Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise

Udar Mittal and Nam Phamdo, *Senior Member, IEEE*

***Abstract*—A signal/noise Karhunen–Loeve transform (KLT) based approach for enhancing speech degraded by colored noise is proposed. The noisy speech frames are classified into speech-dominated frames and noise-dominated frames. In the speech-dominated frames, the signal KLT matrix is used and in the noise dominated frames, the noise KLT matrix is used. The approach does not require noise whitening and hence works well even with narrowband noise. A two-dimensional objective measure which captures both the speech distortion and the noise shaping characteristics of the algorithm is proposed. This measure indicates that the proposed method performs better noise shaping than a modified form of the signal subspace approach proposed by Ephraim and Van Trees and the standard spectral subtraction method. Informal listening tests show that the proposed algorithm does not suffer from the problem of residual musical noise and performs better noise masking than the signal subspace approach.**

***Index Terms*—Colored noise, Karhunen–Loeve transform (KLT), speech enhancement.**

I. Introduction

THE MAIN objective of speech enhancement algorithms is to improve the performance of speech communication systems in a noisy environment. Speech enhancement methods attempt to improve either the subjective quality of the speech to reduce listener fatigue or improve the intelligibility of noisy speech. Further, these algorithms also improve the performance of other speech processing systems (such as an automatic speech recognizer) developed for a noise free environment.

Spectral subtraction [1] is a traditional method for enhancing speech degraded by an additive stationary background noise in a single channel system. The major drawback of this method is the characteristic of the residual noise called *musical* noise. It comprises of tones of random frequencies. Various modifications of spectral subtraction [2] based on the noise masking property of the human ear have been developed. The noise masking properties are modeled by calculating a *noise masking threshold*. A listener tolerates additive noise as long as it remains below the masking threshold.

Another drawback of speech enhancement methods is the distortion of the useful signal. The compromise between signal distortion and the level of residual noise is a well-known problem in speech enhancement. Both of these can not be minimized simultaneously. Minimum mean square error (MMSE) estimates with constraints on speech spectra [3] and constraints on hidden Markov model as well as codebook constrained Weiner filters [5] have been proposed. Ephraim and Van Trees [6], [7] proposed a signal-subspace based spectral domain constrained estimator which controls the harmful components of residual noise while minimizing the signal distortion. This estimator minimizes the signal distortion while keeping the energy of the residual noise in each spectral component below some threshold. This allows shaping of the residual noise spectrum so that its perception can be minimized. The main drawback of this method is that it deals only with white noise. Though, for nonwhite broadband noise, Ephraim and Van Trees have proposed whitening of the noise, i.e., premultiplying the noisy speech signal by the square root of the noise covariance matrix's inverse. The spectral domain constraints are applied after the noise whitening has been performed and hence it does not ensure that the residual noise has spectral shape similar to that of the clean speech. Further, this method requires inverting the noise covariance matrix, hence it can not be used for narrowband noise.

In [8], Jensen *et al.* proposed a quotient singular value decomposition (QSVD) based approach for colored noise. This approach also requires prewhitening (it is an integral part of the algorithm). Moreover, QSVD algorithms are computationally intensive and do not provide noise shaping.

In order to provide a proper noise shaping for colored noise we propose an approach which does not require noise whitening. We refer to it as the signal/noise KLT based approach. Since the approach does not require matrix inversion, it works well even with narrowband noise.

In Section II, we review the signal subspace based spectral domain constrained estimator [6], [7] for white noise and its modification for colored noise. In this section, we also show the problems associated with noise shaping with this modification. In Section III, a signal/noise KLT based estimator is derived for enhancing signals corrupted by colored noise. Implementation details of this approach are also presented in this section. In Section IV, we compare the signal/noise KLT approach, the signal subspace approach and the spectral subtraction in terms of a two-dimensional spectral distortion measure and informal listening tests. The computational complexity of the proposed approach and signal subspace approach are also compared in this section.

II. Signal Subspace Approach

Our notations are adopted from [6] and [7].

Manuscript received August 4, 1998; revised November 5, 1998. This work was supported by Northrop Grumman Corporation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-Claude Junqua.

The authors are with Department of Electrical and Computer Engineering, State University of New York, Stony Brook, NY 11794-2350 USA (e-mail: udar@sbee.sunysb.edu; phamdo@sbee.sunysb.edu).

Publisher Item Identifier S 1063-6676(00)01710-7.

Notations: In subsequent sections, lower case letters represent K-dimensional column vectors. A $K \times K$ matrix R_s represents the covariance matrix of the vector s. Let $\hat{s}$ be an estimate of s and $\hat{R}_s$ be an estimate of R_s. For a given matrix A, $A^{\#}$ is the conjugate transpose of A and $\mathrm{tr}(A)$ is the trace of A (the sum of the diagonal elements of A). Let

$$R_s = U_s \Lambda_s U_s^{\#} \tag{1}$$

be the eigenvalue decomposition of R_s, $\lambda_s(k)$ be the kth diagonal element of Λ_s and u_{sk} be the corresponding eigenvector. Let y, w, and $z = y + w$ be K-dimensional vectors of the clean speech, the additive noise and the noisy speech signal, respectively.

A. Spectral Domain Constraint (SS-SDC) Estimator for White Noise

In this subsection, we assume that the noise vector w is white. Let $\hat{y} = Hz = Hy + Hw$ be a linear estimator of y. Here, H represents the speech enhancement filter. The residual error signal is given by

$$r = \hat{y} - y = (H - I)y + Hw = r_y + r_w \tag{2}$$

where $r_y$represents the signal distortion and r_w represents the residual noise. In [6] and [7], a linear estimator is obtained which minimizes the signal distortion subject to constraints on the spectrum of the residual noise. This spectrum can be made similar to that of speech and hence residual noise can be masked by the speech signal. Let

$$R_y = U_y \Lambda_y U_y^{\#} \tag{3}$$

be the eigenvalue decomposition of the covariance matrix R_y of y. Let $U_y = [U_1, U_2]$, where

$$U_1 = \{u_{yk} \colon \lambda_y(k) > 0\} \tag{4}$$

and

$$U_2 = \{u_{yk} \colon \lambda_y(k) = 0\}. \tag{5}$$

The kth spectral component of the residual noise is given by $u_{yk}^{\#} r_w$. Here, the filter H is designed by minimizing

$$\epsilon_y = \mathrm{tr}\left(E\left(r_y r_y^{\#}\right)\right) \tag{6}$$

subject to

$$E\left\{\left|u_{yk}^{\#} r_w\right|^2\right\} \leq \alpha_k \sigma_w^2 \tag{7}$$

where σ_w^2 is the noise variance. The constraint constant α_k is zero if $u_{yk} \in U_2$. The filter H thus obtained is given by

$$H = U_y Q U_y^{\#} \tag{8}$$

where

$$Q = \mathrm{diag}\left(\alpha_k^{1/2}\right). \tag{9}$$

B. Modification for Colored Noise

For colored broadband noise, whitening of the colored noise is first performed. Let

$$\tilde{z} = R_w^{-(1/2)} z = R_w^{-(1/2)} y + R_w^{-(1/2)} w = \tilde{y} + \tilde{w}. \tag{10}$$

