Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator

YARIV EPHRAIM, STUDENT MEMBER, IEEE, AND DAVID MALAH, SENIOR MEMBER, IEEE

Abstract—This paper focuses on the class of speech enhancement systems which capitalize on the major importance of the *short-time spectral amplitude* (*STSA*) of the speech signal in its perception. A system which utilizes a minimum mean-square error (MMSE) STSA estimator is proposed and then compared with other widely used systems which are based on Wiener filtering and the "spectral subtraction" algorithm.

In this paper we derive the MMSE STSA estimator, based on modeling speech and noise *spectral components* as statistically independent Gaussian random variables. We analyze the performance of the proposed STSA estimator and compare it with a STSA estimator derived from the Wiener estimator. We also examine the MMSE STSA estimator under uncertainty of signal presence in the noisy observations.

In constructing the enhanced signal, the MMSE STSA estimator is combined with the complex exponential of the noisy phase. It is shown here that the latter is the MMSE estimator of the complex exponential of the original phase, which does not affect the STSA estimation.

The proposed approach results in a significant reduction of the noise, and provides enhanced speech with *colorless* residual noise. The complexity of the proposed algorithm is approximately that of other systems in the discussed class.

I. Introduction

THE problem of enhancing speech degraded by uncorrelated additive noise, when the noisy speech alone is available, has recently received much attention. This is due to the many potential applications a successful speech enhancement system can have, and because of the available technology which enables the implementation of such intricate algorithms. A comprehensive review of the various speech enhancement systems which emerged in recent years, and their classification according to the aspects of speech production and perception which they capitalize on, can be found in [1].

We focus here on the class of speech enhancement systems which capitalize on the major importance of the short-time spectral amplitude (STSA) of the speech signal in its perception [1], [2]. In these systems the STSA of the speech signal is estimated, and combined with the short-time phase of the degraded speech, for constructing the enhanced signal. The "spectral subtraction" algorithm and Wiener filtering are well-known examples [1], [3]. In the "spectral subtraction" algorithm, the STSA is estimated as the square root of the maximum likelihood (ML) estimator of each signal spectral component variance [3]. In systems which are based on Wiener filtering, the STSA estimator is obtained as the modulus of the optimal minimum mean-square error (MMSE) estimator of each signal spectral component [1], [3]. These two STSA estimators were derived under Gaussian assumption.

Since the "spectral subtraction" STSA estimator is derived from an optimal (in the ML sense) variance estimator, and the Wiener STSA estimator is derived from the optimal MMSE signal spectral estimator, both are not optimal *spectral amplitude* estimators under the assumed statistical model and criterion. This observation led us to look for an optimal STSA estimator which is derived directly from the noisy observations. We concentrate here on the derivation of an MMSE STSA estimator, and on its application in a speech enhancement system.

The STSA estimation problem formulated here is that of estimating the modulus of each complex Fourier expansion coefficient[1] of the speech signal in a given analysis frame from the noisy speech in that frame. This formulation is motivated by the fact that the Fourier expansion coefficients of a given signal segment are samples of its Fourier transform, and by the close relation between the Fourier series expansion and the discrete Fourier transform. The latter relation enables an efficient implementation of the resulting algorithm by utilizing the FFT algorithm.

To derive the MMSE STSA estimator, the *a priori* probability distribution of the speech and noise Fourier expansion coefficients should be known. Since in practice they are unknown, one can think of measuring each probability distribution or, alternatively, assume a reasonable statistical model.

In the discussed problem, the speech and possibly also the noise are neither stationary nor ergodic processes. This fact excludes the convenient possibility of obtaining the above probability distributions by examining the long time behavior of each process. Hence, the only way which can be used is to examine independent sample functions belonging to the ensemble of each process, e.g., for the speech process these sample functions can be obtained from different speakers. However, since the probability distributions we are dealing with are time-varying (due to the nonstationarity of the processes), their measurement and characterization by the above way is complicated, and the entire procedure seems to be impracticable.

For the above reasons, a statistical model is used here. This model utilizes asymptotic statistical properties of the Fourier expansion coefficients (see, e.g., [5]). Specifically, we assume

Manuscript received June 9, 1983; revised December 5, 1983 and May 14, 1984. This work was supported by the Technion V.P.R. Fund—the Natkin fund for Electrical Engineering Research.

Y. Ephraim was with the Technion—Israel Institute of Technology, Haifa, Israel. He is now with the Information Systems Laboratory, Stanford University, Stanford, CA 94305.

D. Malah is with the Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel.

[1] The complex Fourier expansion coefficients are also referred here as spectral components.

0096-3518/84/1200-1109$01.00

that the Fourier expansion coefficients of each process can be modeled as statistically independent Gaussian random variables. The mean of each coefficient is assumed to be zero, since the processes involved here are assumed to have zero mean. The variance of each speech Fourier expansion coefficient is time-varying, due to speech nonstationarity.

This Gaussian statistical model is motivated by the central limit theorem, as each Fourier expansion coefficient is, after all, a weighted sum (or integral) of random variables resulting from the process samples. The fact that a central limit theorem exists (under mild conditions) also for strongly mixing processes (i.e., in which sufficiently separated samples are weakly dependent) [4], [5] encourages the use of the Gaussian model in the discussed problem.

The statistical independence assumption in the Gaussian model is actually equivalent to the assumption that the Fourier expansion coefficients are uncorrelated. This latter assumption is commonly justified by the fact that the normalized correlation between different Fourier expansion coefficients approaches zero as the analysis frame length approaches infinity [6].

In our problem, the analysis frame length T cannot be too large due to the quasi-stationarity of the speech signal. Its typical value is 20–40 ms. This may cause the Fourier expansion coefficients to be correlated to a certain degree. Nevertheless, we continue with this statistical independence assumption in order to simplify the resulting algorithm. The case of statistically dependent expansion coefficients is now under investigation. In practice, an appropriate window (e.g., Hanning) is applied to the noisy process, which reduces the correlation between widely separated spectral components, at the expense of increasing the correlation between adjacent spectral components. This is a consequence of the wider main lobe but lower sidelobes of a window function, in comparison to the rectangular window.

It is worthwhile noting that several efforts have been made in the past for measuring the probability distribution of a speech spectral component. It turns out that the answer to the question of what is the correct distribution is controversial, since different investigators arrived at different distributions. For example, Zelinski and Noll [7], [8] observed that a gain normalized cosine transform coefficient (which is closely related to the real part of the Fourier transform coefficient) is approximately Gaussian distributed. On the other hand, Porter and Boll [9] claim that the amplitude of a gain normalized Fourier transform coefficient is gamma-like distributed. However, since in the latter measurements the long-time behavior of the speech signals was examined, this gamma distribution reflects the relative frequency of amplitude appearance rather than the probability density function of the STSA.

In conclusion of the above discussion concerning the statistical model of the speech spectral components, we note that since the true statistical model seems to be inaccessible, the validity of the proposed one can be judged *a posteriori* on the basis of the results obtained here. In addition, the optimality of the estimators derived here is of course connected with the assumed statistical model.

In this paper we derive the MMSE STSA estimator based on the above statistical model, and compare its performance with that of the Wiener STSA estimator. This comparison is of interest since the Wiener estimator is a widely used STSA estimator, which is also derived under the same statistical model.

The Gaussian statistical model assumed above does not take into account the fact that the speech signal is not surely present in the noisy observations. This model results in a Rayleigh distribution for the amplitude of each signal spectral component, which assumes insignificant probability for low amplitude realizations. Therefore, this model can lead to less suppression of the noise than other amplitude distribution models (e.g., gamma) which assume high probability for low amplitude realizations. However, using a statistical model of the latter type can lead to a worse amplitude estimation when the signal is present in the noisy observations.

One useful approach to resolve this problem is to derive an MMSE STSA estimator which takes into account the uncertainty of speech presence in the noisy observations [3], [10], [11]. Such an estimator can be derived on the basis of the above Gaussian statistical model, and by assuming that the signal is present in the noisy observations with probability $p < 1$ only. The parameter p supplies a useful degree of freedom which enables one to compromise between noise suppression and signal distortion. This is of course an advantage in comparison to the use of a gamma type statistical model.

The above approach is applied in this paper, and the resulting STSA estimator is compared with the McAulay and Malpass [3] estimator in enhancing speech. The latter estimator is an appropriately modified ML STSA estimator, which assumes that the signal is present in each noisy spectral component with a probability of $p = 0.5$.

In this paper we also examine the estimation of the complex exponential of the phase of a given signal spectral component. The complex exponential estimator is used in conjunction with the MMSE STSA estimator for constructing the enhanced signal. We derive here the MMSE complex exponential estimator and discuss its effect on the STSA estimation. We show that the complex exponential of the noisy phase is the MMSE complex exponential estimator which does not affect the STSA estimation. Therefore, the noisy phase is used in the proposed system.

The paper is organized as follows. In Section II and Appendix A we derive the MMSE STSA estimator and compare its performance with that of the Wiener STSA estimator. In Section III we extend the MMSE STSA estimator and derive it under uncertainty of signal presence in the noisy spectral components. In Section IV and Appendix B we discuss the MMSE estimation of the complex exponential of the phase. In Section V we discuss the problem of estimating the *a priori* SNR of a spectral component, which is a parameter of the STSA estimator. In Section VI we describe the proposed speech enhancement system and compare it with the other widely used systems mentioned above. In Section VII we summarize the paper and draw conclusions.

