K < 1$. For, if $\alpha \geq \alpha_{\max}$, the optimal filter is given by $G_\mu = I$. This filter provides zero signal distortion while satisfying the constraint in (22). From (33), we see that $\mu \in [0, \infty]$ when $\alpha \in [0, \alpha_{\max}]$, where $\alpha = 0$ implies $\mu = \infty$ and $\alpha = M/K$ results in $\mu = 0$. Hence, μ which satisfies (33) also satisfies (24) and therefore is the Lagrange multiplier for the inequality constrained optimization problem (22). Equation (33) has a single solution since α is a monotonically decreasing continuous function of μ.

B. Spectral Domain Constrained Estimator

We now derive the linear estimator which minimizes the signal distortion subject to constraints on the spectrum of the residual noise. This spectrum can be made similar to that of the speech, and thus, the residual noise can be masked by the speech signal. The kth spectral component of the residual noise is given by $u_k^\# r_w$. For $k = 1, \cdots, M$, we require that the energy in $u_k^\# r_w$ be smaller than or equal to $\alpha_k \sigma_w^2$, where $0 < \alpha_k \leq 1$. For $k = M+1, \cdots, K$, we require that the energy in $u_k^\# r_w$ be zero, since the signal energy in the noise subspace is zero. Hence, the filter H is designed by

$$\begin{aligned} &\min\nolimits_H \epsilon_y^2 \\ \text{subject to: } &E\{|u_k^\# r_w|^2\} \leq \alpha_k\sigma_w^2, \quad k = 1, \cdots, M \\ &E\{|u_k^\# r_w|^2\} = 0, \qquad k = M+1, \cdots, K. \end{aligned} \tag{34}$$

Following an optimization procedure similar to that used in the time domain constrained problem, while taking into account that the matrix H can now have complex entries, it can be shown that the optimal H must satisfy the following gradient matrix equation:

$$HR_y + \sigma_w^2 LH - R_y = 0 \tag{35}$$

where

$$L \triangleq U\Lambda_\mu U^\# \tag{36}$$

and $\Lambda_\mu = \mathrm{diag}(\mu_1, \cdots, \mu_K)$ is a diagonal matrix of Lagrange multipliers. Applying the eigendecomposition (12) of R_y to (35) we obtain

$$(I - Q)\Lambda_y - \sigma_w^2 \Lambda_\mu Q = 0 \tag{37}$$

where $Q \triangleq U^\# HU$. A possible solution to (37) is obtained when Q is diagonal with elements given by

$$q_{kk} = \begin{cases} \frac{\lambda_y(k)}{\lambda_y(k) + \sigma_w^2\mu_k} & k = 1, \cdots, M \\ 0 & k = M+1, \cdots, K. \end{cases} \tag{38}$$

For this Q, we have

$$E\{|u_k^\# r_w|^2\} = \begin{cases} \sigma_w^2 q_{kk}^2 & k = 1, \cdots, M \\ 0 & k = M+1, \cdots, K. \end{cases} \tag{39}$$

If the nonzero constraints in (34) are satisfied with equality, then $\sigma_w^2 q_{kk}^2 = \alpha_k\sigma_w^2$ implies that

$$q_{kk} = (\alpha_k)^{1/2}, \quad k = 1, \cdots, M \tag{40}$$

and

$$\mu_k = \frac{\lambda_y(k)}{\sigma_w^2}[(1/\alpha_k)^{1/2} - 1], \quad k = 1, \cdots, M. \tag{41}$$

Since $\mu_k \geq 0$, the Kuhn–Tucker necessary conditions for the constrained minimization are satisfied by the proposed solution (38). Hence, from (38) and (40) we conclude that the desired H is given by

$$\begin{aligned} H &= UQU^\# \\ Q &= \mathrm{diag}\ \ (q_{11}, \cdots, q_{KK}) \\ q_{kk} &= \begin{cases} \alpha_k^{1/2} & k = 1, \cdots, M \\ 0 & k = M+1, \cdots, K. \end{cases} \end{aligned} \tag{42}$$

From (42) we see that the choice of $\{\alpha_k\}$ completely specifies the gains of the estimator. This is not surprising since the estimator is linear and the spectra of its input and output signals are known. The input noise is white with spectrum σ_w^2, and the nonzero spectrum of the output residual noise is $\alpha_k\sigma_w^2$. In theory, $\{\alpha_k\}$ can be chosen independently of the statistics of the signal and noise. In this case, the second-order statistics of the signal and noise affect the estimator through the KLT only. This is a dual situation to the one that exists in the spectral subtraction approach (see Section IV–C). In the latter case, the second-order statistics of the signal and noise affect only the gain function of the estimator while the transform, i.e., the DFT, is signal independent.

A possible choice for α_k is

$$\alpha_k = \left(\frac{\lambda_y(k)}{\lambda_y(k) + \sigma_w^2}\right)^\gamma \tag{43}$$

where $\gamma \geq 1$ is an experimentally determined constant. This choice makes the spectrum of the residual noise look similar to that of the clean signal. The value of γ controls the suppression level of the noise as well as the resulting signal distortion. When γ increases, the permitted residual noise level decreases and the signal distortion level increases. It is interesting to note that this gain function has been commonly used in the spectral subtraction approach [2].

An alternative choice for α_k which results in a more aggressive noise suppression gain function is given by

$$\alpha_k = \exp\{-\nu\sigma_w^2/\lambda_y(k)\} \tag{44}$$

where $\nu \geq 1$ is an experimentally chosen constant. Similar to γ, the value of ν controls the suppression level of the

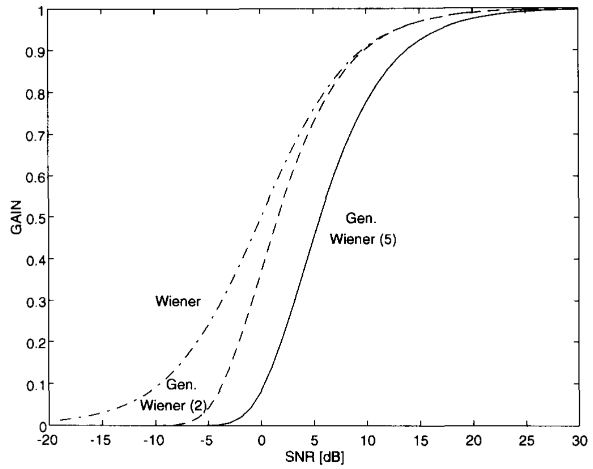


Fig. 2. Wiener and generalized-Wiener gain functions.

noise as well as the resulting signal distortion. The motivation for choosing this gain function is that for $\nu = 2$, the first-order Taylor approximation of $\alpha_k^{-1/2}$ is precisely the inverse of the Wiener gain function $g_1(k)$ given in (32). Hence, we refer to this function as generalized Wiener gain function. The Wiener and generalized Wiener gain functions are depicted in Fig. 2 for $\nu = 2$ and $\nu = 5$. The generalized Wiener gain function with $\nu = 5$ was found particularly useful in speech enhancement. Its performance will be described in Section V.

IV. LMMSE, LS, and Spectral Subtraction

We now study the relations between the TDC linear signal estimator (30) and the LMMSE, LS, and spectral subtraction signal estimators.

A. LMMSE Estimation

The LMMSE signal estimator is a particular case of the TDC linear estimator (30). It is obtained from (30) when $\mu \equiv 1$. The canonical form of the LMMSE estimator shown in Fig. 1 is similar to that given in [11, p. 203] for continuous time signals. The main difference between the two estimators is that in [11], the elimination of KLT components which contain noise only is done implicitly, since for those components,$\lambda_y(m) = 0$ and hence, $g_1(m) = 0$. The scheme in Fig. 1 makes it explicit that only $M < K$ components are expected to contain signal information. This subtle difference is important when $\lambda_y(m)$ is estimated from empirical data, since zero eigenvalues may inevitably have negative estimates. Such estimates will always be obtained when an unbiased nonperfect estimator is used. Thus, the number of zero eigenvalues must be estimated in advance, and the corresponding KLT components must be eliminated. Note that this strategy is made possible by recognizing the fact that some signal eigenvalues must be zero. If this fact is overlooked, then negative estimates of signal eigenvalues can only be mistakenly attributed to flaws of the estimator. This has often been done in the spectral subtraction literature where a similar problem occurs.