Note that $\tilde{w}$ is white with variance $\sigma_{\tilde{w}}^2 = 1$. Let $r_{\tilde{y}} = \tilde{H}\tilde{y} - \tilde{y}$. The filter $\tilde{H}$ is obtained by minimizing

$$\epsilon_{\tilde{y}} = \mathrm{tr}\left[E\left\{r_{\tilde{y}} r_{\tilde{y}}^{\#}\right\}\right] \tag{11}$$

subject to

$$E\left\{\left|u_{\tilde{y}k}^{\#} r_{\tilde{w}}\right|^2\right\} \leq \alpha_k \sigma_{\tilde{w}}^2 \tag{12}$$

where $r_{\tilde{w}} = \tilde{H}\tilde{w}$. The effective filter H is derived from $\tilde{H}$ by

$$H = R_w^{1/2} \tilde{H} R_w^{-(1/2)}. \tag{13}$$

Note that the constrained optimization problem here tries to shape the spectrum of $r_{\tilde{w}}$ like the spectrum of $\tilde{y}$. This shaping does not ensure that the residual noise ($r_w = Hw$) spectrum is shaped like that of y. Further, this approach minimizes the variance $\epsilon_{\tilde{y}}$ of $r_{\tilde{y}}$ rather than minimizing the variance ϵ_y of $r_y = (H - I)y$. Since

$$r_y = R_w^{1/2} r_{\tilde{y}} \tag{14}$$

$$\epsilon_y = \mathrm{tr}\left(R_w^{1/2} E\left(r_{\tilde{y}} r_{\tilde{y}}^{\#}\right) R_w^{1/2}\right). \tag{15}$$

Thus, the filter H derived from $\tilde{H}$ may not minimize ϵ_y. Note that ϵ_y is the square of the Frobenius norm [9] of $R_w^{1/2}(E(r_{\tilde{y}} r_{\tilde{y}}^{\#}))^{1/2}$. Let $\|A\|_f$ represents the Frobenius norm of a matrix A. Since

$$\|AB\|_f^2 \leq \|A\|_f^2 \|B\|_f^2 \tag{16}$$

so

$$\begin{aligned} \epsilon_y &\leq \left\|R_w^{1/2}\right\|_f^2 \left\|\left(E\left(r_{\tilde{y}} r_{\tilde{y}}^{\#}\right)\right)^{1/2}\right\|_f^2 \\ &= \mathrm{tr}(R_w)\mathrm{tr}\left(E\left(r_{\tilde{y}} r_{\tilde{y}}^{\#}\right)\right) \\ &= \mathrm{tr}(R_w)\epsilon_{\tilde{y}}. \end{aligned} \tag{17}$$

Thus, minimizing $\epsilon_{\tilde{y}}$ is equivalent to minimizing the upper bound on ϵ_y.

III. Signal/Noise KLT Based Approach

A. Mathematical Background

The subspace based estimates which exploit the orthogonality between an estimated subspace and parameter dependent subspace has become quite popular in signal processing applications. Pisarenko's method [10] which uses the eigensubspace of a certain covariance matrix to estimate the frequencies of a sine wave in white noise have been applied to a variety of problems, such as linear system identification [12], direction of arrival estimation [11], and enhancement of speech corrupted by white noise [6]. For colored noise, we discussed the noise-whitening method and its disadvantages in the previous section. Here we present a subspace based speech enhancement algorithm for colored noise which does not require noise whitening. Since signal and noise are independent, $R_z = R_y + R_w$. The range space of R_y is the space spanned by the eigenvectors of R_y with nonzero eigenvalues. We call this the signal subspace. Let

$$R_y = U_y \Lambda_y U_y^{\#} \tag{18}$$

be the eigenvalue decomposition of R_y. Note that $U_y^{\#}$ is the KLT matrix of the signal. Thus the subspace based approach is referred to as the KLT based approach. Let $U_y = [U_1, U_2]$, where U_1 denotes the $K \times M$ matrix of eigenvectors of R_y with positive eigenvalues. Note that the span of U_1 is the signal subspace and the span of U_2 is orthogonal to the signal subspace. In case of white noise, the orthogonal subspace is also referred to as the noise subspace.

The subspace based estimates are based on the assumption that the span of R_y can be accurately estimated from the data. This estimate is often called the sampled subspace. Note that $U_1 U_1^{\#}$ and $U_2 U_2^{\#}$ are orthogonal projections on the signal subspace and the orthogonal subspace, respectively. The energy of the signal part in the orthogonal projection $U_2 U_2^{\#} z$ is zero. Hence, this part does not provide any information about the signal y. Least-square estimation methods [13] nullify this component and produce an estimate which is the projection on to the signal subspace. Parametric estimation methods [13] produce an speech parameters estimate by minimizing the norm between the parameter-dependent signal subspace (eigenspace) and the estimated signal subspace or by exploiting the orthogonality between the parameter-dependent subspace and the estimated orthogonal subspace. These estimates are in general nonlinear and also require parametric modeling of speech. The least-square estimates here are linear but they neither reduce the noise energy in the signal subspace nor do they provide noise shaping. We propose a linear method which not only ensures that no residual noise in the orthogonal subspace is present in the processed signal but also ensures proper noise shaping and noise energy reduction in the signal subspace. The mathematical formulation of the constrained optimization problem and its solution is given in Section III-C and III-E.

B. *Frame Classification*

Speech frames can be classified into speech activity frames and silent frames. The speech activity frames can further be divided into voiced speech and unvoiced speech frames. The energy of speech in the voiced speech frames is typically higher than in the unvoiced speech frames and in the silent frames. Since the additive noise is stationary, the segmental signal-to-noise ratio (Seg-SNR) varies from frame-to-frame. A different enhancement approach should be used for frames with different Seg-SNR. To incorporate this, the noisy speech frames are classified into two categories, namely *speech dominated frames* and *noise dominated frames*. The signal KLT matrix is used for speech dominated frames and a noise KLT matrix is used for noise dominated frames. If $\mathrm{tr}(R_z) > (\max(10^{\mathrm{SNR}/20}, 1) + 0.3)\,\mathrm{tr}(R_w)$ then the given frame is classified as speech dominated, otherwise it is noise dominated. The value $(\max(10^{\mathrm{SNR}/20}, 1) + 0.3)$ has been determined empirically. Note that the algorithm depends on the *a priori* knowledge of SNR. This SNR can be evaluated by long term averaging. A sensitivity study of the proposed algorithm to mismatch of the long term SNR is provided in Section IV-B.