II. MMSE Short-Time Spectral Amplitude Estimator

In this section we derive the MMSE STSA estimator under the statistical model assumed in Section I. We also analyze its performance and examine its sensitivity to the *a priori* SNR, which was found to be a key parameter. This performance

and sensitivity analysis is also done for the Wiener STSA estimator, and the two estimators are compared on this basis.

Derivation of Amplitude Estimator

Let $x(t)$ and $d(t)$ denote the speech and the noise processes, respectively. The observed signal $y(t)$ is given by

$$y(t) = x(t) + d(t), \qquad 0 \leq t \leq T \tag{1}$$

where, without loss of generality, we let the observation interval be $[0, T]$. Let $X_k \triangleq A_k \exp(j\alpha_k)$, D_k, and $Y_k \triangleq R_k \exp(j\vartheta_k)$ denote the kth spectral component of the signal $x(t)$, the noise $d(t)$, and the noisy observations $y(t)$, respectively, in the analysis interval $[0, T]$. Y_k (and similarly X_k and D_k) is given by

$$Y_k = \frac{1}{T}\int_0^T y(t) \exp\left(-j\frac{2\pi}{T}kt\right) dt$$

$$k = 0, \pm 1, \pm 2, \cdots. \tag{2}$$

Based on the formulation of the estimation problem given in the previous section, our task is to estimate the modulus A_k, from the degraded signal $\{y(t), 0 \leq t \leq T\}$.

Toward this end, we note that the signal $\{y(t), 0 \leq t \leq T\}$ can be written in terms of its spectral components Y_k by [6]

$$y(t) = \underset{K\to\infty}{\text{l.i.m.}} \sum_{k=-K}^{K} Y_k \exp\left(j\frac{2\pi}{T}kt\right) \qquad 0 \leq t \leq T \tag{3}$$

where l.i.m. means limit in the mean. Moreover, on the basis of the Gaussian statistical model for the spectral components assumed here, the series (3) converges almost surely to $y(t)$ for every $t \in [0, T]$. Therefore, it can be shown that $\{y(t), 0 \leq t \leq T\}$ and $\{Y_0, Y_1, \cdots\}$ bear the same information (up to events whose probability is zero) [12, Appendix D]. This means that the MMSE estimation problem can be reduced to be that of estimating A_k from the infinite countable set of observations $\{Y_0, Y_1, \cdots\}$. In addition, since the spectral components are assumed to be statistically independent, the MMSE amplitude estimator can be derived from Y_k only. In conclusion, the MMSE estimator $\hat{A}_k$ of A_k is obtained as follows:

$$\begin{aligned}\hat{A}_k &= E\{A_k \mid y(t), \quad 0 \leq t \leq T\} \\ &= E\{A_k \mid Y_0, Y_1, \cdots\} \\ &= E\{A_k \mid Y_k\} \\ &= \frac{\int_0^\infty \int_0^{2\pi} a_k p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}{\int_0^\infty \int_0^{2\pi} p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k da_k}\end{aligned} \tag{4}$$

where $E\{\cdot\}$ denotes the expectation operator, and $p(\cdot)$ denotes a probability density function (PDF).

Under the assumed statistical model, $p(Y_k \mid a_k, \alpha_k)$ and $p(a_k, \alpha_k)$ are given by

$$p(Y_k \mid a_k, \alpha_k) = \frac{1}{\pi\lambda_d(k)} \exp\left\{-\frac{1}{\lambda_d(k)} \left|Y_k - a_k e^{j\alpha_k}\right|^2\right\} \tag{5}$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} \exp\left\{-\frac{a_k^2}{\lambda_x(k)}\right\} \tag{6}$$

where $\lambda_x(k) \triangleq E\{|X_k|^2\}$, and $\lambda_d(k) \triangleq E\{|D_k|^2\}$, are the variances of the kth spectral component of the speech and the noise, respectively. Substituting (5) and (6) into (4) gives (see Appendix A)

$$\begin{aligned}\hat{A}_k &= \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} M(-0.5; 1; -v_k) R_k \\ &= \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} \exp\left(-\frac{v_k}{2}\right) \\ &\quad \left[(1 + v_k) I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] R_k.\end{aligned} \tag{7}$$

$\Gamma(\cdot)$ denotes the gamma function, with $\Gamma(1.5) = \sqrt{\pi}/2$; $M(a; c; x)$ is the confluent hypergeometric function [4, eq. A.1.14]; $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively. v_k is defined by

$$v_k \triangleq \frac{\xi_k}{1 + \xi_k} \gamma_k \tag{8}$$

where ξ_k and γ_k are defined by

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \tag{9}$$

$$\gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)} \tag{10}$$

ξ_k and γ_k are interpreted (after McAulay and Malpass [3]) as the *a priori* and *a posteriori* signal-to-noise ratios (SNR), respectively.

A similar expression to (7) was obtained in [14], [15] when the amplitude of a Gaussian sinusoidal random process buried in Gaussian noise is optimally estimated.

It is of interest to examine the asymptotic behavior of $\hat{A}_k$ at high SNR, i.e., at $\xi_k >> 1$. By considering the exponential distribution of v_k [i.e., $p(v_k) = 1/\xi_k \exp(-v_k/\xi_k)$], it is easy to see that $\xi_k >> 1$ implies $v_k >> 1$ with high probability. Therefore, to examine $\hat{A}_k$ at $\xi_k >> 1$, we substitute the following approximation of the confluent hypergeometric function [4, eq. A.1.16b] in (7).

$$M(-0.5; 1; -v_k) \cong \frac{\sqrt{v_k}}{\Gamma(1.5)} \qquad v_k >> 1. \tag{11}$$

We get

$$\begin{aligned}\hat{A}_k &\cong \frac{\xi_k}{1 + \xi_k} R_k \qquad \text{high SNR} \\ &\triangleq A_k^w.\end{aligned} \tag{12}$$

Since we finally estimate the spectral component $X_k = A_k \exp(j\alpha_k)$ by $\hat{X}_k = \hat{A}_k \exp(j\vartheta_k)$, where $\exp(j\vartheta_k)$ is the complex exponential of the noisy phase (see Section IV), we get from (12) the following approximation for the kth signal spectral component estimator:

$$\begin{aligned}\hat{X}_k &\cong \frac{\xi_k}{1 + \xi_k} Y_k \qquad \text{high SNR} \\ &\triangleq X_k^w.\end{aligned} \tag{13}$$

This estimator is in fact the MMSE estimator of the kth signal spectral component, i.e., the Wiener estimator. For this reason, (12) is referred to as a Wiener amplitude estimator.

It is useful to consider the amplitude estimator $\hat{A}_k$ in (7) as being obtained from R_k by a multiplicative nonlinear gain function which is defined by

$$G_{\mathrm{MMSE}}(\xi_k, \gamma_k) \triangleq \frac{\hat{A}_k}{R_k}. \tag{14}$$

From (7) we see that this gain function depends only on the *a priori* and the *a posteriori* SNR, ξ_k and γ_k, respectively. Several gain curves which result from (7) and (14) are shown in Fig. 1. $\gamma_k - 1$ in Fig. 1 is interpreted as the "instantaneous SNR," since $\gamma_k \triangleq R_k^2/\lambda_d(k)$, and R_k is the modulus of the signal plus noise resultant spectral component.

The gain curves in Fig. 1 show an increase in gain as the instantaneous SNR $\gamma_k - 1$ decreases, while the *a priori* SNR ξ_k is kept constant. This behavior is explained below on the basis of the fact that the MMSE estimator compromises between what it knows from the *a priori* information and what it learns from the noisy data.

Let ξ_k result from some fixed values of $\lambda_x(k)$ and $\lambda_d(k)$. The fixed value of $\lambda_x(k)$ determines the most probable realizations of A_k, which are considered by the MMSE estimator. This is due to the fact that $\lambda_x(k)$ is the only parameter of $p(a_k)$ [see (6)]. On the other hand, the fixed value of $\lambda_d(k)$ makes γ_k to be proportional to R_k, since $\gamma_k \triangleq R_k^2/\lambda_d(k)$. Therefore, as ξ_k is fixed and γ_k decreases, the estimator should compromise between the most probable realizations of A_k and the decreasing values of R_k. Since A_k is estimated by $\hat{A}_k = G_{\mathrm{MMSE}}(\xi_k, \gamma_k)R_k$, this can be done by increasing $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$.

Fig. 1 also shows several gain curves corresponding to the Wiener gain function which results from the amplitude estimator (12). This gain function is given by

$$G_w(\xi_k, \gamma_k) = \frac{\xi_k}{1 + \xi_k} \tag{15}$$

and it is independent of γ_k. The convergence of the MMSE and the Wiener amplitude estimators at high SNR is clearly demonstrated in Fig. 1.

It is interesting to note that the same gain curves as those belonging to the gain function $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$ were obtained by a "vector spectral subtraction" amplitude estimation approach [16]. In this approach, the amplitude estimator is obtained from two mutually dependent MMSE estimators of the amplitude and the cosine of the phase error (i.e., the phase $\vartheta_k - \alpha_k$). Since an estimator of the cosine of the phase error is used for estimating the amplitude, this approach is interpreted as a "vector spectral subtraction" amplitude estimation. We conjecture that this coinciding of the gain curves is a consequence of the statistical independence assumption of the real and imaginary parts of each complex Fourier expansion coefficient, which results in the statistical independence of the amplitude and phase. This probably enables one to obtain the MMSE amplitude estimator, by cross coupling the two partial MMSE estimators, of the amplitude and the cosine of the phase error.

