A Perceptually Constrained GSVD-Based Approach for Enhancing Speech Corrupted by Colored Noise

Gwo-Hwa Ju and Lin-Shan Lee, *Fellow, IEEE*

***Abstract*—The singular value decomposition (SVD)-based method for single-channel speech enhancement has been shown to be very useful when the additive noise is white. For colored noise, with this approach, one needs to whiten the noise spectrum prior to SVD-based approach and perform the inverse whitening processing afterwards. A truncated quotient SVD (QSVD)-based approach has been proposed to handle this problem and found very useful. In this paper, a generalized SVD (GSVD)-based subspace approach for speech enhancement is first extended from the concept of the truncated QSVD-based approach, in which the dimension of the signal subspace can be precisely and automatically determined for each frame of the noisy signal. But with this new approach some residual noise is still perceivable under lower signal-to-noise ratio conditions. Therefore a perceptually constrained GSVD (PCGSVD)-based approach is further proposed to incorporate the masking properties of human auditory system to make sure the undesired residual noise to be nearly un-perceivable. Closed-form solutions are obtained for both the GSVD- and PCGSVD-based enhancement approaches. Very carefully performed objective evaluations and subjective listening tests show that the PCGSVD-based approach proposed here can offer improved speech quality, intelligibility and recognition accuracy, whether the noise is stationary or nonstationary, especially when the additive noise is nonwhite.**

***Index Terms*—Auditory masking thresholds, colored noise, generalized singular value decomposition (GSVD), signal subspace, speech enhancement.**

I. Introduction

VOICE quality and intelligibility are always important for communication systems, either wired or wireless, either in human-to-human or human-to-machine interactions. In order to obtain near-transparent speech communications, for example via mobile phones, speech enhancement techniques have been employed to improve the quality and intelligibility of the noise-corrupted speech and/or the speech recognition performance. The corrupting noise sources are usually classified into additive and convolutional. The former very often dominates in real-world applications, and the spectral subtraction (SS) approach has been a very popular example solution for it [1]–[3]. To subtract the noise components from the input noisy speech, the SS algorithm has to estimate the statistics of the additive noise in frequency domain. Under low signal-to-noise ratio (SNR) conditions, a spectral flooring process is usually taken to prevent the over-subtraction situation occurred. However, all such processes very often produce some unnatural residual noise in the enhanced speech, the so-called *musical noise*, due to the inevitable random tone peaks generated in the time-frequency spectrogram. Previous studies have pointed out that this perceivable residual noise can be effectively alleviated by considering the masking effect in human auditory system [4]–[8], i.e., the residual noise will not be perceived if it is under the masking thresholds in human auditory functions.

The singular value decomposition (SVD)-based subspace approach has been found useful for noise reduction in recent years [9]–[12]. With this approach, by diagonalizing the Hankel-form matrices constructed from the noisy speech samples by the SVD, we can properly decompose the vector space for the input speech samples into two orthogonal subspaces. It assumes that the clean speech is presented only in the signal subspace whereas the additive noise spans both the signal and noise subspaces. We can thus discard the noise subspace components and reconstruct the speech signal from those of the signal subspace only. This approach was found very effective while the additive noise is white. But when the noise is not white, a reasonable approach is to whiten the noise spectrum prior to the SVD-based approach and perform the inverse whitening processing afterwards. To avoid such extra processes, a truncated quotient SVD (QSVD)-based approach was proposed [12] to perform the noise whitening with the SVD algorithm together in an integrated enhancement framework, and found very useful in handling colored noise. This truncated QSVD-based approach was then extended in this paper, in which more precise and flexible determination of the dimensions of the signal and noise subspaces became possible for each frame of the noisy signal using well-defined procedures [13]. This extended speech enhancement algorithm was referred to as generalized SVD (GSVD [14])-based approach here. In fact, similar concept using Karhunen–Loève transform (KLT)-based subspace technique have also been proposed recently for enhancing speech corrupted by colored noise [15]–[17].

Although this GSVD-based speech enhancement approach has been shown to provide better performance than the previous SVD-based approach, some *musical noise* is still perceivable in the enhanced speech under lower SNR conditions [12], [13]. This is why the auditory masking thresholds (AMTs) in human auditory functions were further integrated with the above GSVD-based algorithm to establish an improved framework for speech enhancement in this paper [18], referred to as the perceptually constrained GSVD (PCGSVD)-based approach here. Because this PCGSVD-based approach operates in

Manuscript received March 14, 2005; revised January 27, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alex Acero.

G.-H. Ju was with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, R.O.C. He is now with the Telecommunication Laboratories, Chunghwa Telecom Corporation, Ltd., Taoyuan 32617, Taiwan, R.O.C. (e-mail: jgh@cht.com.tw).

L.-S. Lee is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C. (e-mail: lslee@gate.sinica.edu.tw).