B. LS Estimation

The LS estimator of y given z is obtained from

$$\min_{\{y=Vs\}} ||z - y||^2 \tag{45}$$

where V is assumed known. This estimator is given by [12, p. 365]

$$\begin{aligned}\hat{y}_{\text{ls}} &= V(V^{\#}V)^{-1}V^{\#}z \\ &\overset{\triangle}{=} P_V z\end{aligned} \tag{46}$$

where P_V is the projection matrix onto the column space of V or onto the signal subspace [12, p. 366]. Since the signal subspace is also spanned by the columns of U_1, we have that [27, p. 197]

$$P_V = U_1(U_1^{\#}U_1)^{-1}U_1^{\#} = U_1U_1^{\#}. \tag{47}$$

Hence, $\hat{y}_{\text{ls}}$ can be obtained from (30) using $\mu \boxminus 0$. This means that the LS estimator only projects the noisy signal onto the signal subspace, but it does not modify the projected signal. Thus, this estimator results in the lowest possible (zero) signal distortion and in the highest possible residual noise level $\alpha_{\max}\sigma_w^2$.

The relationship expressed in (47) is important, since it shows that the matrix V need not be explicitly known in implementing the LS estimator. The estimator can be obtained from eigendecomposition of the covariance matrix of the noisy vector. This fact has often been overlooked in the speech literature. Signal estimators have been designed using estimates of the parameters of the vectors of V, e.g., the frequencies $\{\omega_m\}$ and the damping factors $\{\rho_m\}$, obtained from the noisy signal (see, e.g., [9] and [10]). The latter approach requires detailed specification of the linear model, and the solution of a hard parameter estimation problem. More importantly, the signal estimate obtained in this way is not necessarily optimal in the LS sense or in any other known sense.

For completeness of the discussion, we note that an LS estimate of V can be obtained from joint LS estimates of s and V [28]. The latter are obtained from

$$\min_{s,V} ||z - Vs||^2. \tag{48}$$

The minimization of (48) over s results in

$$\hat{s}_{\text{ls}} = (V^{\#}V)^{-1}V^{\#}z. \tag{49}$$

On substituting $\hat{s}_{\text{ls}}$ in (48), the estimate of V results from

$$\max_V \operatorname{tr}\{P_V zz^{\#}\} \tag{50}$$

where maximization is performed over the feasible set of parameters of V. The estimation of V by (50) requires a nontrivial maximization in an M-dimensional space. A simpler approach to this problem that requires 1-D minimization is the popular MUSIC approach [22], [13]. This approach capitalizes on orthogonality of the column vectors of V and the vectors which span the noise subspace.

C. Spectral Subtraction Estimation

In spectral subtraction signal estimation the DFT is first applied to the noisy signal. Spectral components whose estimated variance is smaller than or equal to the noise spectral component variance are nulled. The surviving spectral components are modified by a gain function and the inverse DFT is applied.

The nulling of noisy spectral components by the spectral subtraction approach is analogous to elimination of the noise subspace. This can be explained if we assume that the DFT has the desired properties of the KLT. This will be the case, for example, when the covariance matrix of the noisy signal is circulant [21]. Under this assumption, some of the spectral components of the clean signal must have zero variance. When the variance of such component is estimated by subtracting the variance of the noise spectral component from the estimated variance of the noisy spectral component, as is done in the spectral subtraction approach, negative estimate is highly likely since the estimator is not perfect. In fact, any unbiased estimator will provide negative variance estimates 50% of the time. Hence, nulling of noisy spectral components which result in negative signal spectral variance estimates is equivalent to nulling spectral components which carry no signal information or spectral components in the noise subspace.

Let $D^{\#}$ denote the $K \times K$ matrix that represents the DFT. The spectral components of the noisy signal are given by the vector $D^{\#}z$. Similarly, the spectral components of the noise process are given by the vector $D^{\#}w$. Assume that only $M < K$ components of $D^{\#}z$ are such that each has variance greater than that of the corresponding component of $D^{\#}w$. If the columns of D that result in those M components are arranged in a $K \times M$ matrix denoted by D_1, and $D = [D_1, D_2]$, then the spectral subtraction estimator can be written as

$$\hat{y}_{\text{sps}} = \frac{1}{K} D \begin{bmatrix} G_{\text{sps}} & 0 \\ 0 & 0 \end{bmatrix} D^{\#} z \tag{51}$$

$$= \frac{1}{K} D_1 G_{\text{sps}} D_1^{\#} z \tag{52}$$

where G_{sps} is an $M \times M$ diagonal gain matrix. When the Wiener gain function is used, the mth diagonal element of G_{sps} is given by

$$g_{\text{sps}}(m) = \frac{\hat{f}_z(m) - \hat{f}_w(m)}{\hat{f}_z(m)} \tag{53}$$

where $\hat{f}_z(m) = E\{|(D^{\#}z)_m|^2\}/K$ denotes the variance of the mth spectral component of the noisy signal, and $\hat{f}_w(m) = E\{|(D^{\#}w)_m|^2\}/K$ denotes the variance of the mth spectral component of the noise process. Note that $\hat{f}_z(m)$ and $\hat{f}_w(m)$ are estimates of the power spectral densities of the signal and noise, respectively. The name "spectral subtraction" was originally motivated by the subtraction of the noise power spectral density from the power spectral density of the noisy signal.

The major difference between the LMMSE ((30) with $\mu = 1$) and the spectral subtraction estimator (52) is in the transformation used to decompose the vector space of the noisy signal. The LMMSE uses the KLT while the spectral subtraction estimator uses the DFT. The two transforms coincide if the covariance matrix R_y is circulant [21]. Hence, under this condition, the spectral subtraction approach (52), (53) is optimal in the LMMSE sense. If R_y is not circulant but Toeplitz, then $\hat{y}_{\text{sps}}$ is only asymptotically optimal in the LMMSE sense. This is proven in the Appendix, where it is shown that if the signal and noise are stationary, $\hat{y}_{\text{lmmse}} - \hat{y}_{\text{sps}}$ converges in the mean to zero as $K \to \infty$, i.e.

$$\lim_{K\to\infty} \frac{1}{K} E\{||\hat{y}_{\text{lmmse}} - \hat{y}_{\text{sps}}||^2\} = 0. \tag{54}$$

D. Estimation from Empirical Data

The TDC linear estimator (25) and the spectral subtraction estimator (51) assume exact knowledge of the second-order statistics of the noisy signal and the noise process. In practice, this information is not available and is estimated from the noisy data. We now show that, under stationary and ergodic conditions, the implementable version of the linear estimator converges w.p. 1 to (25). In addition, the implementable version of the spectral subtraction estimator converges in probability to the LMMSE estimator ((25) with $\mu = 1$). This means that, under these conditions, the spectral subtraction estimator is asymptotically optimal in the LMMSE sense.

Consider estimation of the clean vector y_t at time t from the corresponding noisy vector z_t. Let $\hat{R}_{z_t}$ denote the estimate of the covariance matrix of z_t. This estimate can be obtained from the empirical covariance of $2T + 1$ nonoverlapping vectors of the noisy signal in the neighborhood of z_t, i.e., $\{z_{t-T}, \cdots, z_t, \cdots, z_{t+T}\}$, as shown in Fig. 3. Let Z_n be a K-dimensional vector which comprises the *samples* $[n, \cdots, n + K - 1]$ of the noisy signal. Note from Fig. 3 that $z_t = Z_{(t-1)K+1}$, $t = 1, 2, \ldots$. Then, $\hat{R}_{z_t}$ can be obtained from

$$\hat{R}_{z_t} = \frac{1}{2TK} \sum_{n=(t-T-1)K+1}^{(t+T-1)K} Z_n Z_n^{\#}. \tag{55}$$

Similarly, let $\hat{R}_{w_t}$ be the estimate of the noise covariance R_{w_t}. This estimate can be obtained from a segment of the noisy signal which is known to contain noise only. If the signal and noise are assumed stationary and ergodic, then $\hat{R}_{z_t}$ and $\hat{R}_{w_t}$ converge w.p. 1 as $T \to \infty$ to the true covariances R_{z_t} and R_{w_t}, respectively. Hence, as $T \to \infty$, the estimate of the covariance matrix of the clean vector y_t

$$\hat{R}_{y_t} \triangleq \hat{R}_{z_t} - \hat{R}_{w_t} \tag{56}$$

approaches the true covariance R_y of y_t w.p. 1, and the implementable version of the linear estimator

$$\hat{H}_{\text{TDC}} = \hat{R}_{y_t}(\hat{R}_{y_t} + \mu\hat{R}_{w_t})^{-1} \tag{57}$$

approaches (25) w.p. 1.