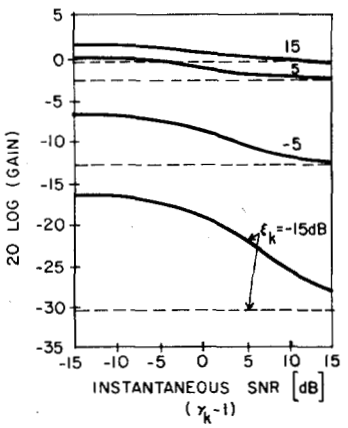


Fig. 1. Parametric gain curves describing (a) MMSE gain function $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$ defined by (7) and (14) (solid lines), and (b) Wiener gain function $G_w(\xi_k, \gamma_k)$ defined by (15) (dashed line).

Error Analysis and Sensitivity

The MMSE amplitude estimator (7) is derived under the implicit assumption that the *a priori* SNR ξ_k and the noise variance $\lambda_d(k)$ are known. However, in the speech enhancement problem discussed here, these parameters are unknown in advance, as the noisy speech alone is available. Therefore, they are replaced by their estimators in a practical system (see Section V). For this reason it is of interest to examine the sensitivity of the amplitude estimator to inaccuracy in these parameters.

We found that the *a priori* SNR is a key parameter in the discussed problem, rather than the noise variance which is easier to estimate. Therefore, we examine here the sensitivity of the MMSE amplitude estimator to the *a priori* SNR ξ_k only. In addition, for similar reasons, we are interested here especially in the sensitivity at low *a priori* SNR (i.e., $\xi_k << 1$).

We present here a sensitivity analysis which is based on the calculation of the mean-square error (MSE) and the bias associated with the amplitude estimator (7) when the *a priori* SNR ξ_k is perturbated. This sensitivity analysis provides also an error analysis, since the latter turns out to be a particular case of the former.

A similar problem to the above one arises in the Wiener amplitude estimator, which depends on the *a priori* SNR parameter [see (12)]. Since the Wiener estimator is widely used in speech enhancement systems, we give here a sensitivity analysis for this estimator as well, and compare it with the MMSE amplitude estimator.

Let ξ_k^* denote the nominal *a priori* SNR, and $\tilde{\xi}_k \triangleq \xi_k^* + \Delta\xi_k$ denote its perturbated version. The MMSE amplitude estimator which uses the perturbated ξ_k is obtained from (7), and is given by

$$\hat{A}_k = \Gamma(1.5)\frac{\sqrt{\tilde{v}_k}}{\gamma_k} M(-0.5; 1; -\tilde{v}_k)R_k \tag{16}$$

where $\tilde{v}_k$ is defined by

$$\tilde{v}_k = \frac{\tilde{\xi}_k}{1 + \tilde{\xi}_k}\gamma_k. \tag{17}$$

Similarly, the Wiener amplitude estimator with the perturbated ξ_k is obtained from (12), and is given by

$$A_k^w = \frac{\tilde{\xi}_k}{1+\tilde{\xi}_k} R_k. \tag{18}$$

To calculate the residual MSE in the amplitude estimation (16) for low *a priori* SNR values, it is most convenient to expand $M(a;c;x)$ in (16) by the following series [4, eq. A.1.14]:

$$M(a,c,x) = \sum_{r=0}^{\infty} \frac{(a)_r}{(c)_r} \frac{x^r}{r!}$$

$$= 1 + \frac{a}{c}\frac{x}{1!} + \frac{a(a+1)}{c(c+1)}\frac{x^2}{2!} + \cdots \tag{19}$$

where $(a)_r \triangleq a(a+1)\cdots(a+r-1)$, and $(a)_0 \triangleq 1$. By so doing, and using the fact that γ_k is exponentially distributed, i.e.,

$$p(\gamma_k) = \frac{1}{1+\xi_k^*} \exp\left(-\frac{\gamma_k}{1+\xi_k^*}\right) \qquad \gamma_k \geq 0, \tag{20}$$

we get the normalized residual MSE $\epsilon_{\mathrm{MMSE}}(\xi_k^*, \tilde{\xi}_k)$ by

$$\epsilon_{\mathrm{MMSE}}(\xi_k^*, \tilde{\xi}_k) \triangleq E\{[A_k - \hat{A}_k]^2\}/E\{[A_k - E(A_k)]^2\}$$

$$= \frac{1}{1-\pi/4}\left\{1 + \frac{\pi}{4}\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\frac{1}{\xi_k^*}\sum_{r,l=0}^{\infty} \cdot \frac{(-0.5)_r(-0.5)_l}{(1)_r(1)_l}\frac{1}{r!l!}\left(\frac{1+\xi_k^*}{1+\tilde{\xi}_k}\right)^{r+l} \cdot (-\tilde{\xi}_k)^{r+l}\Gamma(r+l+1) - 2\frac{\pi}{4}\left(\frac{\xi_k^*}{1+\xi_k^*}\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\right)^{1/2}\frac{1}{\xi_k^*}\sum_{r,l=0}^{\infty} \cdot \frac{(-0.5)_r(-0.5)_l}{(1)_r(1)_l}\frac{1}{r!l!}\left(\frac{1+\xi_k^*}{1+\tilde{\xi}_k}\right)^{l} \cdot (-\xi_k^*)^r(-\tilde{\xi}_k)^l\Gamma(r+l+1)\right\}. \tag{21}$$

It can be shown by using the Lebesgue monotonic convergence theorem and the Lebesgue dominated convergence theorem [17] that the commutation of the expectation and limit operations needed in the calculation of (21) are vaid for $\xi^* < 1$ and $\tilde{\xi} < (1-\xi^*)/2\xi^*$. Therefore, the resulting expression in (21) is also valid in that domain.

The normalized residual MSE $\epsilon_w(\xi_k^*, \tilde{\xi}_k)$ resulting in the Wiener amplitude estimation (18) can be calculated similarly, and is given by

$$\epsilon_w(\xi_k^*, \tilde{\xi}_k) \triangleq E\{[A_k - A_k^w]^2\}/E\{[A_k - E(A_k)]^2\}$$

$$= \frac{1}{1-\pi/4}\left\{1 + \left(\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\right)^2\left(\frac{1+\xi_k^*}{\xi_k^*}\right) - 2\Gamma(1.5)\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\frac{1}{(\xi_k^*)^{1/2}}\sum_{r=0}^{\infty}\frac{(-0.5)_r}{(1)_r}\frac{1}{r!} \cdot (-\xi_k^*)^r\Gamma(r+1.5)\right\} \tag{22}$$

which is valid for $\xi^* < 1$.

For low SNR the above expressions can also be used to calculate the nominal residual MSE, which corresponds to the MSE when the *a priori* SNR is known exactly. This can be done by substituting $\tilde{\xi}_k = \xi_k^*$ in (21) and (22).

For very low SNR values, $\epsilon_{\mathrm{MMSE}}(\xi_k^*, \tilde{\xi}_k)$ and $\epsilon_w(\xi_k^*, \tilde{\xi}_k)$ can be approximated by considering terms of up to third order only in the infinite sums of (21) and (22). Fig. 2 shows the residual MSE obtained in this way as a function of the nominal *a priori* SNR ξ_k^*, and for several values of $\Delta\xi_k/\xi_k^*$. A number of conclusions can be drawn now. First, note from (21) and (22) that the nominal normalized MSE in the MMSE estimation cannot be greater than unity, while in the Wiener amplitude estimation it can be as high as $1/(1-\pi/4)$. Second, both estimators seem to be insensitive to small perturbations in the nominal *a priori* SNR ξ_k^* value. Finally, it is interesting to note that both estimators are more sensitive to underestimates of the *a priori* SNR than to its overestimates. In addition, by using an overestimate of ξ_k^* in the Wiener amplitude estimation, the residual MSE decreases. This surprising fact can be explained by noting that the Wiener estimator is not an MMSE amplitude estimator under the assumed model and criterion. Therefore, using an erroneous value of ξ_k^* can either increase or decrease the MSE.

The operational conclusion of the above error analysis is that on the basis of an MSE criterion, it is more appropriate to use an overestimate of the *a priori* SNR than to use an underestimate of it. It is satisfying to note that a similar conclusion was drawn in [18] on a perceptual ground. In [18] it was found that when the speech spectral component variance is estimated (for the spectral subtraction algorithm purposes) by the "power-spectral subtraction" method (see Section V), then it is useful to use a "spectral floor" which masks the "musical noise." This "spectral floor" is a positive threshold value which is used as the estimated variance when the "power spectral subtraction" method results in an estimate which is lower than that threshold. Therefore, the "spectral floor" is an overestimate of the signal spectral component variance, and also of its *a priori* SNR.

We turn now to the calculation of the normalized bias of each estimator when the *a priori* SNR is perturbated. The normalized bias is defined here as the ratio between the expected value of the amplitude estimation error and the expected value of the amplitude.