C. *Signal KLT Approach for Speech Dominated Frames*

As mentioned in Section II-B, for colored noise, the noise whitening method neither ensures proper spectral shaping of residual noise nor does it minimize the signal distortion ϵ_y. In order to provide proper shaping of the spectrum of the residual noise (r_w) and minimization of the variance (ϵ_y) of the signal distortion (r_y), we propose a constrained optimization problem given by

$$\min_H \epsilon_y \tag{19}$$

subject to constraints

$$E\left(\left|u_{yk}^{\#} r_w\right|^2\right) \le \alpha_k \sigma_w^2(k), \qquad k = 1, 2, \cdots, K \tag{20}$$

where $\sigma_w^2(k)$ is the kth diagonal element of $U_y^{\#} R_w U_y$. We further restrict that the matrix H is of the form

$$H = U_y Q U_y^{\#} \tag{21}$$

where Q is a diagonal matrix.

The constraints in (20) provide noise shaping. Note that, if $\alpha_k = \lambda_y(k)/\sigma_w^2(k)$ then the spectrum of r_w is shaped exactly like the spectrum of the clean speech provided that equality is achieved in (20). We need to elaborate on (21). Consider the signal parts: y of z and $\hat{y}_s = Hy$ of $\hat{y} = Hz$. Note that, if H satisfies (21) then the eigenvectors of R_y and $R_{\hat{y}_s}$ are identical. This not only provides lesser spectral distortion between y and $\hat{y}_s$, it also provides shaping of the residual noise spectrum like the spectrum of $\hat{y}_s$. Let q_{kk} be the kth diagonal element of Q. From (20) and (21) we obtain

$$q_{kk} \le \alpha_k^{1/2}. \tag{22}$$

Now,

$$\begin{aligned}\epsilon_y &= \mathrm{tr}\big(E\left\{r_y r_y^{\#}\right\}\big)\\ &= \mathrm{tr}\big(U_y(I-Q)\Lambda_y(I-Q)U_y^{\#}\big)\\ &= \sum_{k=1}^{K} \lambda_y(k)(1-q_{kk})^2.\end{aligned} \tag{23}$$

Thus

$$q_{kk} = \min(1, \alpha_k^{1/2}) \tag{24}$$

minimizes (23) subject to constraints (22). Note that for white noise, the filter H thus obtained is identical to the one obtained in [6] and [7].

1) Description of the Algorithm: We assume that the covariance matrix, R_w, of the noise vector, w, is known. This covariance matrix can be estimated from the noise-only (silent) frames. We further assume that the noise is stationary. Let R_z

be the covariance matrix of z. Since speech and noise are independent,

$$R_z = R_y + R_w. \tag{25}$$

The covariance matrix R_z can be estimated from the noisy signal. Let $\hat{R}_z$ be an estimate of the covariance matrix of noisy speech. In Section III-E, a method for estimating this covariance matrix is presented. We estimate the covariance matrix R_y of clean speech by

$$\hat{R}_y = \hat{R}_z - R_w. \tag{26}$$

We will be using the eigenvectors of $\hat{R}_y$ to get an estimate of y. Since $\hat{R}_y$ is an estimate of the covariance matrix of y, hence it should be positive definite. But the subtraction based estimate of (26) does not ensure positive definiteness of $\hat{R}_y$. This effect is predominant mainly in noise dominated frames. Let

$$\hat{R}_y = U_y \Lambda_y U_y^{\#} \tag{27}$$

be the eigenvalue decomposition of $\hat{R}_y$. Let M be the number of eigenvalues of $\hat{R}_y$ strictly greater than zero. Let $U_y = [U_1, U_2]$, where U_1 denotes the $K \times M$ matrix of eigenvectors of $\hat{R}_y$ with positive eigenvalues, i.e.

$$U_1 = \{u_{yk} \colon \lambda_y(k) > 0\}. \tag{28}$$

Let

$$z_T = U_y^{\#} z = U_y^{\#} y + U_y^{\#} w = y_T + w_T. \tag{29}$$

The covariance matrix R_{w_T} of w_T is $U_y^{\#} R_w U_y$. Let $\sigma_{w_T}^2(k)$ be the kth diagonal element of R_{w_T}. The linear filter H used for estimating y is:

$$H = U_y Q U_y^{\#} \tag{30}$$

where Q is a diagonal matrix. As in [6] and [7], we choose the constraints

$$\alpha_k = \begin{cases} \exp\left(-\dfrac{\nu \sigma_{w_T}^2(k)}{\lambda_y(k)}\right), & k = 1, 2, \cdots, M \\ 0, & \text{otherwise,} \end{cases} \tag{31}$$

where ν is a predetermined constant. Smaller values of ν tend to leave the signal unchanged while larger values tend to suppress both the residual noise as well as the residual signal. Since $\alpha_k \le 1$, from (24) we get

$$q_{kk} = \begin{cases} \alpha_k^{1/2}, & k = 1, 2, \cdots, M \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

The estimate $\hat{y}$ is now obtained as

$$\hat{y} = Hz. \tag{33}$$

D. Noise KLT Approach for Noise Dominated Frames

Noise dominated frames have larger noise energy than clean speech energy. Thus the chances that the estimate $\hat{R}_y$ in (26) is nonpositive definite are much larger than in speech dominated frames. Moreover, noise dominated frames are either nonspeech activity frames or unvoiced speech frames. Thus, the spectrum of speech is flatter in these frames. Hence, the eigenvectors of the noise covariance matrix R_w provide a good approximation of the eigenvectors of R_y. The above observation is based on the empirical data. Let

$$R_w = U_w \Lambda_w U_w^{\#} \tag{34}$$

be the eigenvalue decomposition of R_w. Now, the filter H is obtained by minimizing

$$\epsilon_y = \mathrm{tr}\big(E\left(r_y r_y^{\#}\right)\big) \tag{35}$$

subject to

$$E\left(\left|u_{wk}^{\#} r_w\right|^2\right) \le \alpha_k \lambda_w(k) \tag{36}$$

and the restriction that the filter H is of the form

$$H = U_w Q U_w^{\#} \tag{37}$$

where Q is a diagonal matrix. The solution of the above constrained optimization problem is

$$q_{kk} = \min\left(1, \alpha_k^{1/2}\right). \tag{38}$$

Note that, if we assume that the eigenvectors of R_y and R_w are identical, then the restriction (37) can be dropped and H is obtained as a solution to the optimization problem with constraints given in (36).