To prove convergence of the spectral subtraction estimator to the LMMSE estimator, we first show that the implementable version of the spectral subtraction estimator converges w.p. 1 to (51). Then we combine this result with (54). The implementable version of the spectral subtraction estimator uses estimates of the spectral variances $\hat{f}_{z_t}(m)$ and $\hat{f}_{w_t}(m)$, obtained

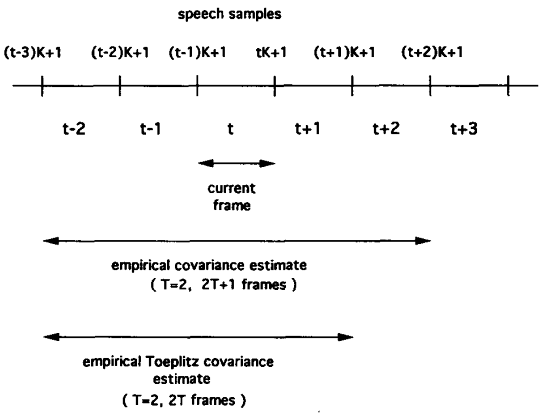

speech samples
(t-3)K+1
(t-2)K+1
(t-1)K+1
tK+1
(t+1)K+1
(t+2)K+1
t-2
t-1
t
t+1
t+2
t+3
current frame
empirical covariance estimate (T=2, 2T+1 frames)
empirical Toeplitz covariance estimate (T=2, 2T frames)

Fig. 3. Signal vectors and samples.

from sample averages as was done in estimating R_{z_t} and R_{w_t}, respectively. Under stationary and ergodic assumptions, as $T \to \infty$, these estimates converge w.p. 1 to the true variances $\hat{f}_{z_t}(m)$ and $\hat{f}_{w_t}(m)$, respectively. Hence, the implementable version of the spectral subtraction estimator converges w.p. 1 to (51).

Let $\tilde{y}_{\rm sps}$ denote the implementable spectral subtraction estimate of y. Almost sure convergence of this estimate to $\hat{y}_{\rm sps}$ implies convergence in probability. Hence, for every $\epsilon > 0$ and $\delta > 0$, there exists T_1 such that for all $T > T_1$

$$\Pr\{|\tilde{y}_{\rm sps} - \hat{y}_{\rm sps}| > \delta\} < \epsilon. \tag{58}$$

Similarly, convergence in the mean of $\hat{y}_{\rm sps}$ to $\hat{y}_{\rm lmmse}$ implies convergence in probability. Hence, there exists K_1, such that for all $K > K_1$

$$\Pr\{|\hat{y}_{\rm sps} - \hat{y}_{\rm lmmse}| > \delta\} < \epsilon. \tag{59}$$

Using the triangle inequality, the union bound, and (58) and (59), in that order, we have that

$$\begin{aligned}
&\Pr\{|\hat{y}_{\rm lmmse} - \tilde{y}_{\rm sps}| \le 2\delta\} \\
&\quad \ge \Pr\{|\hat{y}_{\rm lmmse} - \hat{y}_{\rm sps}| \le \delta, |\hat{y}_{\rm sps} - \tilde{y}_{\rm sps}| \le \delta\} \\
&\quad = 1 - \Pr\{|\hat{y}_{\rm lmmse} - \hat{y}_{\rm sps}| > \delta \cup |\hat{y}_{\rm sps} - \tilde{y}_{\rm sps}| > \delta\} \\
&\quad \ge 1 - [\Pr\{|\hat{y}_{\rm lmmse} - \hat{y}_{\rm sps}| > \delta\} + \Pr\{|\hat{y}_{\rm sps} - \tilde{y}_{\rm sps}| > \delta\}] \\
&\quad \ge 1 - \epsilon - \epsilon = 1 - 2\epsilon.
\end{aligned} \tag{60}$$

Hence

$$\Pr\{|\tilde{y}_{\rm sps} - \hat{y}_{\rm lmmse}| > 2\delta\} \le 2\epsilon \tag{61}$$

and $\tilde{y}_{\rm sps}$ converges in probability to $\hat{y}_{\rm lmmse}$ as $T \to \infty$ and $K \to \infty$.

Since both estimators $\tilde{y}_{\rm TDC}$ (the implementable version of (25)) and $\tilde{y}_{\rm sps}$ approach the linear estimator (25), it is important to establish the rate at which each estimator converges to $\hat{y}_{\rm TDC}$. The estimator with the larger rate of convergence will be better. This subject is currently under investigation. We note, however, that convergence of $\tilde{y}_{\rm sps}$ is achieved when both $T \to \infty$ and $K \to \infty$, while convergence of $\tilde{y}_{\rm TDC}$ is achieved when only $T \to \infty$.

V. Implementation and Evaluation

In this section, we discuss issues related to the implementation of the linear signal estimators developed here. Furthermore, we evaluate the performance of the best estimator and compare it with a version of the spectral subtraction estimator [29].

A. Implementation

In implementing the TDC linear estimator (31), (32), the permissible noise level coefficient α must be specified for each vector, and the corresponding Lagrange multiplier μ must be calculated using (33). Clearly the optimal α must be signal dependent since speech is not strictly stationary. One strategy for minimizing the effects of the residual noise is to choose α so that the level of the residual noise stays below the masking threshold of the auditory system. Thus, the residual noise becomes inaudible [30]. Unfortunately, this approach may result in a high level of signal distortion, especially when the masking threshold is low. Furthermore, this approach can not be presently tested, since estimation of masking thresholds in the time domain is not fully understood.

An alternative approach to implementing the TDC linear estimator is to specify μ and to evaluate its effect on α. This approach is more intuitive since the effects of μ on the residual noise and signal distortion are better understood. When μ increases from zero to infinity, the level of the residual noise decreases while the level of signal distortion increases. A useful approach for choosing μ that compromises between signal distortion and residual noise is to make μ dependent on the signal-to-noise ratio (SNR) in each frame of the noisy signal [31]. A simpler approach which has also been found to be useful is to set μ to a fixed value which is greater than 1. Typical useful values are $\mu = 2 - 3$.

The SDC estimator (43) is significantly easier to implement than the TDC estimator since the Lagrange multipliers were analytically calculated. The SDC estimator is also significantly more powerful than the TDC estimator since it allows constraints in the perceptually significant spectral domain. In addition, the TDC can be considered a particular case of the SDC if, for example, α_k is chosen to be

$$\alpha_k = \left(\frac{\lambda_y(k)}{\lambda_y(k) + \mu\sigma_w^2}\right)^2. \tag{62}$$

For these reasons, our main focus will be on the SDC estimator (42). For the given choice of gain function (44), this estimator depends only on one fixed parameter ν, whose value was experimentally chosen to be $\nu = 5$.

In implementing the linear signal estimators, one must have good estimates of R_{z_t} and R_{w_t}, the covariance matrices of the vectors of the noisy signal and the noise process, at time t, respectively. Furthermore, a good estimate of the dimension M_t of the signal subspace at time t is required. The covariance matrix R_{w_t} is used to whiten the input noise as explained in Section II-A. In general, R_{w_t} is estimated from vectors of the noisy signal during which speech is absent. If the noise is stationary, then estimation can be performed from an initial segment of the noisy signal which was recorded before speech was present. When the noise is not stationary, a speech/noise detector must be used, and the noise covariance is estimated and updated from nonspeech frames of the noisy signal. Thus, in principle, the problem of estimation R_{z_t} and R_{w_t} is the same.

Other parameters which must be specified are those characterizing the analysis conditions, i.e., the frame length K, the overlap duration between adjacent frames, the type of analysis and synthesis windows, and the number $2T + 1$ of nonoverlapping K-dimensional vectors of the noisy signal from which R_{z_t} is estimated. The SDC linear estimator was applied to frames of the noisy signal which overlapped each other by 50%. In order to preserve the whiteness of the input noise, i.e., $R_{w_t} = \sigma_w^2 I$, only a rectangular analysis window could be used. The enhanced vectors were Hanning windowed and combined using the overlap and add synthesis approach [32].

We next address the problem of estimating R_{z_t}, and specifying the values of the related parameters K and T. The problem of estimating M_t is addressed in Section V-A-2).

1) Estimating R_{z_t}: As mentioned in Section IV-D, the covariance matrix of the vector z_t can be estimated from the vectors $\{Z_n, n = (t-T-1)K+1, \cdots, (t+T-1)K\}$ using the empirical covariance. This estimate can be rewritten as

$$\hat{R}_{z_t} = D_t D_t^{\#} \tag{63}$$

where

$$D_t \triangleq \frac{1}{(2TK)^{1/2}}[Z_{(t-T-1)K+1}, \cdots, Z_{(t-1)K+1}, \cdots, Z_{(t+T-1)K}] \tag{64}$$

is a $K \times 2KT$ data matrix [13]. If this covariance estimate is used, then the linear estimator can be implemented by applying either eigendecomposition to $\hat{R}_z$ or singular value decomposition (SVD) to the data matrix D_t [13].