The normalized bias $B_{\mathrm{MMSE}}(\xi_k^*, \tilde{\xi}_k)$ of the MMSE estimator at low *a priori* SNR is obtained by using (16), (19), and (20). It is equal to

$$B_{\mathrm{MMSE}}(\xi_k^*, \tilde{\xi}_k) \triangleq E\{A_k - \hat{A}_k\}/E\{A_k\}$$

$$= 1 - \left(\frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\frac{1}{\xi_k^*}\right)^{1/2}\sum_{r=0}^{\infty}\frac{(-0.5)_r}{(1)_r} \cdot \left(\frac{1+\xi_k^*}{1+\tilde{\xi}_k}\right)^r(-\tilde{\xi}_k)^r \tag{23}$$

and is valid for $\xi^*\tilde{\xi} < 1$. The normalized bias $B_w(\xi_k^*, \tilde{\xi}_k)$ of the Wiener amplitude estimator is easily obtained from (18), and is given by

$$B_w(\xi_k^*, \tilde{\xi}_k) \triangleq E\{A_k - A_k^w\}/E\{A_k\}$$

$$= 1 - \frac{\tilde{\xi}_k}{1+\tilde{\xi}_k}\left(\frac{1+\xi_k^*}{\xi_k^*}\right)^{1/2}. \tag{24}$$

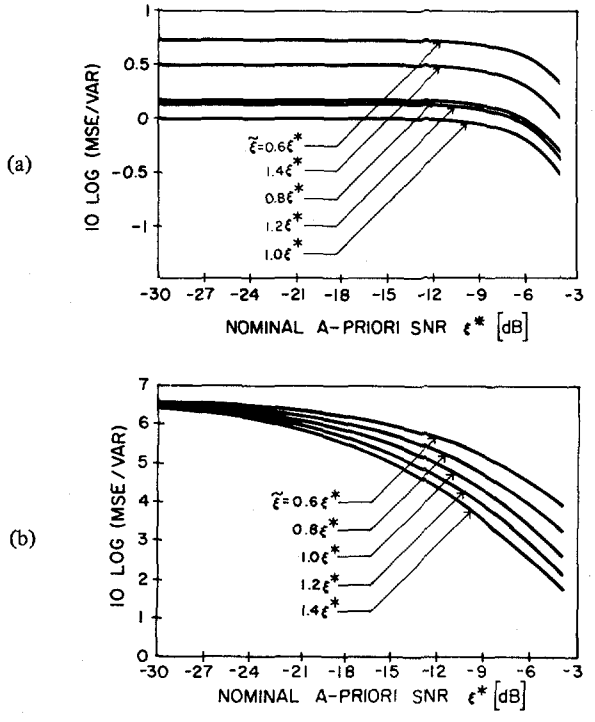


Fig. 2. Normalized MSE in amplitude estimations for perturbated values of the *a priori* SNR. (a) MMSE estimator (16). (b) Wiener estimator (18).

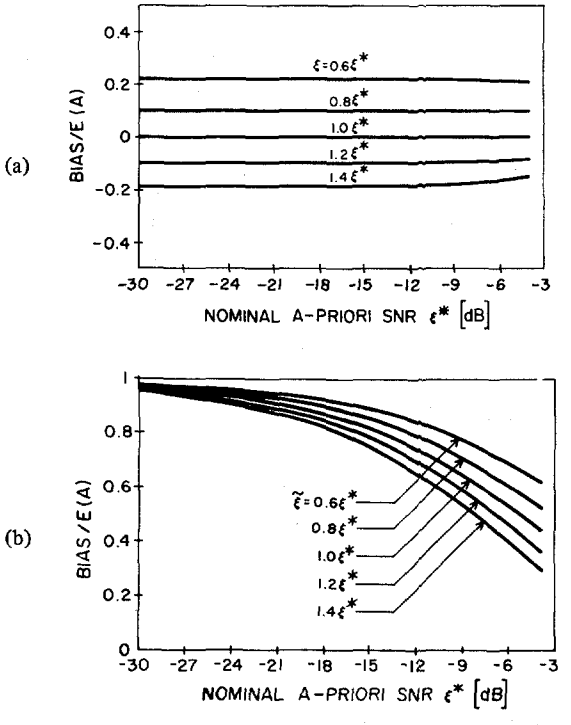


Fig. 3. Normalized bias of amplitude estimators for perturbated values of the *a priori* SNR. (a) MMSE estimator (16). (b) Wiener estimator (18).

Fig. 3 shows the bias of the MMSE and the Wiener amplitude estimators as a function of ξ_k^*: $B_{\mathrm{MMSE}}(\xi_k^*, \tilde{\xi}_k)$ in Fig. 3 is calculated by using terms of up to the third order in the infinite sum in (23).

III. MMSE Amplitude Estimator Under Uncertainty of Signal Presence

In this section we derive the MMSE amplitude estimator under the assumed Gaussian statistical model, and uncertainty of signal presence in the noisy observations. By so doing we extend the amplitude estimator derived in Section II, as will be clarified later.

Signal absence in the noisy observations $\{y(t), 0 \leq t \leq T\}$ is frequent, since speech signals contain large portions of silence. This absence of signal implies its absence in the noisy spectral components as well. However, it is also possible that the signal is present in the noisy observations, but appears with insignificant energy in some noisy spectral components, which are randomly determined. This is a typical situation when the analyzed speech is of voiced type, and the analysis is not synchronized with the pitch period.

The above discussion suggests two statistical models for speech absence in the noisy spectral components. In the first one, speech is assumed to be either present or absent, with given probabilities, in all of the noisy spectral components. The reasoning behind this model is that signal presence or absence should be the same in all of the noisy spectral components, since the analysis is done on a finite interval. In the second model which represents the other extreme, a statistically independent random appearance of the signal in the noisy spectral components is assumed. As is implied by the above discussion, this model is more appropriate for voiced speech signals when weak signal spectral components are considered as if they were absent.

These two models, and the resulting MMSE amplitude estimators based upon them, are examined in details in [19]. We found that the estimator whose derivation is based on the second model is especially successful in speech enhancement applications. Therefore, we present its derivation in this section.

The idea of utilizing the uncertainty of signal presence in the noisy spectral components for improving speech enhancement results was first proposed by McAulay and Malpass [3]. In their work they actually capitalize on the above second model of signal absence, and modify appropriately an ML amplitude estimator. In Section VI we compare the speech enhancement results of the McAulay and Malpass amplitude estimator with those of the one which we derive here.

Derivation of Amplitude Estimator Under Signal Presence Uncertainty

The MMSE estimator, which takes into account the uncertainty of signal presence in the noisy observations, was developed by Middleton and Esposito [10]. Based on our second model for signal absence, in which statistically independent random appearances of the signal in the noisy spectral components is assumed, and on the statistical independence of the spectral components assumed in Section I, this MMSE estimator is given by [19]

$$\hat{A}_k = \frac{\Lambda(Y_k, q_k)}{1 + \Lambda(Y_k, q_k)} E\{A_k \mid Y_k, H_k^1\} \tag{25}$$

where $\Lambda(Y_k, q_k)$ is the generalized likelihood ratio defined by

$$\Lambda(Y_k, q_k) = \mu_k \frac{p(Y_k \mid H_k^1)}{p(Y_k \mid H_k^0)} \tag{26}$$

with $\mu_k \triangleq (1 - q_k)/q_k$, and q_k is the probability of signal absence in the kth spectral component. H_k^0 and H_k^1 denote the two hypotheses of signal absence and presence, respectively, in the kth spectral component. $E\{A_k \mid Y_k, H_k^1\}$ is the MMSE amplitude estimator when the signal is surely present in the kth spectral component. This is in fact the estimator (7). Therefore, in order to derive the new amplitude estimator (25), we need to calculate the additional function $\Lambda(Y_k, q_k)$ only. This can be easily done by using the Gaussian statistical model assumed for the spectral components, or equivalently, by using (5) and (6). We get

$$\Lambda(Y_k, q_k) = \mu_k \frac{\exp(v_k)}{1 + \xi_k} \tag{27}$$

where ξ_k in (27) is now defined by

$$\xi_k \triangleq \frac{E\{A_k^2 \mid H_k^1\}}{\lambda_d(k)}. \tag{28}$$

This definition agrees with its previous definition in (9), since there the signal is implicitly assumed to be surely present in the noisy spectral components.

It is more convenient to make $\Lambda(Y_k, q_k)$ and the resulting amplitude estimator a function of $\eta_k \triangleq E\{A_k^2\}/\lambda_d(k)$ which is easier to estimate than ξ_k. η_k is related to ξ_k by

$$\begin{aligned}\eta_k &\triangleq \frac{E\{A_k^2\}}{\lambda_d(k)} \\ &= (1 - q_k)\frac{E\{A_k^2 \mid H_k^1\}}{\lambda_d(k)} \\ &= (1 - q_k)\,\xi_k. \end{aligned} \tag{29}$$

Thus, by considering $\Lambda(Y_k, q_k)$ in (27) as $\Lambda(\xi_k, \gamma_k, q_k)$, and using $E\{A_k \mid y_k, H_k^1\} = G_{\mathrm{MMSE}}(\xi_k, \gamma_k)R_k$, where $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$ is the gain function defined by (7) and (14), the amplitude estimator (25) can be written as

$$\begin{aligned}\hat{A}_k &= \frac{\Lambda(\xi_k, \gamma_k, q_k)}{1 + \Lambda(\xi_k, \gamma_k, q_k)} G_{\mathrm{MMSE}}(\xi_k, \gamma_k)R_k \Big|_{\xi_k = \eta_k/(1-q_k)} \\ &\triangleq G^D_{\mathrm{MMSE}}(\eta_k, \gamma_k, q_k)R_k. \end{aligned} \tag{30}$$

Note that if $q_k = 0$, then $\Lambda/(1 + \Lambda)$ in (30) equals unity, and also $\eta_k = \xi_k$. In this case $G^D_{\mathrm{MMSE}}(\eta_k, \gamma_k, q_k)$ turns out to be equal to $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$. Thus, the amplitude estimator (7) can be considered as a particular case of the amplitude estimator (30).