Digital Object Identifier 10.1109/TASL.2006.876868

the generalized singular domain, whereas the conventional auditory masking thresholds (AMTs) are well defined in frequency domain, the previously proposed transformation between the frequency domain and the eigen domain [19] is extended to be performed between the frequency domain and the generalized singular domain [18], with which a closed-form solution for the PCGSVD-based speech enhancement approach is obtained. Experimental results based on various objective and subjective tests (e.g., time/frequency domain evaluations, speech recognition accuracies, paired-utterance listening comparison, mean opinion score (MOS) rating, etc.) show this proposed PCGSVD-based approach can effectively alleviate the phenomenon of *musical noise* in the previous GSVD-based approach, enhance the quality and intelligibility of the processed speech, and improve the accuracy of the speech recognition system, regardless of whether the additive noise is stationary or not, especially when the noise is nonwhite.

The rest of the paper is organized as follows. The framework for the GSVD-based speech enhancement approach is first summarized in Section II. The procedures for obtaining the AMTs and transforming them to the generalized singular domain, and the proposed PCGSVD-based approach are then presented in Section III. Experiments, objective/subjective performance evaluation results, and some discussions are offered in Section IV, with the conclusions finally given in Section V. Detailed derivations of the closed-form solutions for the GSVD- and PCGSVD-based approaches are presented in the Appendix.

II. Summary of the GSVD-Based Speech Enhancement Approach

Two series of the GSVD-based approach are summarized here (GSVD-MVE and GSVD-LSE), with the difference from the previously proposed truncated QSVD-based approach [12] clearly indicated. Let y_i be the ith sample of the input noisy speech signal $\boldsymbol{y_I}$, expressed as the sum of the samples d_i and n_i of the clean speech $\boldsymbol{d_I}$ and the noise $\boldsymbol{n_I}$:

$$y_i = d_i + n_i, \quad i = 1, 2 \cdots \tag{1}$$

and the goal here is to estimate $\boldsymbol{d_I}$ from $\boldsymbol{y_I}$. Fig. 1 depicts the framework of the GSVD-based speech enhancement approach, which includes five phases as described next [13].

A. Phase (I): Framer, Nonspeech Detector, and Buffer

The input speech $\boldsymbol{y_I}$ is first segmented into overlapped frames $\mathbf{y}$ with window length M, and then the following enhancement process is repeated for each frame. A voice activity detection (VAD) algorithm is used to identify and accumulate the nonspeech frames in the input signal.

B. Phase (II): Construction of the Hankel-Form Sample Matrices

To employ the subspace concept for speech enhancement, two series of Hankel-form sample matrices of order $L \times K$, as in Fig. 2, are constructed. H_Y for each noisy speech frame $\mathbf{y}$ and $H_{\tilde{N}}$ for the latest buffered nonspeech frame $\tilde{\boldsymbol{n}}$, where $L+K-1$ equals the frame size M and in general K is much smaller than L. From (1), it is clear that the matrix H_Y can be represented as the summation of two other Hankel-form sample matrices H_D and H_N, $H_Y = H_D + H_N$, which are, respectively, constructed from the clean speech frame and the real noise frame. Under noise-free conditions, the column dimension of the matrix H_Y (in this case H_N is a zero matrix with $L \times K$ zeros and thus $H_Y = H_D$) is chosen such that H_Y is rank deficient, i.e., rank$(H_Y) < K$ [12]. Both H_D and H_N are unknown, yet H_N can be approximated by $H_{\hat{N}} \equiv \alpha H_{\tilde{N}}$, where α can be either a constant or a time-varying variable. With the GSVD algorithm as described below, we can estimate H_D and thus the clean speech frame from the matrices H_Y and $H_{\hat{N}}$. In other word, we can apply the GSVD algorithm and the subspace concept to the rank deficient least squares (LS) problem to estimate clean speech signal from the noisy observations.

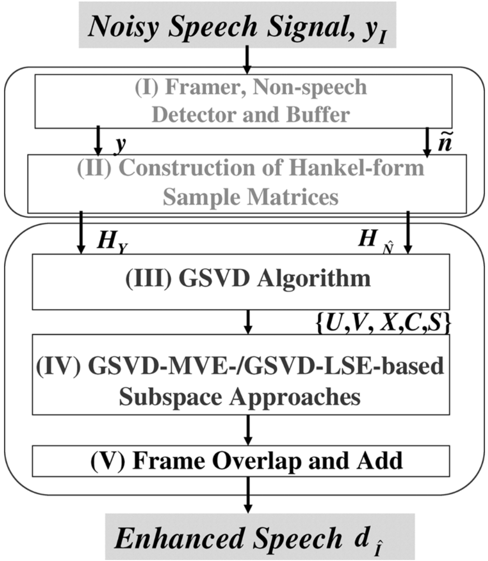

Noisy Speech Signal, y_I
(I) Framer, Non-speech Detector and Buffer
y
$\tilde{n}$
(II) Construction of Hankel-form Sample Matrices
H_Y
$H_{\hat{N}}$
(III) GSVD Algorithm
$\{U,V,X,C,S\}$
(IV) GSVD-MVE-/GSVD-LSE-based Subspace Approaches
(V) Frame Overlap and Add
Enhanced Speech $d_{\hat{I}}$

Fig. 1. Framework of the GSVD-based speech enhancement approach.