The SVD approach requires less computations since it does not involve explicit estimation of the covariance matrix. The disadvantage of the SVD approach is that it applies to the empirical covariance estimate only. Thus, it does not allow the use of other, more structured, covariance estimates which may be more appropriate for speech signals. The need for such estimates results from the fact that speech signals are not strictly stationary. Thus, the amount of data from which the covariance can be estimated is limited and, therefore, the number of estimated covariance entries must be minimized.

One such covariance estimate which was found to be useful in our speech enhancement application is the empirical Toeplitz covariance. This estimate was constructed from the first K samples of the biased autocorrelation function estimate [33]. The latter estimate was obtained from $2TK$ samples of the noisy signal at instants $(t-T-1)K+1, \ldots, (t+T-1)K$, as is demonstrated in Fig. 3. This estimator was efficiently implemented using the FFT algorithm.

In order to implement the empirical Toeplitz covariance estimator, the values of K and T must be chosen. If the speech signal was strictly stationary, T and K would have been chosen to be as large as possible, because large KT guarantees good estimate of R_{z_t} [33]. In addition, the improvement in SNR obtained by the signal subspace approach is proportional to K/M, since the signal subspace dimension M is fixed, and the noise is evenly distributed in the entire K-dimensional space. Since speech signals are not strictly stationary, however, the values of K and T are restricted by the following constraints. First, $2TK$, the total number of speech samples used in estimating the Toeplitz covariance matrix, must be smaller than the period during which the signal can be considered stationary. Typically, this number equals 300–400 samples at 8 kHz sampling rate. Second, $K > M$ should be chosen so that the SNR improvement expected from the existence of the signal subspace can be utilized. Third, the larger T is, the more accurate $\hat{R}_{z_t}$ is since, on the average, there are $2T$ samples of the noisy signal per each estimated autocorrelation sample. Finally, K should be chosen small to reduce computational complexity in performing eigendecomposition of the estimated covariance.

In this work, we have obtained best results using $T = 5$ and $K = 40$. This amounts to estimating the covariance matrix from 400 samples of the noisy signal. For this value of K, only a few frames resulted in estimated M which was equal to K. This is not a problem since, for those frames, the signal occupies the entire space and the noise subspace is null.

2) Estimating M_t: Numerous approaches for estimating the order of a model were reported in the literature (see, e.g., [34]–[40]). In our case, the order is the dimension of the signal subspace, or the dimension of the vector s in the linear regression model (6) for the noisy signal. The most popular approach is the minimum description length (MDL) of Rissanen [34]. This approach was shown by Schwarz [36] to be optimal in the minimum probability of error sense in detecting the order of the model, assuming that it has an underlying exponential (Koopman–Darmois) probability distribution. A recent approach by Merhav *et al.* [38], [39] appears most promising. It guarantees minimization of the probability of underestimation of the order, uniformly for all processes in the given class, while maintaining exponentially decaying probability of overestimation of the order. The approach was applied to exponential probability density functions (pdf's) in [39]. These approaches require maximum likelihood (ML) estimation of the parameter set of the model.

The approach of [39] is applicable here if the pdf of the noise is assumed exponential. A particular case is that of Gaussian noise. The parameter set in this case is the vector s and the noise variance σ_w^2 in the model (6). Hence, there are $M + 1$ parameters in the model. The estimator of M as obtained from the vector z is given by

$$\hat{M} = \arg \min_{1 \le m \le \bar{M}} \left\{ \max_{s,\sigma_w} \frac{1}{K} \log p_{\bar{M}+1}(z|s,\sigma_w^2) - \max_{s,\sigma_w} \frac{1}{K} \log p_{m+1}(z|s,\sigma_w^2) < \delta \right\} - 1 \tag{65}$$

where $p_{m+1}(z|s,\sigma_w^2)$ denotes the pdf of the noisy signal z given s and σ_w^2, assuming that the dimension of the signal subspace is m, $\bar{M}$ denotes the maximum possible dimension of the signal subspace, and δ determines the exponential rate of decay of the probability of overestimation of M. This estimator chooses the smallest order for which the likelihood of the model of order m is close to that of the model with the maximum order up to a factor of δ.

Assuming Gaussian noise in the model (6), and a signal subspace dimension m, we now derive $\max_{s,\sigma_w} \log p_{m+1}(z|s,\sigma_w^2)$ by finding the ML estimates of s and σ_w^2. Since $Vs = U_{1,m}s'$ for some $s' \in R^m$, where $U_{1,m}$ is the $K \times m$ matrix of the eigenvectors $\{u_1, \cdots, u_m\}$ of R_z, we can equivalently find the ML estimates of s' and σ_w^2. From (49), the ML estimate of s' is given by $\hat{s'} = U^{\#}_{1,m}z$. Substituting $\hat{s'}$ in $\log p_{m+1}(z|s', \sigma_w^2)$ and maximizing over σ_w^2 gives

$$\hat{\sigma}_w^2(m) = \frac{1}{K}||U_{2,K-m}U^{\#}_{2,K-m}z||^2 \tag{66}$$

where $U_{2,K-m}$ is an $K \times (K-m)$ matrix of the eigenvectors $\{u_{m+1}, \cdots, u_K\}$ of R_z. Furthermore

$$\max_{s',\sigma_w} \frac{1}{K} \log p_{m+1}(z|s', \sigma_w^2) = \text{const} - \frac{1}{2} \log \hat{\sigma}_w^2(m). \tag{67}$$

Hence, from (65) the estimator of M is obtained from

$$\hat{M} = \arg \min_{1 \le m \le \bar{M}} \left\{ \frac{1}{2} \log \hat{\sigma}_w^2(m) - \frac{1}{2} \log \hat{\sigma}_w^2(\bar{M}) < \delta \right\} - 1. \tag{68}$$

Note that $\hat{\sigma}_w^2(m)$ represents the energy of the noisy signal in the noise subspace assuming that its dimension is $K - m$. As m increases, this energy decreases. The estimate $\hat{M}$ is chosen as the smallest value of m for which this energy is close (up to δ) to the minimum possible energy of the noisy signal in the noise subspace.

In principle, $\bar{M}$ can be as large as K. Since, however, we must guarantee that $\lambda_z(k) > \sigma_w^2$ for all $k \le \bar{M}$, the value of $\bar{M}$ is estimated as the number of eigenvalues of the estimated covariance R_z which exceed $\hat{\sigma}_w^2$. In our implementation of the estimator (68), we have used $\delta = .0025 \log \hat{\sigma}_w^2(\bar{M})$.

3) Implementation Summary: The linear estimator to be evaluated here is the SDC estimator defined by (42) and (44) with $\nu = 5$. This estimator was applied to frames of $K = 40$ samples of the noisy signal, at 8 kHz sampling rate, which overlapped each other by 50%. A rectangular analysis window was used. A Hanning window was used in the overlap and add synthesis procedure. The estimator for y_t was constructed from eigenvectors and eigenvalues of the estimated empirical Toeplitz covariance matrix of the noisy signal R_{z_t}. This covariance estimator was implemented using $T = 5$. The dimension M of the signal subspace was estimated from (68) using the eigenvectors of the estimated covariance.

B. Spectral Subtraction

The version of the spectral subtraction estimator implemented here was described in [29]. The estimator was applied to Trapezoidal windowed vectors of $K = 256$ samples of the noisy signal at 8 kHz sampling rate, which overlapped each other by $\Delta K = 76$ samples. The gain function is given by

$$g_{\text{sps}}(k) = \left(\frac{\max\{|Z(\omega_k)|^2/K - \mu\hat{\sigma}_w^2, \beta\hat{\sigma}_w^2\}}{|Z(\omega_k)|^2/K} \right)^{1/2} \tag{69}$$

where $Z(\omega_k)$ is the kth sample of the DFT of the noisy vector z, $\hat{\sigma}_w^2$ is an estimate of the variance of the white input noise, μ controls the amount of subtracted noise, and $\beta\hat{\sigma}_w^2$ represents a "spectral floor." In terms of our standard notation, $|Z(\omega_k)|^2/K$ in (69) represents an empirical estimate (periodogram) of $\hat{f}_z(k)$ in (53). The gain function in (69) is different from the implementable version of (53) in three aspects. First, an adjustable amount of input noise is subtracted from the estimated spectrum of the noisy vector. This has a similar effect as the Lagrange multiplier μ in the TDC linear estimation. The goal here is the same, i.e, to balance residual noise and signal distortion. Second, the minimum value that the estimated spectrum of the clean signal can have is $\beta\hat{\sigma}_w^2$, rather than zero. This results in the so-called "spectral floor" which was found useful in partial masking of the annoying musical residual noise. Third, the gain function represents the square root of the Wiener gain function. We obtained best subjective performance using $\mu = 2$ and $\beta = 0.005$.