Several gain curves which result from $G^D_{\mathrm{MMSE}}(\eta_k, \gamma_k, q_k)$ in (30) are described in Fig. 4 for $q_k = 0.2$. It is interesting to compare these gain curves with those of $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$ which are depicted in Fig. 1. Especially it is interesting to see the different trend of the gain curves in each pair corresponding to the same value of the *a priori* SNR when this value is high. The decrease in gain as γ_k decreases and η_k is high, for the case in which $q_k > 0$, is in contrast to the increase in gain for the case in which $q_k = 0$ [i.e., for $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$]. This is probably a result of favoring the hypothesis of signal absence by the amplitude estimator (30) in such a situation.

We conclude this section by noting that the estimator (25) which takes into account the signal presence uncertainty could be obtained from the estimator (4), which assumes that the signal is surely present, if $p(a_k, \alpha_k)$ in (4) is chosen appropriately. This can be done by using

$$p(a_k, \alpha_k) = (1 - q_k)p(a_k, \alpha_k \mid H_k^1) + q_k\,\delta(a_k, \alpha_k) \tag{31}$$

where $p(a_k, \alpha_k \mid H_k^1)$ is the joint PDF of A_k and α_k when the signal is surely present, and $\delta(a_k, \alpha_k)$ is a Dirac function. Under the Gaussian assumption used here, $p(a_k, \alpha_k \mid H_k^1)$ is given by (6). This is an interesting interpretation of the estimator (25), which was originally derived in [10] by minimizing the mean-square estimation error. It also indicates that the estimator derived by using (4) with the above $p(a_k, \alpha_k)$ [or equivalently (25)] is the MMSE estimator for a class of *a priori* PDF's which differ in the probability assumed for signal absence.

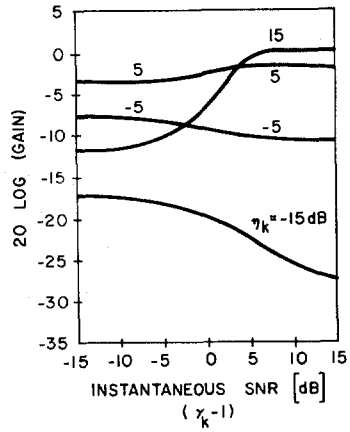


Fig. 4. Parametric gain curves describing the MMSE gain function $G^D_{\mathrm{MMSE}}(\eta_k, \gamma_k, q_k)$ defined by (30) for $q_k = 0.2$.

IV. MMSE Complex Exponential Estimator

In the previous sections we gave the motivation for using an optimal STSA estimator of the speech signal, and derived such an estimator under an assumed statistical model. In this section we concentrate on the derivation of an optimal MMSE estimator of the complex exponential of the phase under the same statistical model. This estimator is combined with the MMSE STSA estimator for constructing the enhanced signal.

We show that the MMSE complex exponential estimator has a nonunity modulus. Therefore, combining it with an optimal amplitude estimator results in a new amplitude estimator which is no longer optimal. On the other hand, the MMSE complex exponential estimator whose modulus is constrained to be unity, and therefore does not affect the amplitude estimation, is the complex exponential of the noisy phase.

We also show in this section that the optimal estimator of the principle value of the phase is the noisy phase itself. This result is of interest, although it does not provide another estimator for the complex exponential than the above constrained one. Its importance follows from the fact that it is unknown which one, the phase or its complex exponential, is more important in speech perception. Therefore, the optimal estimators of both of them should be examined.

Derivation of MMSE Complex Exponential Estimator

Based on the statistical model assumed in Section I, the MMSE estimator of the complex exponential $e^{j\alpha_k}$, given the

noisy observations $\{y(t), 0 \leq t \leq T\}$, is given by

$$\begin{aligned} e^{j\tilde{\alpha}_k} &= E\{e^{j\alpha_k} \mid y(t), 0 \leq t \leq T\} \\ &= E\{e^{j\alpha_k} \mid Y_0, Y_1, \cdots\} \\ &= E\{e^{j\alpha_k} \mid Y_k\} \\ &= E\{e^{-j\varphi_k} \mid Y_k\} e^{j\varphi_k} \\ &= [E\{\cos\varphi_k \mid Y_k\} - jE\{\sin\varphi_k \mid Y_k\}]\, e^{j\varphi_k} \end{aligned} \tag{32}$$

where φ_k is the phase error which is defined by $\varphi_k \triangleq \vartheta_k - \alpha_k$, and ϑ_k is the noisy phase. $E\{\sin\varphi_k \mid Y_k\}$ and $E\{\cos\varphi_k \mid Y_k\}$ can be easily calculated for the Gaussian statistical model assumed here (see Appendix B). We obtain

$$E\{\sin\varphi_k \mid Y_k\} = 0 \tag{33}$$

and

$$\begin{aligned} e^{j\tilde{\alpha}_k} &= E\{\cos\varphi_k \mid Y_k\} e^{j\vartheta_k} \\ &= \Gamma(1.5)\sqrt{v_k}\; M(0.5; 2; -v_k)\, e^{j\vartheta_k} \\ &= \Gamma(1.5)\sqrt{v_k}\exp(-v_k/2)\,[I_0(v_k/2) + I_1(v_k/2)]\, e^{j\vartheta_k} \\ &\triangleq \cos\tilde{\varphi}_k\, e^{j\vartheta_k}. \end{aligned} \tag{34}$$

The combination of the MMSE estimator $e^{j\tilde{\alpha}_k}$ with an independently derived amplitude estimator $\hat{A}_k$ results in the following estimator $\tilde{X}_k$ for the kth spectral component:

$$\tilde{X}_k = \hat{A}_k \cos\tilde{\varphi}_k\, e^{j\vartheta_k}. \tag{35}$$

The modulus of the spectral estimator $\tilde{X}_k$ represents now a new amplitude estimator which is not optimal if $\hat{A}_k$ is optimal. That is, improving the estimation of the complex exponential of the phase (in comparison with the use of the complex exponential of the noisy phase) adversely affects the amplitude estimation.

It is worthwhile to further investigate the estimator (35) when $\hat{A}_k$ is the MMSE estimator from (7). We now show that this estimator is nearly equivalent to the Wiener spectral estimator X_k^w, which is given by (13). On the one hand, this fact implies that $\tilde{X}_k$ is a nearly MMSE spectral estimator, since the Wiener spectral estimator is MMSE. On the other hand, this fact enables us to estimate the degradation in the amplitude estimation by using the error analysis of the previous section.

To show that $\tilde{X}_k$ in (35) and X_k^w in (13) are nearly equivalent, we compare their gain curves for the SNR values which are of interest here. Several of these gain curves are shown in Fig. 5. The closeness of the gain curves, which correspond to the same value of ξ_k, implies that the two estimators $\tilde{X}_k$ and X_k^w are nearly equivalent.

Due to the major importance of the STSA in speech perception, it is of interest to derive an MMSE estimator of the complex exponential of the phase which does not affect the amplitude estimation. To derive this estimator, which we denote by $e^{j\hat{\alpha}_k}$, we solve the following constrained optimization problem:

$$\min_{e^{j\hat{\alpha}_k}} E\{|e^{j\alpha_k} - e^{j\hat{\alpha}_k}|^2\}$$

$$\text{subject to } |e^{j\hat{\alpha}_k}| = 1. \tag{36}$$

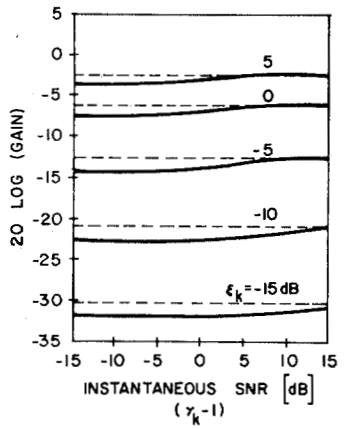


Fig. 5. Parametric gain curves resulting from (a) combined spectral estimator (35) when $\hat{A}_k$ is the MMSE amplitude estimator (7) (solid lines) (b) Wiener spectral estimator (13) (dashed lines).

Using the Lagrange multipliers method, we get

$$e^{j\hat{\alpha}_k} = e^{j\vartheta_k}. \tag{37}$$

That is, the complex exponential of the noisy phase is the MMSE complex exponential estimator which does not affect the amplitude estimation.

Optimal Phase Estimator

The optimal estimator of the principle value of the phase is derived here by minimizing the following distortion measure [20]

$$E\{1 - \cos(\alpha_k - \hat{\alpha}_k)\}. \tag{38}$$

This measure is invariant under modulo 2π transformation of the phase α_k, the estimated phase $\hat{\alpha}_k$, and the estimation error $\alpha_k - \hat{\alpha}_k$. For small estimation errors, (38) is a type of least-square criterion, since $1 - \cos\beta \cong \beta^2/2$ for $\beta << 1$.