1) Description of the Algorithm: Let $z_W = U_w^{\#} z$, thus

$$R_{z_W} = U_w^{\#} R_z U_w \tag{39}$$

is the covariance matrix of z_W. Define

$$\hat{R}_{y_W} = \beta R_{z_W} - \Lambda_w \tag{40}$$

where β is a constant. Note that $\beta > 1$ corresponds to under-compensation of noise while $\beta < 1$ corresponds to over-compensation. We choose $\beta = 0.8$ if the previous eight frames are noise dominated, i.e., we do over-compensation of noise since this frame contains mostly noise. Otherwise we choose $\beta = 1.3$. These values of β were determined empirically.

Let $\Phi = \mathrm{Diag}(\hat{R}_{y_W})$. Let M be the number of elements of Φ which are greater than 0. Let ϕ_k be the kth diagonal element of Φ. Let $U_w = [U_1, U_2]$, where

$$U_1 = \{u_{wk} \colon \phi_k > 0\} \tag{41}$$

and

$$U_2 = \{u_{wk} \colon \phi_k \le 0\}. \tag{42}$$

The linear filter H used for estimating y is:

$$H = U_w Q U_w^{\#}. \tag{43}$$

We choose

$$\alpha_k = \begin{cases} \exp\left(-\dfrac{2\nu \lambda_w(k)}{\phi_k}\right), & k = 1, 2, \cdots, M \\ 0, & \text{otherwise.} \end{cases} \tag{44}$$

Note that (44) differs from (31) by a factor of two in the exponent. This ensures more noise suppression and better noise shaping. Noise is more perceptible in noise dominated frames. More noise suppression decreases noise perception in such frames. Since $\alpha_k \leq 1$, from (38) we get

$$q_{kk} = \alpha_k^{1/2}. \tag{45}$$

The estimate $\hat{y}$ is now obtained as

$$\hat{y} = Hz. \tag{46}$$

E. Estimation of Covariance Matrix

Empirical Toeplitz covariance matrices are found to be quite useful in speech enhancement applications [6]. We estimate the covariance matrix from the estimates of the first K autocorrelation coefficients of noisy speech z. Let z_t denote the sampled value of z at time instant t. The kth autocorrelation coefficient at time instant t is evaluated from $K(2T-1)$ samples which include $(T-1)K$ past samples and TK future samples. Thus, R_z at time instant t is evaluated from the $K(2T-1)$-dimensional vector $[z_{t-(T-1)K+1}, \cdots, z_t, \cdots, z_{t+TK}]$. The kth autocorrelation coefficient is now calculated as

$$R_z(k) = \frac{1}{K(2T-1)} \sum_{i=1}^{K(2T-1)-k} \cdot z_{t-(T-1)K+i} z_{t-(T-1)K+i+k}. \tag{47}$$

For implementation purpose we choose $K = 40$ and $T = 4$. The noise covariance matrix R_w is also estimated on similar lines from silent speech frames. The estimation of the covariance matrices R_z and R_w utilizes a rectangular window of size 280 samples. This is similar to [6]. The use of a rectangular window maintains the second order statistics of the samples in the window.

F. Implementation Summary

The linear estimate thus obtained was applied to noisy speech sampled at rate 8 kHz. The frame size is $K = 40$ samples. The frames overlap each other by 50% (20 samples). A Hanning window was used in the overlap-and-add synthesis procedure.

IV. Performance Evaluation

The proposed signal/noise KLT based approach (SNK), the modified signal subspace approach (SS-SDC) [6] and spectral subtraction (SS) [1] were tested and compared in enhancing speech signals degraded by tank noise, helicopter noise and first-order autoregressive [AR(1)] noise with parameter 0.9 at SNR of 10, 5, and 0 dB. For SNK, the constants α_k are fixed as in (31) and (44) with $\nu = 0.5$, 2.0, 3.0, 3.5, 4.0, 5.0. 7.0, and 8.0. For SS-SDC, we fixed

$$\alpha_k = \exp\left(-\frac{\nu\sigma_w^2}{\lambda_{\tilde{y}}(k)}\right) \tag{48}$$

with $\nu = 0.5$, 1, 2, 3, 4, 5, 6, and 7. For both SNK and SS-SDC, the same analysis and synthesis methods and the same covariance matrix estimation method were used.

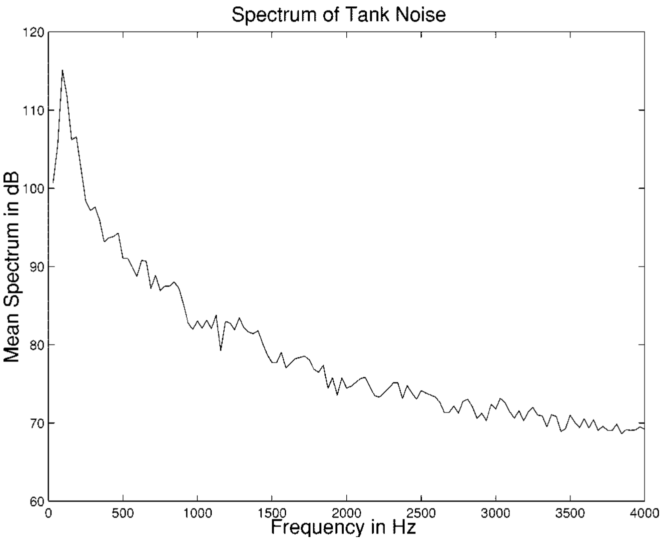

Fig. 1. The magnitude spectrum of tank noise.

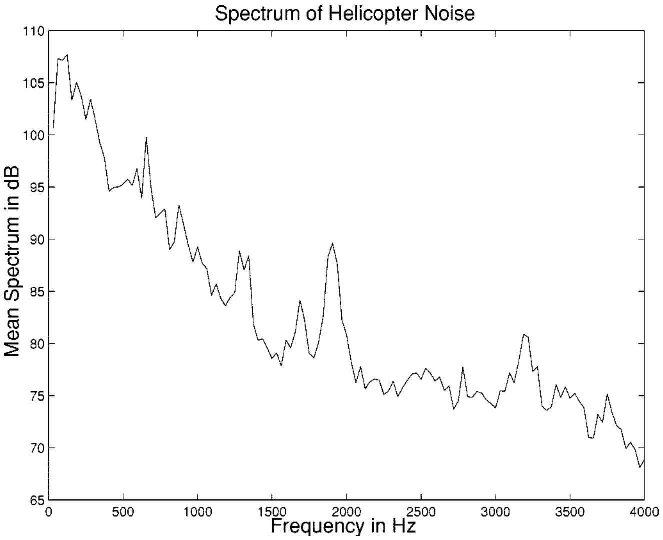

Fig. 2. The magnitude spectrum of helicopter noise.