C. Performance Evaluation

The proposed linear estimator and the spectral subtraction estimator were tested and compared in enhancing speech signals which have been degraded by computer generated additive white Gaussian noise at 10 dB input SNR. The SNR of the noisy signal is defined as

$$\text{SNR} = 10 \log_{10} \frac{\sum_{t=1}^{N} \sum_{k=1}^{K} y_t^2(k)}{\sum_{t=1}^{N} \sum_{k=1}^{K} (z_t(k) - y_t(k))^2} \tag{70}$$

where N is the number of frames in the given sentence, and $y_t(k)$ denotes the kth sample of y_t. Speech material which consists of two sentences spoken by three male speakers and three female speakers (a total of 12 sentences) was used. One of the sentences, "Why were you away a year, Roy?" contains vowels and glides only, and the other, "His vicious father had seizures" contains fricatives only.

The evaluation was performed by a group of 16 listeners. Four subjects were individuals working on different aspects of speech coding and enhancement. These subjects were familiar with the sentences. The other 12 subjects were six students and six professors at George Mason University. The authors were obviously excluded from this test. These 12 subjects were not familiar with the sentences. All subjects claimed to have normal hearing. Their ages ranged from 23 to 40 years old.

Each subject participated in two listening sessions. The goal of the first session was to compare the signal subspace approach with the plain noisy speech. The goal of the second session was to compare the signal subspace approach with the spectral subtraction approach. In each session, 12 pairs of sentences, each representing two different processing methods, were presented to the subject through headphones. The subject was asked to compare the two sentences and vote for one of them. The order of the sentences in each pair was randomized. The comparison was subjective based on the perceived amount and nature of residual noise, possible distortion and nature of the processed speech, etc. No listening fatigue effects were taken into account since each session was relatively short. In comparing the two sentences in a pair, the subjects could listen to them as many times as they wished. To minimize any bias, the subjects were not informed which versions of the speech material they would be comparing.

In the first session, 14 subjects preferred the speech material enhanced by the signal subspace approach over the nonprocessed noisy speech. On the average, the subjects in this group voted in favor of the signal subspace approach for 83.9% of the sentences with standard deviation of 16.5%. The 79% empirical confidence interval of the average score associated with the individual listeners was 83.9 ± 16.1%. This is the smallest symmetric interval around the mean which contains 79% of the scores in favor of the signal subspace obtained from the listeners. The 100% empirical confidence interval of the avearge score associated with the individual listeners was 83.9 ± 25.6%. The general consensus in this session was that the quality of the enhanced signal is far better than that of the raw noisy signal due to the reduction in the level of the input noise. For those sentences where the enhanced signal was preferred, the benefit of noise reduction was worth the slight distortion introduced by the noise removal algorithm. For the other sentences where the noisy signal was preferred, the distortion in the enhanced signal was more noticeable and/or the perception of the noise was tolerable. The two subjects who preferred the noisy signals over the enhanced signals did so for 66.7% of the sentences on the average, with standard deviation of 11.8%. These individuals preferred the "natural" sound of the raw signal and were not bothered by the presence of the noise. One even mentioned that she used to work in a bar and noise does not bother her.

In the second session, the same 14 individuals who preferred the signal subspace processing over the raw noisy signal also preferred the signal subspace processing over the spectral subtraction approach. On the average, the signal subspace approach was preferred for 98.2% of the sentences with standard deviation of 3.6%. The 79% empirical confidence interval of the average score associated with the individual listeners was 98.2 ± 1.8%. The 100% empirical confidence interval of the avearge score associated with the individual listeners was 98.2 ± 6.6%. The major complaint in this comparison was the noticeable annoying musical residual noise in the spectral subtraction approach. Such noise was not present in the sentences processed by the signal subspace approach. Nine subjects felt that the two approaches contribute a comparable amount of distortion to the speech signals during the noise removal process. The other five subjects felt that some sentences processed by the signal subspace approach are slightly more muffled than the sentences processed by the spectral subtraction approach. One subject also indicated that the tonal residual noise in the spectral subtraction approach may be especially destructive over the telephone, since it can easily be confused with tonal signals used in the network. The two subjects who preferred the spectral subtraction processing over the signal subspace approach did so for 83.3% of the sentences with standard deviation of 11.8%. These individuals were not bothered by the musical noise and felt that the spectral subtraction approach provides crisper enhanced signals.

Fig. 4 shows wide-band spectrograms of the clean signal, noisy signal, and the enhanced signals obtained by the spectral subtraction and the signal subspace approaches for the sentence, "Why were you away a year, Roy?" All signals were pre-emphasized to enable display of low energy formants. These spectrograms evidently demonstrate that both the signal subspace approach and the spectral subtraction approach extracted the formants which were not fully buried under the noise with similar success. The musical noise obtained in the spectral subtraction approach is clearly shown as narrow-band short duration signals which are randomly distributed in time and frequency. No such residual noise can be seen for the signal subspace approach. The SNR of the spectral subtraction processed sentence is 14.25 dB, and the SNR of the signal subspace processed sentence is 14.67 dB. The SNR of the enhanced signal is defined similarly to the SNR of the noisy signal in (70) with z being replaced by $\hat{y}$.

The improvement in SNR achieved by the spectral subtraction and the signal subspace approach was similar. The spectral subtraction approach elevated the SNR of the noisy signals by 4.21–6.02 dB, and the signal subspace approach improved the SNR of the noisy signals by 4.22–5.74 dB.

VI. Comments

A comprehensive framework for nonparametric speech enhancement was developed. The basic principle is to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace. Enhancement is performed by removing the noise subspace and estimating the signal from the remaining subspace. Linear estimation is performed with the goal of minimizing signal distortion while masking the residual noise by the signal.

Signal estimation is performed on a frame-by-frame basis. The vector space of the noisy signal is decomposed by applying an estimate of the KLT to the noisy signal. This estimate is obtained from eigendecomposition of a Toeplitz covariance estimate of the noisy vector. This covariance is calculated using the FFT. An estimate of the dimension of the signal subspace is obtained using the Merhav *et al.* approach.

The linear estimation signal subspace theory developed here was found useful in interpreting many aspects of the popular spectral subtraction speech enhancement approach. It was proven that a version of the spectral subtraction approach, which uses the Wiener gain function, is optimal in an asymptotic (large sample) linear minimum mean square error sense, assuming that the speech and noise are stationary. The existence of the signal subspace was used to explain why nulling of weak spectral components of the noisy signal is necessary. The estimation criterion proposed here confirms the intuition that subtracting more noise than actually exists balances signal distortion and residual noise level as conjectured in the spectral subtraction literature. Thus, this work provides a theoretical basis for the heuristically derived spectral subtraction approach.

The proposed signal subspace approach was judged better than the spectral subtraction in our speech enhancement application where the noise was additive and white. It provided