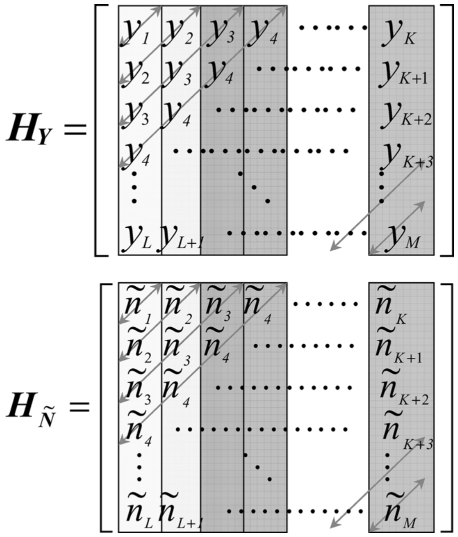

$H_Y =$
y_1 y_2 y_3 y_4 y_K
y_2 y_3 y_4 y_{K+1}
y_3 y_4 y_{K+2}
y_4 y_{K+3}
y_L y_{L+1} y_M
$H_{\tilde{N}} =$
$\tilde{n}_1$ $\tilde{n}_2$ $\tilde{n}_3$ $\tilde{n}_4$ $\tilde{n}_K$
$\tilde{n}_2$ $\tilde{n}_3$ $\tilde{n}_4$ $\tilde{n}_{K+1}$
$\tilde{n}_3$ $\tilde{n}_4$ $\tilde{n}_{K+2}$
$\tilde{n}_4$ $\tilde{n}_{K+3}$
$\tilde{n}_L$ $\tilde{n}_{L+1}$ $\tilde{n}_M$

Fig. 2. Construction of the Hankel-form sample matrices H_Y and $H_{\tilde{N}}$.

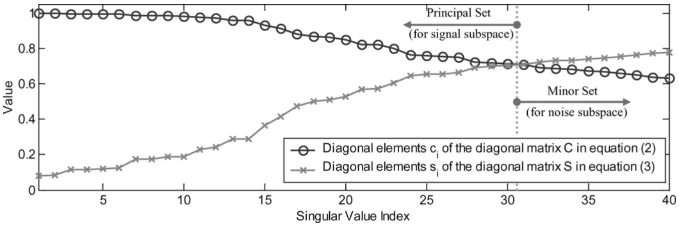


Fig. 3. Typical example of the computation results of the diagonal elements c_i (circle) and s_i (x-mark) of the matrices C and S, respectively, in (2) and (3) for a voiced frame with $K = 40$.

C. Phase (III): GSVD Algorithm

The GSVD algorithm is useful in several constrained least squares problems [14]. With GSVD, a nonsingular matrix $X \in \mathcal{R}^{K \times K}$ and two real $L \times K$ matrices U and V, whose columns are, respectively, orthonormal vectors, can be found to transform both H_Y and $H_{\hat{N}}$ into nonnegative, bounded diagonal matrices C and $S \in \mathcal{R}^{K \times K}$ simultaneously

$$\begin{aligned} U^T H_Y X &= C \\ &= \operatorname{diag}(c_1, \ldots, c_K),\ 1 \geq c_1 \geq \cdots \geq c_K \geq 0 \quad (2) \\ V^T H_{\hat{N}} X &= S \\ &= \operatorname{diag}(s_1, \ldots, s_K),\ 1 \geq s_K \geq \cdots \geq s_1 \geq 0 \quad (3) \\ &\text{subject to} \quad C^T C + S^T S = I_K \\ &\text{or} \quad c_i^2 + s_i^2 = 1,\ 1 \leq i \leq K \quad (4) \end{aligned}$$

where the superscript "T" means the transpose, the diagonal elements c_i and s_i, or the transformed components, of the matrices C, and S are arranged in descending and ascending orders, respectively, and $I_K \in \mathcal{R}^{K \times K}$ is an identity matrix. The constraint in (4) is helpful to achieve efficient numerical solutions here [14], [20], and useful in determining the precise dimension of the signal subspace of the matrices H_Y and $H_{\hat{N}}$ as presented below, which was not discussed at all in the previously proposed truncated QSVD-based speech enhancement approach. The values c_i/s_i, $i = 1, 2 \cdots K$, and the columns of the matrix X are, respectively, referred to as the generalized singular values and generalized singular vectors of the matrices H_Y and $H_{\hat{N}}$. The conventional SVD algorithm can be considered as a special case of the above GSVD algorithm if $H_{\hat{N}} = I_K$.