For SS, the filter $H(w)$ is obtained by

$$H(w) = \frac{\max(0, |Z(w)| - \beta|W(w)|)}{|Z(w)|} \tag{49}$$

where $Z(w)$ is the Fourier spectrum of noisy speech z and $|W(w)|$ is the average magnitude spectrum of noise w. The sampling rate is 8 kHz. The overlapped-add-synthesis procedure with 50% overlapping Hanning window is used. The frame size is 256 samples (128 samples overlapping). The values of β used were 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, and 2.2.

In all three methods, the noise covariance matrix R_w and noise spectrum $|W(w)|$ were estimated from the same silent speech segment.

A. Objective Measure

We are interested in a distortion measure which measures both speech distortion as well as noise shaping. A two-dimensional spectral distortion measure is proposed. Let a and b be two N-dimensional vectors. To compute the spectral distortion

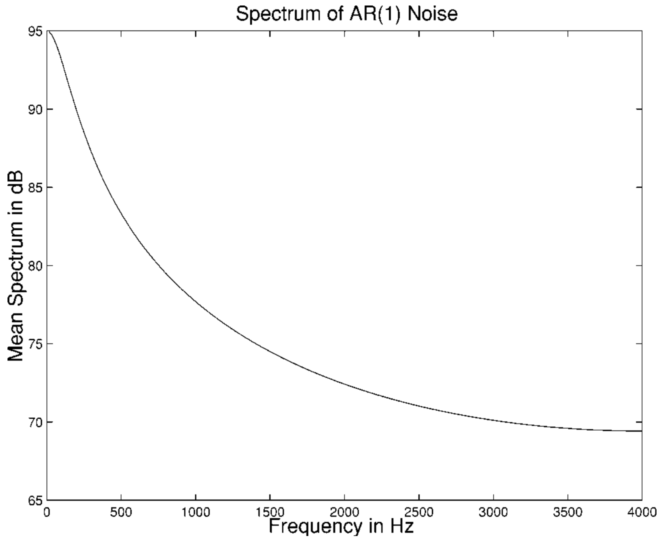

Fig. 3. The magnitude spectrum of AR(1) noise with parameter 0.9.

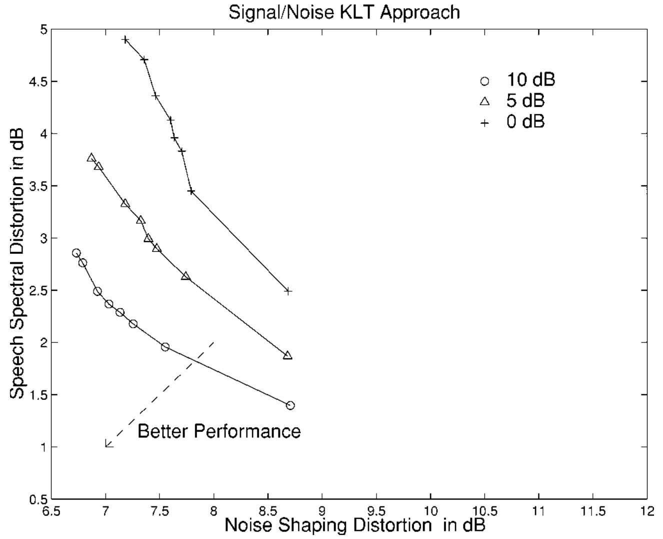

Fig. 4. Two-dimensional spectral distortion measure for the proposed estimator when the signal is degraded by tank noise at SNR = 10, 5, and 0 dB.

between a and b, each vector is first normalized to have unit energy (0 dB). A white noise vector, c, with –30 dB energy is then added to each vector. The addition of white noise prevents computation of log(0) in a logarithmic distortion measure. Let $\tilde{a} = a/||a|| + c$ and $\tilde{b} = b/||b|| + c$ be the normalized and noise-added versions of a and b, respectively. $\tilde{a}$ and $\tilde{b}$ are divided into nonoverlapping frames of length 64. A 256-point DFT is computed for each frame (192 zeros are padded). Let $\tilde{A}_p(k)$ and $\tilde{B}_p(k)$ be the kth frequency components of the pth frame of $\tilde{a}$ and $\tilde{b}$, respectively.

The spectral distortion between a and b is calculated as

$$S(a, b) = \frac{1}{P}\frac{1}{256}\sum_{i=1}^{P}\sum_{k=0}^{255} 20\left|\log|\tilde{A}_p(k)| - \log|\tilde{B}_p(k)|\right| \quad \text{(in decibels)} \tag{50}$$

where $P = \lfloor N/64 \rfloor$ is the number of frames.

Since these methods are linear, $\hat{y} = Hz = Hy + Hw$ can be decomposed into the signal part, Hy, and the noise part, Hw. Ideally, we want H to be such that $S(y, Hy)$ and $S(y, Hw)$ are both zero. Note that, when $H = I$ then $S(y, Hy) = 0$. We call $S(y, Hy)$ the *speech spectral distortion* and $S(y, Hw)$ the *noise shaping distortion*. We compare the three approaches by plotting $(S(y, Hy), S(y, Hw))$ for various values of ν and β. For comparison purposes the sentence "my cap is off for the judge" (spoken by an adult male) is used. Samples from tank, helicopter and AR(1) noise with parameter 0.9 were added to this sentence at SNR of 10, 5, and 0 dB. The samples of tank and helicopter noise are recorded from a Bradley tank and a Bell helicopter, respectively. The AR(1) noise is generated by passing artificially generated Gaussian noise through a first-order all-pole filter. Figs. 1–3 show the spectrum of tank noise, helicopter noise and AR(1) noise with parameter 0.9, respectively. Fig. 4 shows the two-dimensional spectral distortion plot for the proposed algorithm with signal

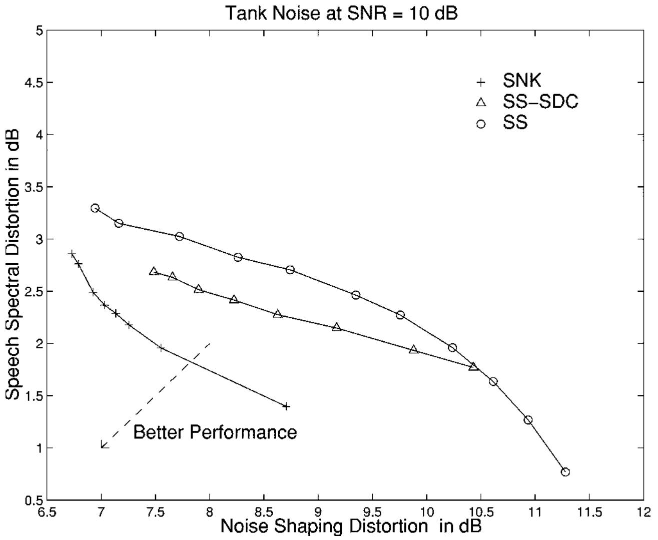

Fig. 5. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by tank noise at SNR = 10 dB.