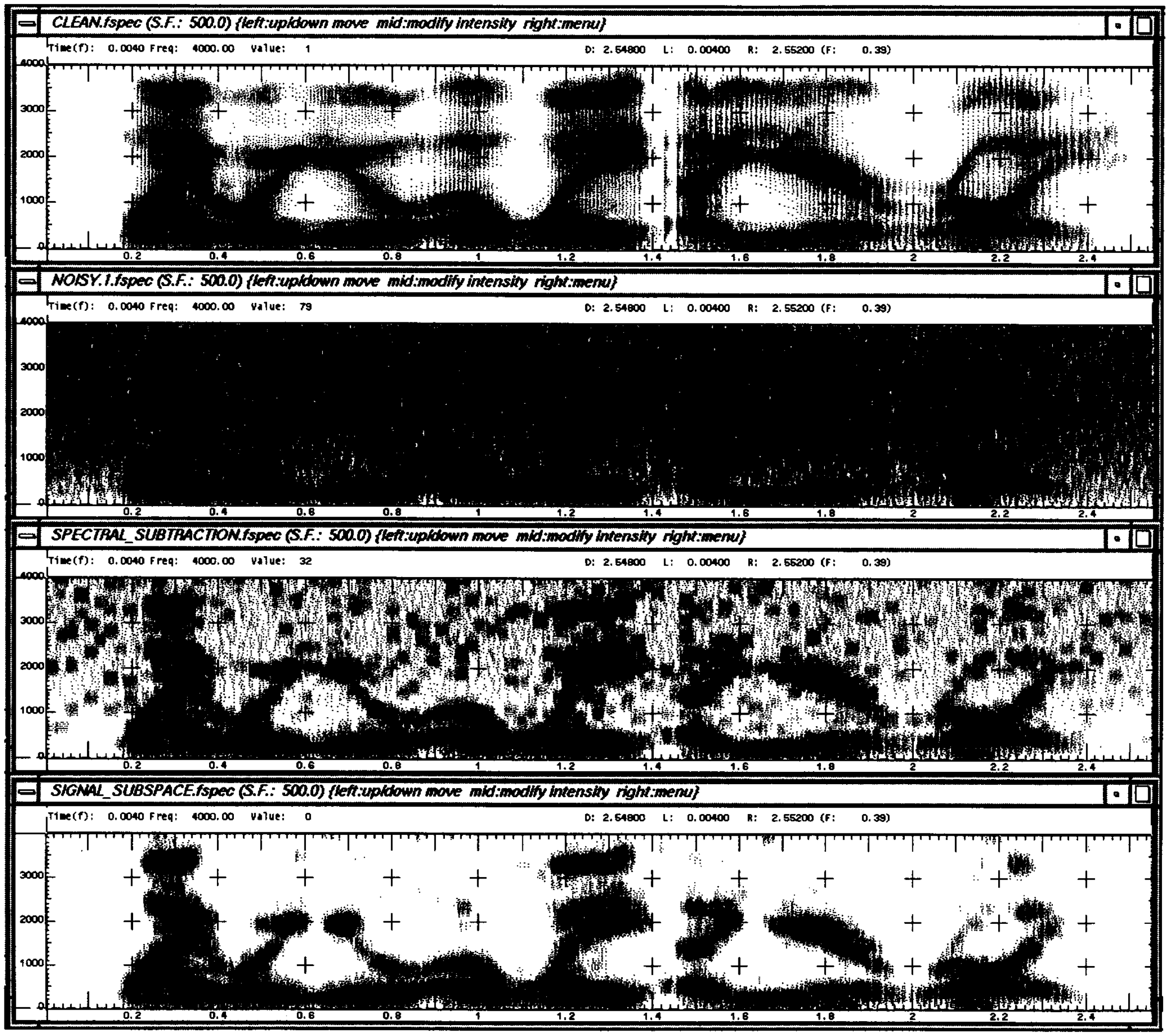


Fig. 4. Spectrograms of the clean signal, noisy signal, spectral subtraction enhanced signal, and signal subspace enhanced signal.

enhanced signals with comparable distortion to that obtained in the spectral subtraction approach but with essentially no musical residual noise.

The major difference between the spectral subtraction approach and the signal subspace approach is in the transform used to decompose the vector space of the noisy signal. The theoretically optimal transform is the KLT. The spectral subtraction approach uses the DFT, while the signal subspace approach was implemented using an empirical estimate of the KLT. While this transform was implemented using the FFT, it is still more computationally expensive than signal independent transforms such as the DFT, since it involves eigendecomposition of the empirical covariance. Future work should focus on studying other signal independent transforms which can well approximate the KLT of the speech signal. Important examples are the discrete cosine transform (DCT) and the discrete wavelet transform. The DCT was proposed as a good approximation to the theoretical KLT, especially when the signal can be modeled as a first-order Markov process [41].

Note that the signal decomposition approach proposed here is completely different from that proposed in [42]–[44]. Here, the Euclidean space of the noisy signal is decomposed into a signal-plus-noise subspace and a noise subspace. In [42]–[44], signal decomposition aims at estimating probabilistic models, i.e., hidden Markov models, for the signal and noise using composite modeling of the noisy data.

Appendix

We prove that the spectral subtraction estimator which uses the Wiener gain function converges in the mean to the LMMSE estimator as $K \to \infty$. No signal subspace details are involved in this proof. Hence, the noise is not necessarily assumed white. Since estimated components of the speech power spectral density $\{\hat{f}_y(m)\}$ are allowed to be zero, nulling of appropriate noisy spectral components which carry no signal information is automatically performed.

The spectral subtraction estimator is given by

$$\begin{aligned} \hat{y}_{\text{sps}} &= \frac{1}{K} D G_{\text{sps}} D^{\#} z \\ &\triangleq H_{\text{sps}} z \end{aligned} \tag{A.1}$$

where G_{sps} is a diagonal matrix with (k,k) element given by

$$g_{\text{sps}}(k,k) = \frac{\hat{f}_y(k)}{\hat{f}_y(k) + \hat{f}_w(k)}, \quad k = 0, \cdots, K-1. \tag{A.2}$$

The LMMSE estimator is given by

$$\begin{aligned} \hat{y}_{\text{lmmse}} &= R_y (R_y + R_w)^{-1} z \\ &\triangleq H_{\text{lmmse}} z. \end{aligned} \tag{A.3}$$

We prove that

$$\lim_{K \to \infty} \frac{1}{K} E\{||\hat{y}_{\text{lmmse}} - \hat{y}_{\text{sps}}||^2\} = 0 \tag{A.4}$$

under the following assumptions.

1) The signal and noise are stationary, i.e., for every K, R_y and R_w are Toeplitz

$$R_y = \{r_y(l-n), \quad l, n = 1, \ldots, K\} \tag{A.5}$$

$$R_w = \{r_w(l-n), \quad l, n = 1, \ldots, K\}. \tag{A.6}$$

2) The sequences $\{r_y(m)\}$ and $\{r_w(m)\}$ are absolutely summable

$$\sum_{m=-\infty}^{\infty} |r_y(m)| < \infty \tag{A.7}$$

$$\sum_{m=-\infty}^{\infty} |r_w(m)| < \infty. \tag{A.8}$$

3) The power spectral density associated with $r_w(m)$

$$f_w(\theta) = \sum_{m=-\infty}^{\infty} r_w(m) e^{-j\theta m} \tag{A.9}$$

satisfies

$$f_w(\theta) \ge m_{f_w} > 0. \tag{A.10}$$

Assumption 2 guarantees the existence and finiteness of the Fourier transforms $f_y(\theta)$ and $f_w(\theta)$ of $r_y(m)$ and $r_w(m)$, respectively. The least upper bounds of $f_y(\theta)$ and $f_w(\theta)$ are denoted by M_{f_y} and M_{f_w}, respectively. From Assumptions 2 and 3, we have that

$$0 \le f_y(\theta) \le M_{f_y} \tag{A.11}$$

$$m_{f_w} \le f_w(\theta) \le M_{f_w}. \tag{A.12}$$

Proof:

$$\begin{aligned} &\frac{1}{K} E\{||\hat{y}_{\text{lmmse}} - \hat{y}_{\text{sps}}||^2\} \\ &= \frac{1}{K} E\{||(H_{\text{lmmse}} - H_{\text{sps}})z||^2\} \\ &= \frac{1}{K} \text{tr}\{(H_{\text{lmmse}} - H_{\text{sps}}) R_z (H_{\text{lmmse}} - H_{\text{sps}})^{\#}\} \\ &= \frac{1}{K} \text{tr}\{R_z (H_{\text{lmmse}} - H_{\text{sps}})^{\#} (H_{\text{lmmse}} - H_{\text{sps}})\} \\ &\le |R_z| \cdot |(H_{\text{lmmse}} - H_{\text{sps}})^{\#} (H_{\text{lmmse}} - H_{\text{sps}})| \\ &\le ||R_z|| \cdot ||(H_{\text{lmmse}} - H_{\text{sps}})|| \cdot |(H_{\text{lmmse}} - H_{\text{sps}})| \\ &\le ||R_z|| \cdot (||H_{\text{lmmse}}|| + ||H_{\text{sps}}||) \cdot |(H_{\text{lmmse}} - H_{\text{sps}})| \end{aligned} \tag{A.13}$$

where $|R_z|$ is the weak or the Hilbert–Schmidt norm of R_z defined by [21, eq. (2.13)]

$$\begin{aligned} |R_z| &= \left(\frac{1}{K} \text{tr}(R_z^{\#} R_z)\right)^{1/2} \\ &= \left(\frac{1}{K} \sum_{l,n=0}^{K-1} |r_z(l,n)|^2\right)^{1/2} \end{aligned} \tag{A.14}$$

and $||R_z||$ is the strong norm of R_z defined by

$$||R_z|| = \lambda_{\max}(R_z) \tag{A.15}$$

where $\lambda_{\max}(R_z)$ is the maximal eigenvalue of R_z. The first inequality in (A.13) results from the Schwarz inequality. Specifically, for $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$, $i, j = 1, \ldots, K$, let a_i be the ith row of A and b_j be the jth column of B. Then, applying twice the Schwarz inequality, we obtain

$$\begin{aligned} \frac{1}{K} \text{tr}\{AB\} &= \frac{1}{K} \sum_{i=1}^{K} a_i b_i \\ &\le \frac{1}{K} \sum_{i=1}^{K} ||a_i|| \cdot ||b_i|| \\ &\le \frac{1}{K} \left(\sum_{i=1}^{K} ||a_i||^2\right)^{1/2} \left(\sum_{i=1}^{K} ||b_i||^2\right)^{1/2} \\ &= \left(\frac{1}{K} \sum_{i,j=1}^{K} |a_{ij}|^2\right)^{1/2} \left(\frac{1}{K} \sum_{i,j=1}^{K} |b_{ij}|^2\right)^{1/2} \\ &= |A| \cdot |B|. \end{aligned} \tag{A.16}$$

The second inequality in (A.13) results from $|R_z| \le ||R_z||$ [21, eq. (2.15)] and from [21, eq. (2.18)]. The third inequality in (A.13) results from the triangular inequality for the strong norm [21, eq. (2.16)].