D. Phase (IV): GSVD-MVE- and GSVD-LSE-Based Subspace Approaches

The diagonal elements of the matrix C in (2) can be partitioned into two sets, the principal set [in which the transformed (or projected) noisy speech components obtained in (2) are dominant (i.e., for those c_i s.t. $c_i > s_i$, $i = 1, 2 \cdots m$, $m \leq K$), associated with the signal subspace of H_Y and $H_{\hat{N}}$] and the minor set [in which the transformed noise components obtained in (3) are dominant (i.e., for those s_i s.t. $s_i \geq c_i$, $i = m + 1$, $m + 2 \cdots K$), associated with the noise subspace of H_Y and $H_{\hat{N}}$]. The dimensions of the signal subspace and the noise subspace are, respectively, denoted as m and $K - m$ here, where the variable m is the number of the coefficients c_i or s_i such that $c_i > s_i$. In other words, because c_i are arranged in descending order and s_i in ascending order, with the constraints in (4), $c_i^2 + s_i^2 = 1$, we have $c_i > s_i$ for $1 \leq i \leq m$ and $s_i \geq c_i$ for $m + 1 \leq i \leq K$, and it can be shown that the value of m is proportional to the instantaneous SNR of each speech frame. In this way the dimension of the signal subspace m can be precisely, flexibly, and automatically determined for each Hankel-form matrix H_Y and therefore for each frame of the noisy speech signal. This is the major difference of the newly extended GSVD-based approach from the previously proposed truncated QSVD-based approach, in which the dimensions of the signal and noise subspaces are empirically determined considering the SNR conditions based on the concept of "parsimonious order" [12], [21]. The signal subspace for H_Y and $H_{\hat{N}}$ is then constructed by the first m row vectors of the matrix X^{-1}, whereas the rest $K - m$ rows of X^{-1} span the noise subspace of H_Y and $H_{\hat{N}}$. Fig. 3 illustrates an example of the computation results for respective diagonal elements c_i and s_i of the matrices C and S obtained in (2)–(4) for a typical example of 512-sample voiced frame corrupted by white noise at 5-dB SNR, with column size $K = 40$ for the matrices H_Y and $H_{\hat{N}}$. It can be found that in this frame of noisy speech $m = 30$, because $c_i > s_i$ for $i = 1, 2 \cdots 30$ but $c_i \leq s_i$ for $i = 31, 32 \cdots 40$, so the first 30 row vectors of the matrix X^{-1} are used to construct the signal subspace and the rest, $i = 31, 32 \cdots 40$, are used to construct the noise subspace. However, of course m can be different for different frames of noisy speech. This is how the dimensions m and $K - m$ are determined precisely and automatically for each frame of noisy speech. The minimum variance estimation (MVE) algorithm, a linear estimator with the lowest residual noise level [9], is then used to estimate the matrix H_D by finding a transformation matrix $P \in \mathcal{R}^{K \times K}$ which minimizes the Frobenius distance between the two matrices $H_Y P$ and H_D (by approximating the rank of H_D by m)

$$P^* = \underset{\substack{P \in \mathcal{R}^{K \times K} \& \\ \operatorname{Rank}(H_D) = m}}{\operatorname{argmin}} \|H_Y P - H_D\|_F^2 \quad (5)$$

where $\|\mathcal{A}\|_F^2 \equiv \sum_{i=1}^{L} \sum_{j=1}^{K} |a_{i,j}|^2$ is the Frobenius norm square of a matrix $\mathcal{A} \in \mathcal{R}^{L \times K}$, and $a_{i,j}$ is an element of the matrix $\mathcal{A}$. $P^* \in \mathcal{R}^{K \times K}$ is the estimated result of the transformation matrix P. A closed-form solution for estimating H_D based on (5) can be obtained by weighting the components of the principal set and nulling those of the minor set of the matrix C [obtained in (2)]

$$H_{\hat{D}} = H_Y P^* = U C' X^{-1} = \sum_{i=1}^{m} c_i' \boldsymbol{u}_i \check{\boldsymbol{x}}_i \quad (6)$$

where the matrices U and X are those in (2) and (3), the vectors $\boldsymbol{u}_i \in \mathcal{R}^{L \times 1}$ and $\check{\boldsymbol{x}}_i \in \mathcal{R}^{1 \times K}$ are the ith column and row, $i = 1, 2 \cdots K$, of the matrices U and X^{-1}, respectively. From (2), (3), and (6), we know that the matrices H_Y, $H_{\hat{N}}$, and $H_{\hat{D}}$ can be transformed onto the same vector space (spans by the row vectors of X^{-1}). The matrix C' is diagonal with diagonal elements c_i' given as follows:

$$c_i' = \left\{ \begin{array}{ll} c_i \left(1 - \frac{s_i^2}{c_i^2}\right), & 1 \leq i \leq m \\ 0, & m + 1 \leq i \leq K \end{array} \right\} \quad (7)$$

where c_i and s_i, $i = 1, 2 \cdots m$, are those in (2) and (3). The proof of (6) and (7) is given in the Appendix. Fig. 4 is the same

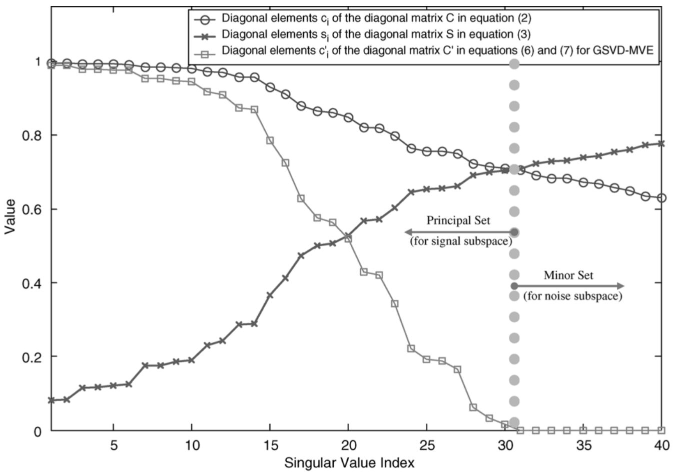


Fig. 4. Typical example of the computation results of the diagonal elements c_i (circle), s_i (x-mark), and c'_i (square, for GSVD-MVE) of the matrices C, S, and C' (for GSVD-MVE), respectively, in (2), (3), and (6) and (7) for a voiced frame with $K = 40$.