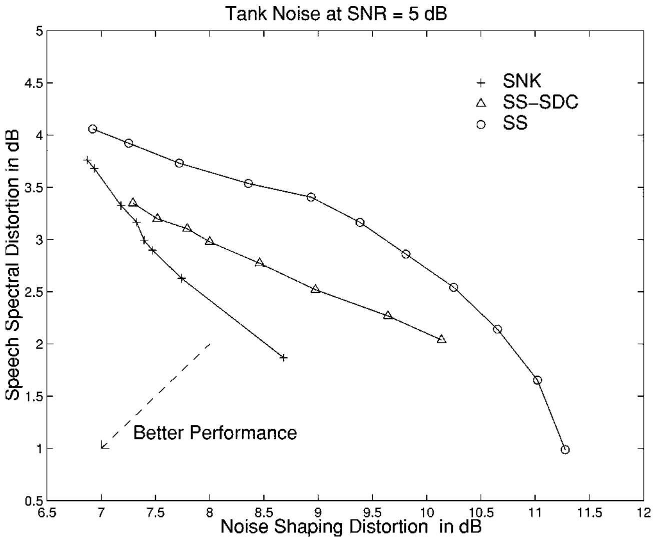

Fig. 6. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by tank noise at SNR = 5 dB.

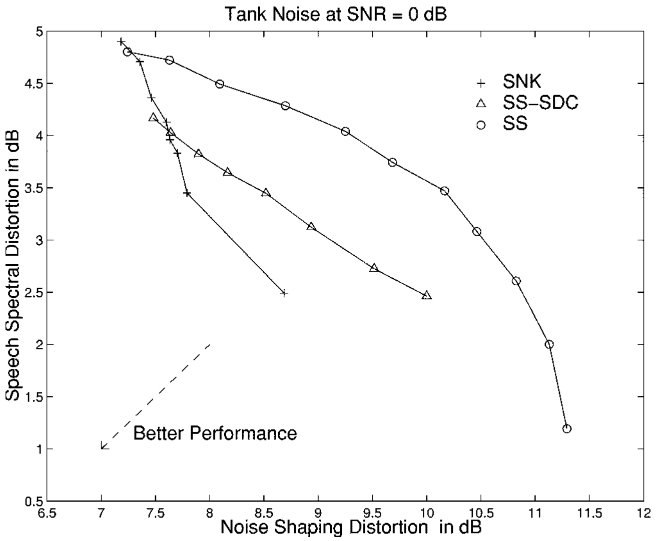

Fig. 7. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by tank noise at SNR $= 0$ dB.

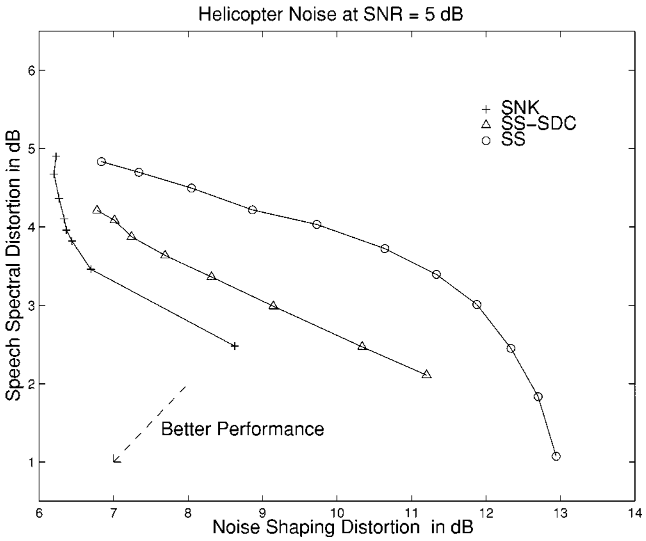

Fig. 9. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by helicopter noise at SNR = 5 dB.

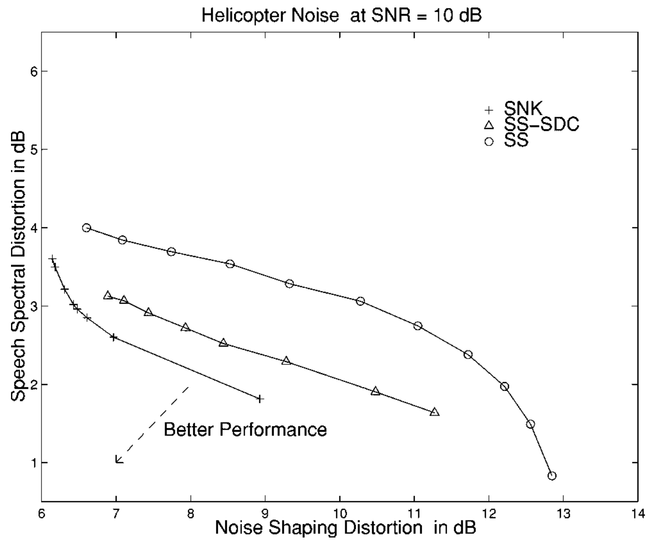

Fig. 8. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by helicopter noise at SNR $= 10$ dB.

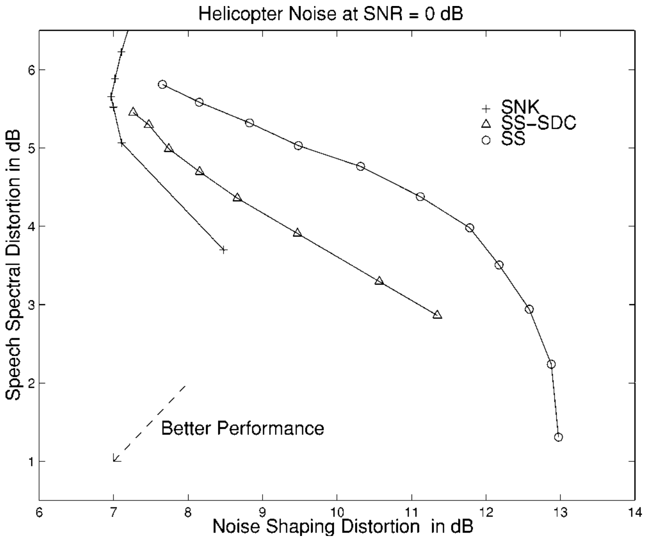

Fig. 10. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by helicopter noise at SNR = 0 dB.