We show that under Assumptions 1–3, $||R_z||$ is bounded for all K, and H_{lmmse} and H_{sps} are asymptotically equivalent [21, eq. (2.20)]. This equivalence is denoted by

$$H_{\text{lmmse}} \sim H_{\text{sps}} \tag{A.17}$$

and it implies that

$$||H_{\text{lmmse}}||, ||H_{\text{sps}}|| \le M_H < \infty \tag{A.18}$$

$$\lim_{K \to \infty} |H_{\text{lmmse}} - H_{\text{sps}}| = 0. \tag{A.19}$$

Hence, the proof follows from (A.13).

From Assumptions 1–3

$$R_z = R_y + R_w \tag{A.20}$$

and this matrix is Toeplitz. The power spectral density associated with R_z is given by

$$f_z(\theta) = f_y(\theta) + f_w(\theta) \tag{A.21}$$

and it satisfies

$$0 < m_{f_w} \leq f_z(\theta) \leq M_{f_y} + M_{f_w} \triangleq M_{f_z} < \infty. \tag{A.22}$$

Since the eigenvalues of R_z are smaller than or equal to M_{f_z} [21, lemma 4.1], we conclude that

$$||R_z|| \leq M_{f_z} < \infty. \tag{A.23}$$

To prove that $H_{\text{lmmse}} \sim H_{\text{sps}}$, recall that

$$H_{\text{lmmse}} = R_y R_z^{-1} \tag{A.24}$$

$$H_{\text{sps}} = \frac{1}{K} D G_{\text{sps}} D^{\#}. \tag{A.25}$$

Since H_{sps} is circulant, and R_y and R_z are Toeplitz, proving that $H_{\text{lmmse}} \sim H_{\text{sps}}$ is a straightforward application of the results from [21].

Let $C(\hat{f}_y)$ be a circulant matrix with top row elements given by

$$\hat{c}_y(m) = \frac{1}{K}\sum_{k=0}^{K-1} \hat{f}_y(k) e^{j\frac{2\pi}{K}km}, \quad m = 0, \ldots, K-1. \tag{A.26}$$

We now show that

$$R_y \sim C(\hat{f}_y). \tag{A.27}$$

Since $f_y(\theta)$ is bounded, $||R_y||$ is finite. Furthermore

$$\begin{aligned}\hat{f}_y(k) &= \frac{1}{K} E\{|(D^{\#}y)_k|^2\} \\ &= \frac{1}{K}(D^{\#} R_y D)_{k,k} \\ &= \frac{1}{K}\sum_{l,n=0}^{K-1} r_y(l-n) e^{-j\frac{2\pi}{K}k(l-n)} \\ &= \sum_{m=-(K-1)}^{K-1}\left(1 - \frac{|m|}{K}\right) r_y(m) e^{-j\frac{2\pi}{K}km} \qquad \text{(A.28)} \\ &= \int_{-\pi}^{\pi} f_y(\phi) F_K\left(\frac{2\pi}{K}k - \phi\right)\frac{d\phi}{2\pi} \qquad \text{(A.29)}\end{aligned}$$

where

$$F_K(\theta) \triangleq \frac{1}{K}\frac{\sin^2(\theta K/2)}{\sin^2(\theta/2)} \tag{A.30}$$

is the Fejer kernel [33]. Since

$$\int_{-\pi}^{\pi} F_K(\theta)\frac{d\theta}{2\pi} = 1 \tag{A.31}$$

$\hat{f}_y(k)$ is bounded from above and below by the least upper bound and the greatest lower bound of $f_y(\theta)$, respectively. Hence, $||C(\hat{f}_y)||$ is bounded. It remains therefore to show that

$$\lim_{K\to\infty} |R_y - C(\hat{f}_y)| = 0. \tag{A.32}$$

From (A.26) and (A.28) we have []

$$\begin{aligned}\hat{c}_y(m) &= \frac{1}{K}\sum_{k=0}^{K-1}\sum_{n=-(K-1)}^{K-1}\left(1 - \frac{|n|}{K}\right) r_y(n) e^{-j\frac{2\pi}{K}k(n-m)} \\ &= \sum_{n=-(K-1)}^{K-1}\left(1 - \frac{|n|}{K}\right) r_y(n)\frac{1}{K}\sum_{k=0}^{K-1} e^{-j\frac{2\pi}{K}k(n-m)} \\ &= \begin{cases} r_y(0) & m = 0 \\ (1 - \frac{m}{K}) r_y(m) + (1 - \frac{K-m}{K}) r_y(m-K) & m > 0. \end{cases}\end{aligned}$$

Let

$$\tilde{c}_y(m) = \begin{cases} \hat{c}_y(m) & m \geq 0 \\ \hat{c}_y(K+m) & m < 0. \end{cases} \tag{A.33}$$

Then

$$\begin{aligned}&|R_y - C(\hat{f}_y)|^2 \\ &\quad = \frac{1}{K}\sum_{l,n=0}^{K-1} |r_y(l-n) - \tilde{c}_y(l-n)|^2 \\ &\quad = \sum_{m=-(K-1)}^{K-1}\left(1 - \frac{|m|}{K}\right)|r_y(m) - \tilde{c}_y(m)|^2 \\ &\quad = \sum_{m=1}^{K-1}\left(1 - \frac{m}{K}\right)\left[\frac{m}{K}(r_y(m) - r_y(m-K))\right]^2 \\ &\qquad + \sum_{m=-(K-1)}^{-1}\left(1 + \frac{m}{K}\right)\left[\frac{m}{K}(r_y(m) - r_y(m+K))\right]^2 \\ &\quad = 2\sum_{m=1}^{K-1}\left(1 - \frac{m}{K}\right)\left[\frac{m}{K}(r_y(m) - r_y(m-K))\right]^2 \\ &\quad \leq 4\sum_{m=1}^{K-1}\left(1 - \frac{m}{K}\right)\left(\frac{m}{K}\right)^2 (r_y^2(m) + r_y^2(m-K)) \\ &\quad \leq 4\sum_{m=1}^{K-1}\left(1 - \frac{m}{K}\right)\left(\frac{m}{K}\right) r_y^2(m) \\ &\quad = 4\sum_{m=1}^{N}\left(1 - \frac{m}{K}\right)\left(\frac{m}{K}\right) r_y^2(m) \\ &\qquad + 4\sum_{m=N+1}^{K-1}\left(1 - \frac{m}{K}\right)\left(\frac{m}{K}\right) r_y^2(m) \\ &\quad \leq 4\left(\frac{N}{K}\right)\sum_{m=1}^{N} r_y^2(m) + 4\sum_{m=N+1}^{K-1} r_y^2(m) \qquad \text{(A.34)}\end{aligned}$$

where the first inequality results from

$$(a-b)^2 \leq 2(a^2 + b^2) \tag{A.35}$$

the second inequality results from the fact that we extended the upper limit of the summation to $K-1$ from $(K-1)/2$ if K is odd and from $K/2$ if K is even, and $N < K$. From Assumption 2

$$\lim_{K\to\infty} \sum_{m=-(K-1)}^{K-1} |r_y(m)|^2 < \infty. \tag{A.36}$$

Hence, for every $\epsilon > 0$ there exists N such that

$$\sum_{m=N+1}^{\infty} |r_y(m)|^2 < \epsilon. \tag{A.37}$$

For this N, letting $K \to \infty$ in (A.34) yields

$$\lim_{K\to\infty} |R_y - C(\hat{f}_y)|^2 \leq 4\epsilon. \tag{A.38}$$

Since ϵ can be arbitrarily small

$$\lim_{K\to\infty} |R_y - C(\hat{f}_y)|^2 = 0. \tag{A.39}$$

We have shown that $R_y \sim C(\hat{f}_y)$. Since R_y and R_w satisfy similar assumptions, $R_w \sim C(\hat{f}_w)$. Similarly

$$R_z \sim C(\hat{f}_y + \hat{f}_w). \tag{A.40}$$

We now show that

$$R_z^{-1} \sim C\left(\frac{1}{\hat{f}_y + \hat{f}_w}\right). \tag{A.41}$$

This will complete the proof since from [21, theorems 2.1, 3]

$$H_{\text{lmmse}} = R_y R_z^{-1} \sim C\left(\frac{\hat{f}_y}{\hat{f}_y + \hat{f}_w}\right) = H_{\text{sps}}. \tag{A.42}$$

From [21, lemma 4.1]

$$||R_z^{-1}|| = \frac{1}{\lambda_{\min}(R_z)} \leq \frac{1}{m_{f_w}} < \infty \tag{A.43}$$

where $\lambda_{\min}(R_z)$ denotes the smallest eigenvalue of R_z. Similarly, applying the argument of (A.29) to the noise process and using (A.10), we conclude that

$$\begin{aligned} ||C^{-1}(\hat{f}_y + \hat{f}_w)|| &= ||C(1/(\hat{f}_y + \hat{f}_w)|| \\ &\leq \frac{1}{\inf_k(\hat{f}_w(k))} \\ &\leq \frac{1}{m_{f_w}} < \infty. \end{aligned} \tag{A.44}$$

Hence, the hypotheses of [21, theorem 2.1] are satisfied and (A.41) holds.