as Fig. 3 for the same example voiced frame, except here what are plotted in addition are the diagonal elements c'_i of the matrix C' obtained from (6) and (7). Apparently, it can be found that c'_i decreases monotonically for $i = 1, 2 \cdots 30$ and becomes zero for $i \geq 31$.

The estimated matrix $H_{\hat{D}}$ obtained here may not have the Hankel-form structure. We can simply average the antidiagonal elements of $H_{\hat{D}}$ to recover the Hankel-form sample matrix $H_{\bar{D}}$ and thus the enhanced speech frames, as depicted in Fig. 5 [12].

For comparison purposes, the matrix H_D can also be estimated using the least squares estimation (LSE) algorithm in the framework of the GSVD-based speech enhancement approach (GSVD-LSE). The simplest estimate of H_D, or $H_{\hat{D}}$, given H_Y, is obtained by approximating H_Y by a matrix of rank m in the least-squares sense [12], where the value of m can be obtained with the same procedure as mentioned previously in this section

$$H_{\hat{D}} = \underset{\mathrm{rank}(H_D)=m}{\mathrm{argmin}} \; \|H_Y - H_D\|_F^2. \tag{8}$$

Again with the GSVD algorithm, the solution for (8) is straightforward

$$H_{\hat{D}} = U \begin{bmatrix} C_1 & 0 \\ 0 & 0 \end{bmatrix} X^{-1} \tag{9}$$

where the diagonal matrix $C_1 \in \mathcal{R}^{m \times m}$ consists of the m most informative diagonal elements of the matrix C (principal set) obtained in (2). The reconstructed Hankel-form sample matrix $H_{\bar{D}}$ and the enhanced speech frames can be similarly obtained.

E. Phase (V): Frame Overlap and Add

Finally, the enhanced speech signal $\boldsymbol{d_{\hat{I}}}$ can be obtained by concatenating the estimated speech frames with the overlap-add method.

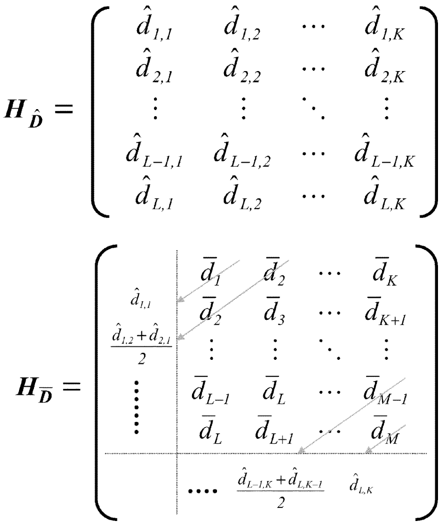

$H_{\hat{D}} =$
$\hat{d}_{1,1}$ $\hat{d}_{1,2}$ $\cdots$ $\hat{d}_{1,K}$
$\hat{d}_{2,1}$ $\hat{d}_{2,2}$ $\cdots$ $\hat{d}_{2,K}$
$\hat{d}_{L-1,1}$ $\hat{d}_{L-1,2}$ $\cdots$ $\hat{d}_{L-1,K}$
$\hat{d}_{L,1}$ $\hat{d}_{L,2}$ $\cdots$ $\hat{d}_{L,K}$
$H_{\bar{D}} =$
$\hat{d}_{1,1}$
$\frac{\hat{d}_{1,2}+\hat{d}_{2,1}}{2}$
$\bar{d}_1$ $\bar{d}_2$ $\cdots$ $\bar{d}_K$
$\bar{d}_2$ $\bar{d}_3$ $\cdots$ $\bar{d}_{K+1}$
$\bar{d}_{L-1}$ $\bar{d}_L$ $\cdots$ $\bar{d}_{M-1}$
$\bar{d}_L$ $\bar{d}_{L+1}$ $\cdots$ $\bar{d}_M$
$\frac{\hat{d}_{L-1,K}+\hat{d}_{L,K-1}}{2}$
$\hat{d}_{L,K}$

Fig. 5. Evaluation of the Hankel-form sample Matrix $H_{\bar{D}}$ from the matrix $H_{\hat{D}}$.

III. PCGSVD-Based Approach

Though the GSVD-based approach mentioned previously has been shown to provide better performance than the other popular enhancement approaches [13], some *musical noise* is still perceivable in the enhanced speech under lower SNR conditions. To obtain better sounding conditions, we further propose to integrate the masking properties of human auditory system into the GSVD-based approach to establish an improved framework for speech enhancement in this paper [18], referred to as the perceptually constrained GSVD (PCGSVD)-based approach here.