degraded by tank noise at SNR = 10, 5, and 0 dB. Figs. 5–7 show the comparison between the three approaches when signal was degraded by tank noise at SNR of 10, 5, and 0 dB, respectively. Note that when ν (or β) is large, $S(y,\, Hy)$ is large while $S(y,\, Hw)$ is small. As ν (or β) is decreased, $S(y,\, Hy)$ decreases while $S(y,\, Hw)$ increases. Also, the performance is improved when the curve moves toward the lower-left corner. Figs. 8–10 show the comparison between the three approaches when signal is degraded by helicopter noise at SNR 10, 5, and 0 dB, respectively. Figs. 11–13 show these comparisons for the AR(1) noise. Note that in Figs. 10 and 13 the noise shaping distortion as well as speech spectral distortion increases for large values of ν. The reason for this observation is as follows. For low SNR and large ν, many spectral components of the signal part, Hy, and the noise part, Hw, are very small. The normalization of the signal to unit energy does not amplify these components significantly. The contribution of these components to the two-dimension measure results in this behavior.

From these figures it can be seen that the proposed SNK approach is superior to both SS-SDC [6], [7] and SS. Furthermore, SS-SDC is always superior to SS.

B. Sensitivity Analysis of Parameter Used for Frame Classification

In Section III-B, we discussed frame classification. As mentioned before if $\mathrm{tr}(Rz) > (\max(10^{\mathrm{SNR}/20}) + 0.3)\,\mathrm{tr}(R_w)$ then the given frame is classified as speech dominated, otherwise it is classified as noise dominated. This classification method requires *a priori* knowledge of SNR. Here, we study the sensitivity of the algorithm to mismatch in SNR. We use the proposed algorithm to enhance speech degraded by tank noise at SNR of 10 dB. Instead of the actual value of SNR, we use various values

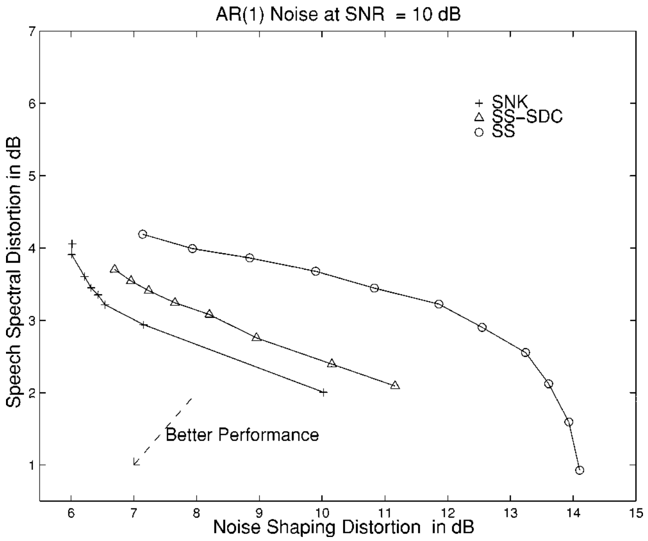

Fig. 11. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by AR(1) noise with parameter 0.9 at SNR = 10 dB.

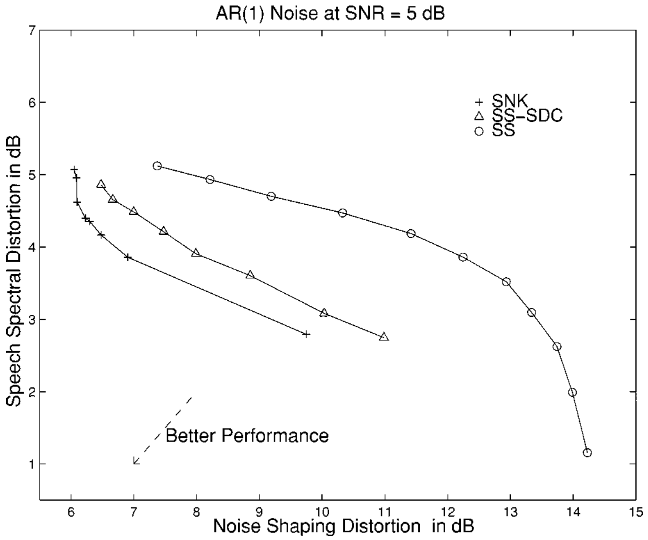

Fig. 12. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by AR(1) noise with parameter 0.9 at SNR = 5 dB.

of the estimated SNR's such as 0, 5, 10, 15, and ∞ dB for the purpose of frame classification. Fig. 14 shows the two-dimension measure for various values of the estimated SNR's. Note that the plots corresponding to 5 and 15 dB are close to the plot corresponding to the actual SNR. The plot which corresponds to 0 and ∞ dB (all frames are speech dominated) are considerably inferior to the other three. Thus the proposed algorithm is not very sensitive to the frame classification parameter as long as the estimated SNR is within ±5 dB of the true SNR.

C. Informal Listening Comparisons

The main aim of the algorithm is to mask the noise spectrum by the speech spectrum and thereby making the noise imperceptible. Informal listening comparisons indicate that both SNK and SS-SDC approaches provide noise masking.

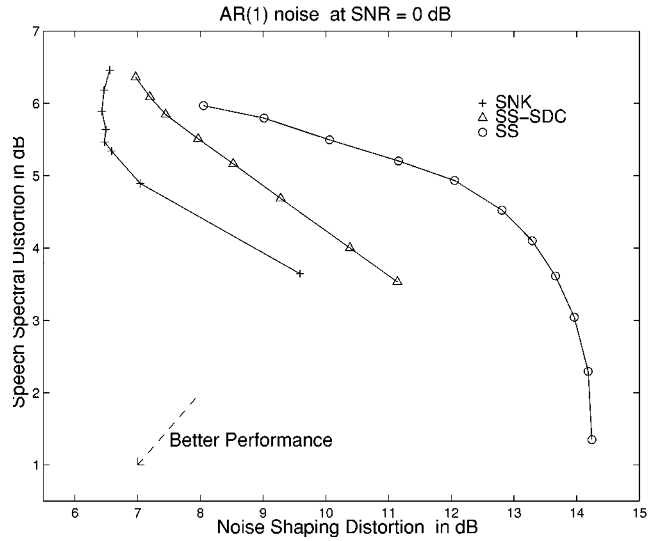

Fig. 13. Comparison between the proposed estimator (SNK), signal subspace based spectral domain constrained estimator (SS-SDC), and spectral subtraction (SS) when signal is degraded by AR(1) noise with parameter 0.9 at SNR = 0 dB.