The proof cannot be generalized to show convergence of $\hat{y}_{\text{TDC}} - \hat{y}_{\text{sps}}$ to zero since the multiplicative noise factor μ can be infinite when $\alpha = 0$. In this case $||R_z||$ is unbounded. For any other fixed finite value of μ used by both estimators, the proof can trivially be extended to show convergence in the mean of $\hat{y}_{\text{TDC}} - \hat{y}_{\text{sps}}$ to zero.

Acknowledgment

The authors wish to thank Dr. N. Merhav for critical reading of the manuscript and his useful comments. They also thank Dr. P. Kroon for his help in evaluating the algorithm, and Mr. D. Talkin for his help in preparing the spectrograms. The individuals who participated in the listening tests are greatly acknowledged.

References

[1] J. S. Lim, Ed., *Speech Enhancement.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[3] J. Makhoul *et al.*, *Removal of Noise From Noise-Degraded Speech Signals*, Panel on removal of noise from a speech/noise National Research Council. Washington, DC: National Academy, 1989.

[4] D. O'Shaughnessy, "Enhancing speech degraded by additive noise or interfering speakers," *IEEE Commun. Mag.*, pp. 46–52, Feb. 1989.

[5] S. F. Boll, "Speech enhancement in the 1980's: Noise suppression with pattern matching," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, Eds.). New York: Marcel Dekker, 1992.

[6] Y. Ephraim, "Statistical model based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[7] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 512–530, Oct. 1979.

[8] J. L. Flanagan, "Parametric coding of speech spectra," *J. Acoust. Soc. Amer.*, vol. 68, no. 2, pp. 412–419, Aug., 1980.

[9] T. F. Quatieri and R. J. McAulay, "Phase coherence in speech reconstruction for enhancement and coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1989, pp. 207–210.

[10] ——, "Noise reduction using a soft-decision sine-wave vector quantizer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 821–824.

[11] H. L. Van Trees, *Detection, Estimation and Modulation Theory*, (Part I). New York: Wiley, 1968.

[12] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis.* New York: Addison-Wesley, 1990.

[13] S. Haykin, *Adaptive Filter Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1991.

[14] A. Feit, "Intelligibility enhancement of noisy speech signals," M.Sc. thesis, Dept. Elect. Eng., Technion-Israel Institute of Technology, July 1973. (in Hebrew).

[15] M. R. Weiss, E. Aschkenasy and T. W. Parsons, "Processing speech signals to attenuate interference," in *Proc. IEEE Symp. Speech Recognition*, Pittsburgh, PA, Apr. 1974, pp. 292–293.

[16] J. S. Lim, "Evaluation of correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 5, pp. 471–472, Oct. 1978.

[17] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

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

[19] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[20] M. M. Sondhi, C. E. Schmidt, and L. R. Rabiner, "Improving the quality of a noisy speech signal," *Bell Syst. Tech. J.*, vol. 60, no. 8, pp. 1847–1859, Oct. 1981.

[21] R. M. Gray, "Toeplitz and Circulant Matrices: II," Stanford Electron. Lab., Tech. Rep. 6504-1, Apr. 1977.

[22] R. O. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation," Ph.D dissertation, Stanford Univ., Stanford, CA, 1981.

[23] N. Tishby, "A dynamical systems approach to speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 365–368.

[24] B. Townshend, "Nonlinear prediction of speech signals," in SFI Studies in the Sciences of Complexity, vol. 13, *Nonlinear Modeling and Forecasting* (M. Casdagli and S. Eubank, Eds.). Reading, MA: Addison-Wesley, 1991.

[25] Y. Bresler and A. Macovski, "Exact maximum likelihood parameter estimation of superimposed exponential signals in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, no. 5, pp. 1081–1089, Oct. 1986.

[26] D. G. Luenberger, *Linear and Nonlinear Programming.* Reading, MA: Addison-Wesley, 1984.

[27] P. Lancaster and M. Tismenetsky, *The Theory of Matrices.* New York: Academic, 1985, 2nd ed.

[28] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternate projection," *IEEE Trans. Acoust., Speech Signal Processing*, vol. 36, no. 10, pp. 1553–1560, Oct. 1988.

[29] G. S. Kang and L. J. Fransen, "Quality improvement of LPC processed noisy speech by using spectral subtraction," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, no. 6, pp. 939–942, June 1989.

[30] J. D. Johnston and K. Brandenburg, "Wide-band coding-perceptual considerations for speech and music," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, Eds.). New York: Marcel Dekker, 1992.

[31] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1993, pp. II-355–II-358.

[32] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[33] M. B. Priestley, *Spectral Analysis and Time Series.* London: Academic, 1989.

[34] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[35] ——, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.

[36] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[37] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.

[38] N. Merhav, M. Gutman, and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1014–1019, Sept. 1989.

[39] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1109–1114, Sept. 1985.

[40] Q. Wu and D. R. Fuhrmann, "A parametric method for determining the number of signals in narrow-band direction finding," *IEEE Trans. Signal Processing*, vol. 39, pp. 1848–1857, Aug. 1991.

[41] P. Yip and K. R. Rao, *Discrete Cosine Transform: Algorithms, Advantages, Applications.* Boston, MA: Academic, 1990.

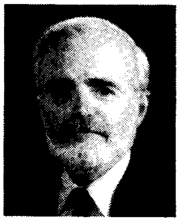
[42] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 845–848.

[43] M. Kadirkamanathan and A. P. Varga, "Simultaneous model re-estimation from contaminated data by composed hidden Markov modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1991, pp. 897–900.

[44] M. J. F. Gales and S. Young, "An improved approach to the hidden Markov model decomposition of speech and noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1992, pp. I-233–I-236.

Yariv Ephraim (S'82–M'84–SM'90–F'94) received the D.Sc degree in electrical engineering from the Technion, Israel Institute of Technology, Haifa, Israel, in 1984.

From 1984 to 1985, he was a Rothschild Post-Doctural Fellow at the Information Systems Laboratory, Stanford University, Stanford, CA. From 1985 to 1993, he was a member of Technical Staff at AT&T Bell Laboratories, Murray Hill, NJ. Since 1991, he has been at George Mason University, Fairfax, VA, where he is currently an Associate Professor of Electrical and Computer Engineering.



Harry L. Van Trees (M'57–SM'73–F'74–LF'94) received the B.Sc. degree from the U.S. Military Academy at West Point in 1952 and the Sc.D. degree from the Massachusetts Institute of Technology (M.I.T.), Cambridge, in 1961.

From 1961 to 1975, he was with the Electrical Engineering Department at M.I.T., becoming a Full Professor in 1969. During this period, he was active in graduate course development and was the leader of a research group working in detection and estimation theory and radar/sonar theory. Since 1988, he has been with George Mason University, Fairfax, VA, where he is currently a Distinguished Professor of Information Technology, Electrical and Systems Engineering. In June of 1989, he founded the Center of Excellence in Command, Control, Communications, and Intelligence, which now has 18 faculty members associated with it. The Center has a research program which includes work in sensing and data fusion, command support systems, communications, modeling and simulation, C3 architectures, and information systems. He has served as Chief Scientist of the Defense Communication Agency and Chief Scientist of the U.S. Air Force. He was Principle Deputy Assistant Secretary of Defense (C3I). He served as an Executive Vice President and President of M/A-COM Goverment Systems Division. He is the author of a three-volume set of books on detection, estimation, and modulation theory. These books contain a unified approach to communications, radar, sonar, and seismic applications. The first volume is a classic in its field and is used in graduate schools throughout the world; it is currently in its 27th printing.

Dr. Van Trees has received the Presidential Award for Meritorious Executive, the Distinguished Civilian Service Awards, and the AFCEA Gold Medal for Engineering.