In this section, first we briefly summarize the procedure for evaluating the human auditory masking thresholds (AMTs) in frequency domain, following that we offer two series of PCGSVD-based approach (PCGSVD-MVE and PCGSVD-LSE). Furthermore, because the PCGSVD-based approach operates in the generalized singular domain, a transformation of AMTs between the frequency domain and the generalized singular domain is proposed [18], with which closed-form solutions for the PCGSVD-MVE- and PCGSVD-LSE-based subspace approaches can be obtained. Finally, in Section III-F, we discuss the influence of the scaling factor α, as defined in Section II-B for obtaining the matrix $H_{\hat{N}}$, on the performance evaluation of the proposed enhancement approaches.

A. Evaluation of the Auditory Masking Thresholds

Noise masking is a well-known psychoacoustic property of the human auditory system that has been applied with good success to speech and audio coding in order to partially or totally mask the distortion introduced in the coding processes [4], [8]. Masking effect happens when the human auditory system is incapable of distinguishing two signals close enough in time or frequency domains. The maximum allowable level of noise spectrum (or distortion spectrum) below which the distortion is not discernible by a human listener is referred to as the masking threshold. This is obtained by the minimum threshold of audibility for a given masker signal. Both temporal and simultaneous masking properties of human perception have been investigated, but here we only use the simultaneous masking properties evaluated in frequency domain in the PCGSVD-based approach proposed in this paper. The evaluation procedure for simultaneous AMTs is briefly described as follows.

The perceptible frequency range for human auditory system (20 Hz–20 kHz) is usually modeled by 25 critical bands. The magnitude square of the discrete Fourier transform (DFT) components of the clean speech signal can be summed in each critical band, and then convolved with a spreading function to consider the cross correlation between the critical bands. This spread sequence is further divided by a set of relative threshold values based on the noise-like or tone-like nature for each critical band of the input speech frame. The AMTs are finally obtained by renormalizing the above sequence to compensate for the gain modification of the convolution process, and make sure they are not below the absolute masking thresholds of human hearing [4]–[6].

B. Formulation of the PCGSVD-MVE-Based Subspace Approach

The enhancement framework of the PCGSVD-based subspace approach by using MVE algorithm (PCGSVD-MVE) is almost identical to that of the GSVD-MVE-based approach as described in the above section, except for the Phase (*IV*) in Section II-D. Therefore only the Phase (*IV*) of the framework of the PCGSVD-MVE-based approach is presented here. By incorporating the auditory masking effect into the GSVD-MVE-based subspace approach to further suppress the perceivable residual noise, the goal here is to find an optimal transformation matrix P for which not only the Frobenius distance between the two matrices $H_Y P$ and H_D is minimum, but under the constraints that the normalized energies are not greater than the transformed AMTs for the first m projections (m is the dimensionality of the signal subspace of the matrices H_Y and $H_{\hat{N}}$ here) of the residual noise signal (i.e., $H_{\hat{N}}P$), and zero for the rest $K-m$ projections. In other words, the residual noise components cannot be perceived by the human ear

$$P^* = \underset{\substack{P\in\mathcal{R}^{K\times K}\,\& \\ \mathrm{Rank}(H_D)=m}}{\operatorname{argmin}} \quad \|H_Y P - H_D\|_F^2$$
$$\text{subject to} \left\{ \begin{array}{ll} \dfrac{\left\|\boldsymbol{v}_i^T H_{\hat{N}} P\right\|^2}{\|\check{\boldsymbol{x}}_i\|^2} \le \gamma_i, & 1 \le i \le m \\ \dfrac{\left\|\boldsymbol{v}_i^T H_{\hat{N}} P\right\|^2}{\|\check{\boldsymbol{x}}_i\|^2} = 0, & m+1 \le i \le K \end{array} \right\} \tag{10}$$

where the vectors $\boldsymbol{v}_i \in \mathcal{R}^{L\times 1}$ and $\check{\boldsymbol{x}}_i \in \mathcal{R}^{1\times K}$ are, respectively, the ith column and row, $i = 1, 2 \cdots K$, of the matrices V and X^{-1} obtained in (2) and (3), and γ_i is the AMTs but transformed to the generalized singular domain, which can be evaluated by the procedures present below [18]. Everything else is the same as the procedures summarized in Section II.

C. Estimating AMTs Projected Onto the Generalized Singular Domain

Because the PCGSVD-based approach operates in the generalized singular domain, whereas the conventional AMTs are well defined in the frequency domain, the previously proposed transformation between the frequency domain and the eigen domain [19] is used here to perform the transformation between the frequency domain and the generalized singular domain [18], with which closed-form solutions for the PCGSVD-MVE- and PCGSVD-LSE-based subspace approaches can be obtained. The power spectrum of the clean speech signal is required for evaluating the AMTs in frequency domain but the clean speech is not known here. This power spectrum is estimated by the Blackman–Tukey frequency estimation technique [22] as summarized below.