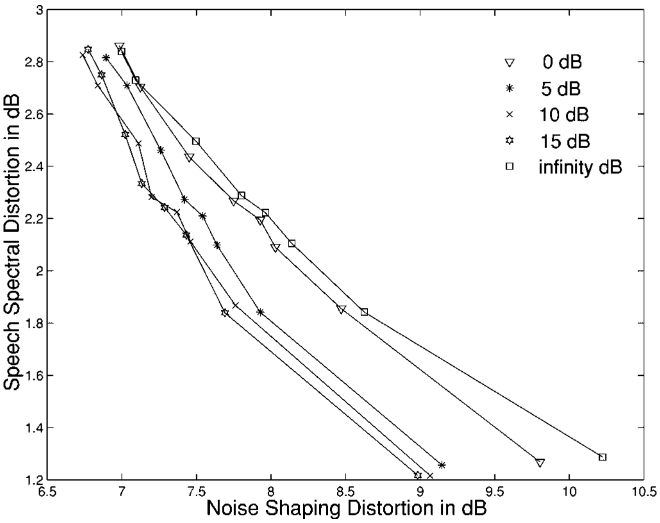

Fig. 14. Sensitivity analysis of the proposed estimator (SNK) to frame classification mismatch when signal is degraded by tank noise at true SNR = 10 dB; the SNR values used for frame classification are 0, 5B, 10 (matched), 15, and ∞ dB.

The SNK approach provides a better noise masking than the SS-SDC approach, i.e., the noise is less perceptible in the SNK approach. In case of SS-SDC, we get peeks (high energy) in the processed speech in certain frames. This might have resulted from the noise whitening in frames which have a dominant frequency (eigenvector) not present in the spectrum (covariance matrix) of noise. The SNK approach does not suffer from this problem. Both SNK and SS-SDC approaches do not suffer from the residual musical noise inherent in the SS approach. Further, the informal listening tests indicate that the noise level is also considerably reduced. The main problem with all three approaches (noticeable only at low SNR's) is the nonstationary behavior of the energy of the residual noise, i.e., the nonuniformity of the residual noise energy from frame-to-frame. When the noise energy is uniform over frames then listeners get tuned

TABLE I
NUMBER OF MATLAB FLOATING POINT OPERATIONS (IN MEGA FLOPS) FOR ENHANCING 2 S OF NOISY SPEECH BY SPECTRAL SUBTRACTION (SS), THE PROPOSED ESTIMATOR (SNK), AND SIGNAL SUBSPACE BASED SPECTRAL DOMAIN CONSTRAINED ESTIMATOR (SS-SDC)

SS	SNK	SS-SDC
39 MFlops	560 MFlops	1260 MFlops

to that noise level. Thus, though the residual noise has a much lower energy than the unprocessed noise, its nonstationarity is more bothersome to some listeners. However, listening to a high level of noise for an extended period of time may result in fatigue as well as long-term auditory damages.

D. Complexity Comparison of SS-SDC and SNK

Both SS-SDC and SNK require an estimate of the covariance matrix R_z. Further both approaches require an eigenvalue decomposition of R_z. In SS-SDC, the covariance matrix $R_{\tilde{z}}$ has to be computed from R_z. This computation requires two matrix multiplications. In SNK, computation of α_k in (31) and (44) requires two matrix multiplications but since only the diagonal elements of the matrix are needed for this computation, this can be reduced to one matrix multiplication and K vector multiplications. Further, in SS-SDC during synthesis, Hanning windowed data have to be multiplied by a whitening matrix. Once an estimate of $\tilde{y}$ is obtained, it has to be multiplied by $R_w^{1/2}$ to get $\hat{y}$. Since SNK does not involve noise whitening, these multiplications are not required in this approach. Table I shows the number of floating point operations (Flops) required in our MATLAB implementation of SNK, SS-SDC, and SS. These numbers were computed by the "flops" instruction in MATLAB. Note that SS approach is ten times faster than SNK while the SNK approach is two times faster than SS-SDC approach.

V. CONCLUSION

A signal/noise KLT based (SNK) approach for speech enhancement in colored noise is proposed. A two-dimensional spectral distortion measure is introduced. This measure indicates that the approach provides better noise shaping than the signal subspace approach and the spectral subtraction approach. Informal listening tests indicate that the algorithm does not suffer from musical noise and provides better noise masking than the signal subspace approach.

ACKNOWLEDGMENT

The authors would like to thank Prof. Y. Ephraim and the anonymous reviewers for their valuable comments.

REFERENCES

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

[2] N. Virag, "Speech enhancement based on masking properties of the auditory system," in *IEEE ICASSP*, 1995, pp. 796–799.

[3] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, Apr. 1991.

[4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, pp. 725–735, Apr. 1992.

[5] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 383–389, Sept. 1996.

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

[7] ——, "A spectrally-based signal subspace approach for speech enhancement," in *IEEE ICASSP*, 1995, pp. 804–807.

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

[9] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD: John Hopkins Univ. Press, 1996.

[10] V. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophys. J. R. Astron. Soc.*, vol. 33, pp. 347–366, 1973.

[11] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Processing*, vol. 39, pp. 1110–1121, May 1991.

[12] E. Moulines, P. Duhamel, J. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Processing*, vol. 43, pp. 516–526, Feb. 1995.

[13] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Berlin, Germany: Springer-Verlag, 1988.

Udar Mittal was born on July 12, 1971, in Dehradun, India. He received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1992, and the M.E. degree in electrical communication engineering from Indian Institute of Science, Bangalore, in 1995. He is currently pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, State University of New York, Stony Brook.

In 1992 and 1993, he was the Telecommunication Department of L&T Ltd., Mysore, India. In 1995, he was a Software Engineer with Motorola India Electronics Limited (MIEL). His research interests include speech coding, speech enhancement, joint source-channel coding, and image coding.

Nam Phamdo (M'89–SM'98) was born in Saigon, Vietnam, on February 20, 1966. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1988, 1989, and 1993, respectively.

He held a graduate fellowship with the Systems Research Center from 1988 to 1992. In 1990, he visited Nippon Telegraph and Telephone (NTT) Human Interface Laboratories, Tokyo, Japan, working on speech coding for digital cellular radio applications. He joined the Department of Electrical Engineering, State University of New York, Stony Brook, as an Assistant Professor in 1993 and became an Associate Professor in 1999. In the summer of 1995, he worked as a summer faculty member at the U.S. Army Research Laboratory, Aberdeen Proving Ground, Aberdeen, MD. His research interests are in speech coding and enhancement, joint source-channel coding, trellis coding, and turbo coding.