The estimator $\hat{\alpha}_k$ which minimizes (38) is easily shown to satisfy

$$tg\,\hat{\alpha}_k = \frac{E\{\sin\alpha_k \mid Y_k\}}{E\{\cos\alpha_k \mid Y_k\}}. \tag{39}$$

By using $\alpha_k = \vartheta_k - \varphi_k$, and $E\{\sin\varphi_k \mid Y_k\} = 0$ [see (33)], it is easy to see that

$$E\{\sin\alpha_k \mid Y_k\} = \sin\vartheta_k \cos\tilde{\varphi}_k \tag{40}$$

$$E\{\cos\alpha_k \mid Y_k\} = \cos\vartheta_k \cos\tilde{\varphi}_k. \tag{41}$$

On substituting (40) and (41) into (39), we get

$$tg\,\hat{\alpha}_k = tg\,\vartheta_k \tag{42}$$

or, alternatively, $\hat{\alpha}_k = \vartheta_k$.

V. *A Priori* SNR Estimation

In this section we address the problem of estimating the *a priori* SNR of a spectral component in a given analysis frame. The *a priori* SNR should be reestimated in each analysis frame, due to the nonstationarity of the speech signal. Two approaches are considered here. In the first, an ML estimator of a speech spectral component variance is utilized. The second approach is based on a "decision-directed" estimation method. Both

approaches assume knowledge of the noise spectral component variance. In practice this variance is estimated from nonspeech intervals which are most adjacent in time to the analysis frame [21], [22]. If the noise is known to be stationary, then it suffices to estimate its spectral component variances once only, from an initial nonspeech interval. We present here the derivation of the above two estimators, and leave for the next section the discussion concerning their application and performance in the proposed speech enhancement system.

Maximum Likelihood Estimation Approach

The ML estimation approach is most commonly used for estimating an unknown parameter of a given PDF [e.g., $\lambda_x(k)$ in (6)], when no *a priori* information about it is available. We now derive the ML estimator of the kth signal spectral component variance in the nth analysis frame. We base the estimation on L consecutive observations $\boldsymbol{Y}_k(n) \triangleq \{Y_k(n), Y_k(n-1), \cdots, Y_k(n-L+1)\}$, which are assumed to be statistically independent. This assumption is reasonable when the analysis is done on nonoverlapping frames. However, in the system used here, overlapping is done (see Section VI). Nevertheless, we continue with this assumption since the statistical dependence is difficult to be modeled and handled. We also assume that the signal and noise kth spectral component variances $\lambda_x(k)$ and $\lambda_d(k)$, respectively, are slowly varying parameters, so that they can be considered constant during the above L observations. Finally, we assume that the kth noise spectral component variance is known.

The ML estimator $\hat{\lambda}_x(k)$ of $\lambda_x(k)$, which is constrained to be nonnegative, is the nonnegative argument which maximizes the joint conditional PDF of $Y_k(n)$ given $\lambda_x(k)$ and $\lambda_d(k)$. Based on the Gaussian statistical model and the statistical independence assumed for the spectral components, this PDF is given by

$$p(\boldsymbol{Y}_k(n)|\lambda_x(k), \lambda_d(k)) = \prod_{l=0}^{L-1} \frac{1}{\pi(\lambda_x(k)+\lambda_d(k))} \cdot \exp\left(-\frac{R_k^2(n-l)}{\lambda_x(k)+\lambda_d(k)}\right) \tag{43}$$

where $R_k(l) \triangleq |Y_k(l)|$. $\hat{\lambda}_x(k)$ is easily obtained from (43), and equals

$$\hat{\lambda}_x(k) = \begin{cases} \frac{1}{L}\sum_{l=0}^{L-1} R_k^2(n-l) - \lambda_d(k) & \text{if nonnegative} \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

This estimator suggests the following estimator for the *a priori* SNR ξ_k.

$$\hat{\xi}_k = \begin{cases} \frac{1}{L}\sum_{l=0}^{L-1} \gamma_k(n-l) - 1 & \text{if nonnegative} \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

where $\gamma_k(l) \triangleq |Y_k(l)|^2/\lambda_d(k)$ is the *a posteriori* SNR in the lth analysis frame.

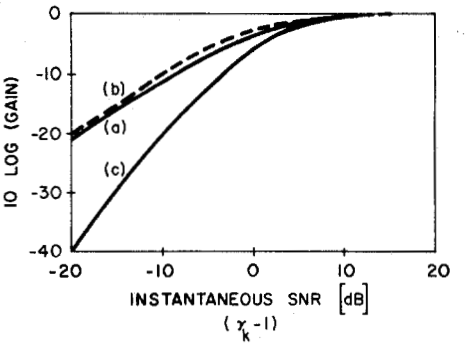


Fig. 6. Gain curves describing (a) MMSE gain function $G_{\mathrm{MMSE}}(\xi_k, \gamma_k)$ defined by (7) and (14), with $\xi_k = \gamma_k - 1$, (b) "spectral subtraction" gain function (46) with $\beta = 1$, and (c) Wiener gain function $G_w(\xi_k, \gamma_k)$ (15) with $\xi_k = \gamma_k - 1$.

It is interesting to consider the ML estimator (44) when $L = 1$. In this case we get the "power spectral subtraction" estimator derived in [3]. The application of the corresponding ξ_k estimator (45) (with $L = 1$) to the MMSE amplitude estimator (7) results in a gain function which depends on γ_k only. Surprisingly, this gain function is almost identical to the "spectral subtraction" gain function for a wide range of SNR values. The "spectral subtraction" gain function is given by (46) [1], and the above near-equivalence occurs when $\beta = 1$.

$$G_{SP}(\gamma_k) \triangleq \frac{A_{SP}(\gamma_k)}{R_k} = \sqrt{1 - \frac{\beta}{\gamma_k}}, \quad \beta \geq 1. \tag{46}$$

This fact is demonstrated in Fig. 6. For comparison purposes, the same figure also shows the gain curve for the Wiener amplitude estimator, which results from (12), and the same *a priori* SNR estimator (i.e., (45) with $L = 1$).

In practice, the running average needed in (45) is replaced by a recursive averaging with a time constant comparable to the correlation time of γ_k. That is, the estimator of ξ_k in the nth analysis frame is obtained by

$$\bar{\gamma}_k(n) = \alpha\bar{\gamma}_k(n-1) + (1-\alpha)\frac{\gamma_k(n)}{\beta}, \quad 0 \leq \alpha < 1, \ \beta \geq 1.$$

$$\hat{\xi}_k(n) = \begin{cases} \bar{\gamma}_k(n) - 1 & \bar{\gamma}_k(n) - 1 \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

β is a correction factor, and here it plays the same role as in the "spectral subtraction" estimator (46). The values of α and β are determined by informal listening, as is explained in Section VI.

"Decision-Directed" Estimation Approach

We now consider the estimation of the *a priori* SNR of a spectral component by a "decision-directed" method. This estimator is found to be very useful when it is combined with either the MMSE or the Wiener amplitude estimator.

Let $\xi_k(n)$, $A_k(n)$, $\lambda_d(k, n)$, and $\gamma_k(n)$ denote the *a priori* SNR, the amplitude, the noise variance, and the *a posteriori* SNR, respectively, of the corresponding kth spectral component in the nth analysis frame. The derivation of the *a priori*

SNR estimator is based here on the definition of $\xi_k(n)$, and its relation to the *a posteriori* SNR $\gamma_k(n)$, as given below:

$$\xi_k(n) = \frac{E\{A_k^2(n)\}}{\lambda_d(k,n)} \tag{48}$$

$$\xi_k(n) = E\{\gamma_k(n) - 1\}. \tag{49}$$

Using (48) and (49) we can write

$$\xi_k(n) = E\left\{\frac{1}{2}\frac{A_k^2(n)}{\lambda_d(k,n)} + \frac{1}{2}[\gamma_k(n) - 1]\right\}. \tag{50}$$

The proposed estimator $\hat{\xi}_k(n)$ of $\xi_k(n)$ is deduced from (50), and is given by

$$\hat{\xi}_k(n) = \alpha\frac{\hat{A}_k^2(n-1)}{\lambda_d(k,n-1)} + (1-\alpha)P[\gamma_k(n) - 1], \quad 0 \leqslant \alpha < 1 \tag{51}$$

where $\hat{A}_k(n-1)$ is the amplitude estimator of the kth signal spectral component in the $(n-1)$th analysis frame, and $P[\cdot]$ is an operator which is defined by

$$P[x] = \begin{cases} x & \text{if } x \geqslant 0 \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

By comparing (50) and (51), we see that $\hat{\xi}_k(n)$ is obtained from (50) by dropping the expectation operator, using the amplitude estimator of the $(n-1)$th frame instead of the amplitude itself in the nth fame, introducing a weighting factor between the two terms of $\xi_k(n)$, and using the operator $P[\cdot]$ defined in (52). $P[\cdot]$ is used to ensure the positiveness of the proposed estimator in case $\gamma_k(n) - 1$ is negative. It is also possible to apply the operator P on the right-hand side of (51) rather than on $\gamma_k(n) - 1$ only. However, from our experience both alternatives give very similar results.

The proposed estimator for $\xi_k(n)$ is a "decision-directed" type estimator, since $\hat{\xi}_k(n)$ is updated on the basis of a previous amplitude estimate.