With (6) and (7) in Section II-D, the $K \times K$-dimensional autocorrelation matrix $R_{\widehat{dd}}$ of the clean speech frame can be probably obtained from the matrix $H_{\hat{D}}$.

$$\begin{aligned} R_{\widehat{dd}} &\cong \frac{1}{M-K+1}(H_{\hat{D}})^T H_{\hat{D}} \\ &= \frac{1}{M-K+1} X^{-T}(C')^2 X^{-1} \end{aligned} \tag{11}$$

where M is the frame size and K is the column dimension of the Hankel-form sample matrices H_Y and $H_{\hat{N}}$. With $R_{\widehat{dd}}$ obtained in (11), the estimated power spectrum $\widehat{\Gamma} \in \mathcal{R}^{J\times 1}$ of the clean speech frame can be approximately obtained by the principal component version of the Blackman–Tukey frequency estimation with Bartlett window [22, pp. 470–471]

$$\widehat{\Gamma} \cong \frac{1}{(M-K+1)K} \sum_{i=1}^{m} \left[c_i \left(1 - \frac{s_i^2}{c_i^2} \right) \right]^2 \boldsymbol{w}_i^T \tag{12}$$

where the elements c_i and $s_i, i = 1, 2 \cdots m$ (m is the dimension of the signal subspace of the matrices H_Y and $H_{\hat{N}}$), are those in (2) and (3) respectively, $\boldsymbol{w}_i \in \mathcal{R}^{1\times J}$ is the vector for the magnitude square of the J-point DFT (J is the number of the AMTs) of the ith row of the matrix X^{-1} [i.e., $\check{\boldsymbol{x}}_i$ in (10)] in (2) and (3), and the value in the bracket of (12) is in fact the value c_i' in (7). With the estimated power spectrum $\widehat{\Gamma}$ in (12), the vector for the AMTs in frequency domain $\Theta \in \mathcal{R}^{J\times 1}$ can then be evaluated from $\widehat{\Gamma}$ by the procedure as mentioned in Section III-A. With Θ, the AMTs projected onto the generalized singular domain can be obtained with similar process to that proposed by Jabloun and Champagne [19].

$$\gamma_i = |\gamma_i'|, \; i = 1, 2 \cdots K \tag{13}$$

where γ_i' is the ith element of the vector $\gamma' \in \mathcal{R}^{K\times 1}$ given as follows:

$$\gamma' = \frac{1}{J} G \Theta \tag{14}$$

where $G \in \mathcal{R}^{K\times J}$ is a transformation matrix whose ith row $i = 1, 2 \cdots K$ is the vector for the magnitude square of the J-point DFT of the ith row vector of the matrix X in (2) and (3).

D. Solution for the PCGSVD-MVE-Based Subspace Approach

With the formulations above in (10)–(14), it can be shown that the estimation of the matrix H_D has a closed-form solution

$$H_{\hat{D}} = H_Y P^* = U C' X^{-1} \tag{15}$$

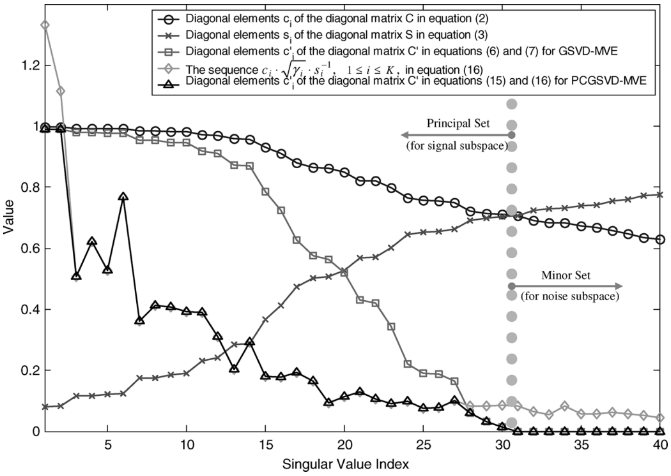


Fig. 6. Typical example of the computation results of the diagonal elements c_i (circle), s_i (x-mark), c_i' (square, for GSVD-MVE), and c_i' (triangle, for PCGSVD-MVE) of the matrices C, S, C' (GSVD-MVE), C' (PCGSVD-MVE) and the sequence $c_i(\sqrt{\gamma_i}/s_i)$, $1 \le i \le K$, (diamond), respectively, in (2), (3), (6), and (7), (15) and (16) for a voiced frame with $K = 40$.

where the diagonal elements c_i', $i = 1, 2 \cdots K$, of the nonnegative diagonal matrix C' are

$$c_i' = \begin{Bmatrix} \min\left[c_i\left(1 - \frac{s_i^2}{c_i^2}\right), c_i\frac{\sqrt{\gamma_i}}{s_i}\right], & 1 \le i \le m \\ 0, & m+1 \le i \le K \end{Bmatrix} \tag{16}$$