By using $\hat{A}_k(n) = G(\hat{\xi}_k(n), \gamma_k(n))\, R_k(n)$, where $G(\cdot,\cdot)$ is a gain function which results from either the MMSE or the Wiener amplitude estimator, (51) can be written in a way which emphasizes its recursive nature. We get from (51)

$$\hat{\xi}_k(n) = \alpha G^2(\hat{\xi}_k(n-1), \gamma_k(n-1))\,\gamma_k(n-1) + (1-\alpha)P[\gamma_k(n) - 1]. \tag{53}$$

Several initial conditions were examined by simulations. We found that using $\xi_k(0) = \alpha + (1-\alpha)P[\gamma_k(0) - 1]$ is appropriate, since it minimizes initial transition effects in the enhanced speech.

The theoretical investigation of the recursive estimator (53) is very complicated due to its highly nonlinear nature. Even for the simple gain function of the Wiener amplitude estimator it was difficult to analyze. Therefore, we examined it by simulation only, and determined in this way the "best" value of α.

VI. System Description and Performance Evaluation

In this section we first describe the proposed speech enhancement system, which was implemented on a general purpose computer (Eclipse S-250). Then we describe the performance of this system, based on informal listening, when each of the STSA estimators discussed in this paper is applied.

System Description

The input to the proposed system is an 8 kHz sampled speech of 0.2–3.2 kHz bandwidth, which was degraded by uncorrelated additive noise. Each analysis frame which consists of 256 samples of the degraded speech, and overlaps the previous analysis frame by 192 samples, is spectrally decomposed by means of a discrete short-time Fourier transform (DSTFT) analysis [23], [24] using a Hanning window. The STSA of the speech signal is then estimated, and combined with the complex exponential of the noisy phase. The estimated DSTFT samples in each analysis frame are used for synthesizing the enhanced speech signal by using the well-known weighted overlap and add method [24].

In applying the MMSE amplitude estimators (7) and (30) in the proposed system, we examined their implementation through exact calculation as well as by using lookup tables. Each lookup table contains a finite number of samples of the corresponding gain function in a prescribed region of (ξ, γ). We found, for example, that when the input SNR is in the range [-5, 5] dB, and the "decision-directed" *a priori* SNR is utilized, it suffices to use 961 samples of each gain function, which are obtained by uniformly sampling the range $-15 \leqslant [(\xi, \gamma - 1)$ or $(\eta, \gamma - 1)] \leqslant 15$ dB. As judged by informal listening, this sampling of the gain functions results in a negligible additional residual noise to the enhanced signal. Therefore, the proposed system operating with the MMSE amplitude estimator can be implemented with a similar complexity to that of other commonly used systems, although a more complicated amplitude estimator is used here.

The proposed system is examined here for enhancing speech degraded by stationary noise. Therefore, the variances of the noise spectral components are estimated only once, from an initial noise segment having a duration of 320 ms. The estimated variances are used in the estimation of γ_k, and ξ_k by either (47) or (51).

Performance Evaluation

In this section we describe the performance of the above speech enhancement systems when each of the STSA estimators considered in this paper is applied. Both *a priori* SNR estimators (i.e., the ML and the "decision-directed") are examined. The values used here for the parameters α and β in (47) and (51) are the best ones found by simulations. Fig. 7 describes a chart of the comparison tests made here.

In each test, speech signals which were degraded by stationary uncorrelated additive wide-band noise with SNR values of 5, 0, and -5 dB were enhanced. The speech material used includes the following sentences, each spoken by a female and a male:

A lathe is a big tool.
An icy wind raked the beach.
Joe brought a young girl.

In addition the sentence "we were away a year ago," spoken by another male, was examined. Six listeners participated in the comparison tests. In each test a pair of the enhanced speech signals were presented to the listeners (through earphones), and they were asked to compare them on the following basis:

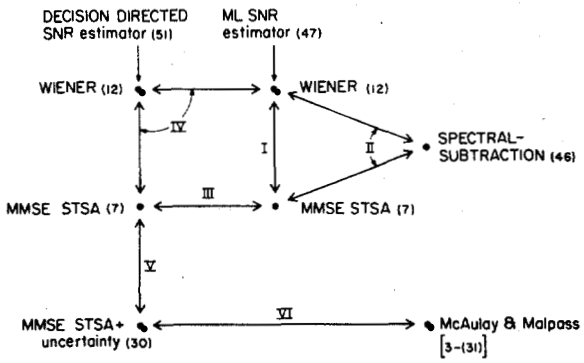


Fig. 7. Comparison listening tests chart.

amount of noise reduction, the nature of the residual noise (e.g., musical versus uniform), and distortion in the speech signal itself.

Let us consider first the tests in which STSA estimators whose derivation is based on the assumption that the speech is surely present in the noisy observations are used.

Case I: Using either the MMSE amplitude estimator (7) or the Wiener amplitude estimator (12), when the *a priori* SNR is estimated by the ML estimator (47) with $\alpha = 0.725, \beta = 2$, gives a very similar enhanced speech quality. A significant reduction of the noise is perceived, but a "musical noise" is introduced. The power of this "musical noise" is very low at the 5 dB SNR value, and it increases as the input SNR decreases. The distortions caused to the speech signal seem to be very small at the high SNR value of 5 dB, and increase as the input SNR decreases. Nevertheless, at the SNR value of -5 dB, the enhanced speech is still very intelligible.

Case II: The enhanced speech obtained by using the "spectral subtraction" amplitude estimator (46) with $\beta = 2$, suffers from a strong "musical noise." This "musical noise" is of higher power level and wider band than the "musical noise" obtained in the above MMSE and Wiener amplitude estimations (Case I). This is especially prominent at the low input SNR values of 0 dB and -5 dB. For this reason, the quality of the enhanced speech obtained by using either the MMSE or the Wiener amplitude estimator is much better than that obtained by using the "spectral subtraction" estimator.

Case III: Using the MMSE amplitude estimator (7) when the *a priori* SNR is estimated by the "decision-directed" estimator (51) with $\alpha = 0.98$ results in a great reduction of the noise, and provides enhanced speech with *colorless* residual noise. This colorless residual noise was found to be much less annoying and disturbing than the "musical noise" obtained when the *a priori* SNR is estimated by the ML estimator (47). As could be judged by informal listening, the distortions in the enhanced speech obtained by using the MMSE amplitude estimator with either the ML or the "decision-directed" *a priori* SNR estimator, are very similar.

Case IV: Using the Wiener amplitude estimator with the "decision-directed" *a priori* SNR estimator and $\alpha = 0.98$ results in a more distorted speech than that obtained by using the recently described MMSE amplitude estimator (Case III). However, the residual noise level in the Wiener estimation is lower than that in the MMSE estimation. Lowering the value of α reduces the distortions of the enhanced speech, but introduces a residual "musical noise" as well. This "musical noise" is probably contributed to by the second term of the "decision-directed" estimator [i.e., $P[\gamma_k(n) - 1]$ in (51)], whose relative weight increases as the value of α decreases. We found that using $\alpha = 0.97$ results in an enhanced speech whose distortion is similar to that obtained by using the MMSE amplitude estimator. In addition, the level of the residual "musical noise" obtained then is lower than that obtained by using the ML *a priori* SNR estimator.

Case V: The MMSE amplitude estimator (30), which takes into account the uncertainty of signal presence in the observed signal, results in a better enhanced speech quality than that obtained by using the MMSE estimator (7). Specifically, by using (30) with $q_k = 0.2$, and the "decision-directed" *a priori* SNR estimator (51) [when $\hat{\xi}_k(n)$ is replaced by $\hat{\eta}_k(n)$] with $\alpha = 0.99$, we get a further reduction of the colorless residual noise, with negligible additional distortions in the enhanced speech signal.

Case VI: The enhanced speech obtained by using the above MMSE amplitude estimator [(30) with $q_k = 0.2$, and (51) with $\alpha = 0.99$], was compared with the enhanced speech obtained by using the McAulay-Malpass amplitude estimator [3] (see Section III). The latter estimator was operated with the "best" value (as judged by informal listening) of the *a priori* SNR parameter, which was found to be 12 dB in our experiment. It was found that the main difference between the two enhanced speech signals is in the nature of the residual noise. When the MMSE estimator is used the residual noise is colorless, while when the McAulay-Malpass estimator is used, musical residual noise results.

VII. Summary and Discussion

We present in this paper an algorithm for enhancing speech degraded by uncorrelated additive noise when the noisy speech alone is available. The basic approach taken here is to optimally estimate (under the MMSE criterion and an assumed statistical model) the short-time spectral amplitude (STSA) and complex exponential of the phase of the speech signal. We use this approach of optimally estimating the two components of the short-time Fourier transform (STFT) separately, rather than optimally estimating the STFT itself, since the STSA of a speech signal rather than its waveform is of major importance in speech perception. We showed that the STSA and the complex exponential cannot be estimated simultaneously in an optimal way. Therefore, we use an optimal MMSE STSA estimator, and combine it with an optimal MMSE estimator of the complex exponential of the phase which does not affect the STSA estimation. The latter constrained complex exponential estimator is found to be the complex exponential of the noisy phase.

In this paper we derive the MMSE STSA estimator and analyze its performance. We showed that the MMSE STSA estimator, and the Wiener STSA estimator which results from the optimal MMSE STFT estimator, are nearly equivalent at high SNR. On the other hand, the MMSE STSA estimator results in significantly less MSE and bias when the SNR is low. This fact supports our approach to optimally estimate the perceptually important STSA directly from the noisy observations rather than deriving it from another estimator (e.g., from the Wiener one).

A MMSE STSA estimator which takes into account the uncertainty of signal presence in the noisy spectral components is also derived in this paper, and examined in enhancing speech.

The MMSE STSA estimator depends on the parameters of the statistical model it is based on. In the proposed algorithm these are the *a priori* SNR of each spectral component, and the variance of each noise spectral component. The *a priori* SNR was found to be a key parameter of the STSA estimator. It is demonstrated here that by using different estimators for the *a priori* SNR, different STSA estimations result. For example, using the "power spectral subtraction" method for estimating the *a priori* SNR results in an STSA estimator which is nearly equivalent to the "spectral subtraction" STSA estimator.

We proposed here a "decision-directed" method for estimating the *a priori* SNR. This method was found to be useful when it is applied to either the MMSE or the Wiener STSA estimator. By combining this estimator with the MMSE STSA estimator which takes into account the uncertainty of signal presence in the noisy observations, we obtained the best speech enhancement results. Specifically, a significant reduction of the input noise is obtained, and the residual noise sounds colorless.

We believe that the full potential of the proposed approach is not yet exploited, although very encouraging results were obtained. Better results may be obtained if the *a priori* SNR estimation could be improved. This issue is now being investigated.

Appendix A

In this Appendix we derive the MMSE amplitude estimator (7). On substituting (5) and (6) into (4), and using the integral representation of the modified Bessel function of the nth order [13, eq. 8.431.5],

$$I_n(z) = \frac{1}{2\pi} \int_0^{2\pi} \cos \beta n \exp(z \cos \beta)\, d\beta \tag{A.1}$$

we obtain

$$\hat{A}_k = \frac{\int_0^{\infty} a_k^2 \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k}{\int_0^{\infty} a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k} \tag{A.2}$$

where v_k is defined by (8), and $\lambda(k)$ satisfying

$$\frac{1}{\lambda(k)} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)}. \tag{A.3}$$

By using [13, eq. 6.631.1, 8.406.3, 9.212.1] we get from (A.2)

$$\hat{A}_k = \lambda(k)^{1/2}\, \Gamma(1.5) M(-0.5; 1; -v_k). \tag{A.4}$$

$\hat{A}_k$, as given by (7), is obtained from (A.4) by using (A.3) and (8)–(10). The equivalent form of $\hat{A}_k$ as given in (7) is obtained by using [4, eq. A.1.31a].

Appendix B

In this Appendix we derive the MMSE estimators of $\cos \varphi_k$, and $\sin \varphi_k$, given the noisy spectral component Y_k.

$$E\{\cos \varphi_k \mid Y_k\} = \int_0^{2\pi} \cos(\vartheta_k - \alpha_k) p(\alpha_k \mid Y_k) d\alpha_k$$

$$= \frac{\int_0^{\infty} \int_0^{2\pi} \cos(\vartheta_k - \alpha_k) p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k\, da_k}{\int_0^{\infty} \int_0^{2\pi} p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k\, da_k}. \tag{B.1}$$

On substituting (5) and (6) into (B.1), and using (A.1), we obtain

$$E\{\cos \varphi_k \mid Y_k\} = \frac{\int_0^{\infty} a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_1\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k}{\int_0^{\infty} a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) I_0\left(2a_k \sqrt{\frac{v_k}{\lambda(k)}}\right) da_k} \tag{B.2}$$

where v_k and $\lambda(k)$ are defined by (8) and (A.3), respectively. By using [13, eq. 6.631.1, 8.406.3, 9.212.1], we get from (B.2)

$$E\{\cos \varphi_k \mid Y_k\} = \Gamma(1.5) \sqrt{v_k}\, M(0.5; 2; -v_k). \tag{B.3}$$

The equivalent form of $E\{\cos \varphi_k \mid Y_k\}$, as given in (34), is obtained by using [4, eq. A.1.31d].

To show that $E\{\sin \varphi_k \mid Y_k\} = 0$ we substitute (5) and (6) into

$$E\{\sin \varphi_k \mid Y_k\} = \frac{1}{p(Y_k)} \int_0^{\infty} \int_0^{2\pi} \sin(\vartheta_k - \alpha_k) \cdot p(Y_k \mid a_k, \alpha_k) p(a_k, \alpha_k) d\alpha_k\, da_k. \tag{B.4}$$

We obtain

$$E\{\sin \varphi_k \mid Y_k\} \sim \int_0^{\infty} a_k \exp\left(-\frac{a_k^2}{\lambda(k)}\right) \int_0^{2\pi} \sin(\vartheta_k - \alpha_k) \cdot \exp\left(\frac{2a_k R_k}{\lambda_d(k)} \cos(\vartheta_k - \alpha_k)\right) d\alpha_k\, da_k \tag{B.5}$$

where $\sim$ denotes proportionality. Now it is easy to see that the inner integral in (B.5) equals zero.

ACKNOWLEDGMENT

The authors wish to thank Prof. I. Bar-David, Prof. M. Zakai, and Dr. M. Sidi for fruitful and helpful discussions in the course of this work. The authors also wish to thank S. Shitz and the anonymous reviewers for critical reading of the manuscript and for their helpful comments.

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586-1604, Dec. 1979.

[2] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972, p. 210.

[3] 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, pp. 137-145, Apr. 1980.

[4] D. Middelton, *Introduction to Statistical Communication Theory*. New York: McGraw-Hill, 1960, ch. 7, appendix 1.

[5] W. A. Pearlman and R. M. Gray, "Source coding of the discrete Fourier transform," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 683-692, Nov. 1978.

[6] W. B. Davenport and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill, 1958, ch. 6.

[7] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, p. 522, Oct. 1979.

[8] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, p. 306, Aug. 1977.

[9] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 1984, pp. 18A.2.1-18A.2.4.

[10] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 434-444, May 1968.

[11] Y. Ephraim and D. Malah, "Speech enhancement under uncertainty of signal presence in the observed signal," Dep. Elec. Eng., Technion, Haifa, Israel, EE pub. 543, July 1983.

[12] T. T. Kadota, "Optimal reception of binary Gaussian signals," *Bell Syst. Tech. J.*, vol. 43, pp. 2767-2810, Nov. 1964.

[13] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic, 1980.

[14] D. Middelton, "The incoherent estimation of signal amplitude in normal noise backgrounds," in *Time Series Analysis*, M. Rosenblatt, Ed. New York: Wiley, 1963, ch. 24.

[15] Y. Ephraim and D. Malah, "Speech enhancement using optimal nonlinear spectral amplitude estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1983, pp. 1118-1121.

[16] —, "Speech enhancement using vector spectral subtraction amplitude estimation," in *Proc. IEEE 13th Conv. Elec. Electron. Eng. in Israel*, Tel-Aviv, Mar. 1983.

[17] M. Loeve, *Probability Theory*, 3rd ed. Princeton, NJ: Van Nostrand, 1963.

[18] 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.

[19] Y. Ephraim, "Enhancement of noisy speech," D.Sc. dissertation, Technion-Israel Inst. Technol., Haifa, 1984.

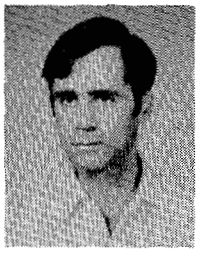
[20] A. S. Wilsky, "Fourier series and estimation on the circle with applications to synchronous communication-Part I: Analysis," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 577-583, Sept. 1974.

[21] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113-120, Apr. 1979.

[22] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 786-794, Aug. 1981.

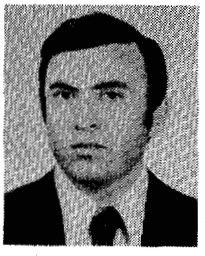
[23] M. R. Portnoff, "Time frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 55-69, Feb. 1980.

[24] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 99-102, Feb. 1980.



Yariv Ephraim (S'81) was born on September 9, 1951. He received the B.Sc., M.Sc., and D.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology, Haifa, in 1977, 1979, and 1984, respectively.

He is currently a Post Doctoral Fellow in the Information Systems Laboratory, Stanford University, Stanford, CA. His present interests are in estimation theory and its applications, and in statistical approaches to speech and image processing.



David Malah (S'67-M'71-SM'84) was born in Poland on March 31, 1943. He received the B.Sc. and M.Sc. degrees in 1964 and 1967, respectively, from the Technion-Israel Institute of Technology, Haifa, and the Ph.D. degree in 1971 from the University of Minnesota, Minneapolis, all in electrical engineering.

During 1971-1972 he was an Assistant Professor at the Department of Electrical Engineering of the University of New Brunswick, Fredericton, N.B., Canada. In 1972 he joined the Department of Electrical Engineering of the Technion, where he is currently an Associate Professor. From 1979 to 1981 he was on sabbatical and leave at the Acoustics Research Department of AT&T Bell Laboratories, Murray Hill, NJ. Since 1975 (except from 1979 to 1981) he has been in charge of a newly established Signal Processing Laboratory which is active in speech and image communication research and real time hardware developments. His main research interests are in digital speech and image coding, speech and image enhancement, and digital signal processing techniques.