respectively for the principal set (c_i', $1 \le i \le m$) and the minor set (c_i', $m+1 \le i \le K$) of the matrix C'. The result in (15) and (16) is derived detailed in the Appendix. From (16), we notice that for the components of the principal set of the matrix C', the singular values c_i' of the matrix $H_{\hat{D}}$ are the smaller value of $c_i(1 - (s_i^2/c_i^2))$ and $c_i(\sqrt{\gamma_i}/s_i)$. The former term is in fact the solution for the GSVD-MVE-based approach as in (7), which is chosen when the normalized projection of the residual noise signal onto the signal subspace of the matrices H_Y and $H_{\hat{N}}$ is below the value of the corresponding transformed AMT [as described in the first constraint of (10)], and thus this projected residual noise component cannot be perceived. Otherwise, the $c_i(\sqrt{\gamma_i}/s_i)$ term in (16) will be chosen. Note that $c_i(\sqrt{\gamma_i}/s_i)$ is proportional to the square root of the ith transformed AMT $\sqrt{\gamma_i}$, but inversely proportional to s_i which is arranged to be increasing with index i as in (3), therefore γ_i is more dominant in this second term for smaller i, which corresponds to more informative signal subspace components. This AMT-related term will be chosen when the ith constraint $1 \le i \le m$ of (10) activates [i.e., equal sign of the first constraint of (10) holds]. In this case, the residual noise component is nearly unperceivable. Fig. 6 is the same as Figs. 3 and 4 for the same example voiced frame, except here what are plotted in addition are the sequence $c_i(\sqrt{\gamma_i}/s_i)$, $1 \le i \le K$, in (16) and the diagonal elements c_i' of the matrix C' for PCGSVD-MVE in (15) and (16). It can be found that for $3 \le i \le 27$ the diagonal elements c_i' for PCGSVD-MVE are upper bounded by $c_i(\sqrt{\gamma_i}/s_i)$, while for $i \ge 31$ they are all zero.

E. PCGSVD-LSE-Based Subspace Approach

For comparison purposes, the PCGSVD-LSE-based subspace approach can be similarly formulated as follows:

$$P^* = \underset{\substack{P \in \mathcal{R}^{K \times K} \,\& \\ \mathrm{Rank}(H_D) = m}}{\operatorname{argmin}} \|H_Y - H_D\|_F^2$$

$$\text{subject to} \begin{Bmatrix} \frac{\left\|\boldsymbol{v}_i^T H_{\hat{N}} P\right\|^2}{\|\tilde{\boldsymbol{x}}_i\|^2} \le \gamma_i, & 1 \le i \le m \\ \frac{\left\|\boldsymbol{v}_i^T H_{\hat{N}} P\right\|^2}{\|\tilde{\boldsymbol{x}}_i\|^2} = 0, & m+1 \le i \le K \end{Bmatrix}. \tag{17}$$

The solution for (17) is similar to (15) and (16)

$$H_{\hat{D}} = H_Y P^* = UC'X^{-1} \tag{18}$$

where the elements c_i', $i = 1, 2 \cdots K$ of the nonnegative diagonal matrix C' are

$$c_i' = \begin{Bmatrix} \min\left[c_i, c_i\frac{\sqrt{\gamma_i}}{s_i}\right], & 1 \le i \le m \\ 0, & m+1 \le i \le K \end{Bmatrix}. \tag{19}$$

F. Estimation of the Scaling Factor α

It is well known that the entropy is a metric of uncertainty for random variables. In the preliminary study, it was found that by setting the value of the scaling factor α, to be used in estimating the matrix $H_{\hat{N}}$ from $H_{\tilde{N}}$ as specified in Section II-B, to be inversely proportional to the spectral entropy of the additive noise in some sense can improve the performances of the PCGSVD-MVE- and PCGSVD-LSE-based approaches. The formulation for the spectral entropy $\mathbf{H}$ is as follows [23]:

$$\mathbf{H} \equiv -\sum_{j=1}^{J} p_j \log p_j \tag{20}$$

where J is the DFT size and p_j, $j = 1, 2 \cdots J$ is the probability density function obtained by normalizing the spectral energy of the jth frequency component $\widehat{N}(j)$ of the estimated noise over all frequency components

$$p_j \equiv \frac{\left|\widehat{N}(j)\right|^2}{\sum_{w=1}^{J}\left|\widehat{N}(w)\right|^2} \qquad 1 \le j \le J. \tag{21}$$

For the PCGSVD-based approach, with smaller value of α (e.g., $\alpha < 1$), less signal distortion was observed in the tests for the estimated speech, especially for the low-energy frequency components, but that also led to increased residual noise. So, there is a tradeoff in choosing the value of α given the best performance. We found that the best value for α has to be inversely proportioned to the spectral flatness of the additive noise in some sense, which is reasonably well measured by the spectral entropy as defined above in (20) and (21). Moreover, if the additive noise is stationary, the evaluation result for the spectral entropy ought to be invariant whether the noise level is high or low. However, changing the value of α does not significantly affect the performances of the GSVD-MVE- and GSVD-LSE-based approaches.

IV. Experiments and Performance Evaluation

The experimental environment was as follows. We randomly selected 30 clean utterances produced by two females and two males from the TIMIT speech corpus with sampling rate 16 kHz for testing [24]. Four types of noise source, "White," "Volvo-car," "Babble (speech-like)," and "Factory," chosen from NOISEX-92 database [25], were artificially added to the test speech with resulting SNR ranged from 20 to -10 dB with 5-dB step size for evaluation. The Babble and Factory noise sources are nonstationary whereas the Volvo-car noise is stationary; all of the three are not white. The following parameter settings were for the GSVD-/PCGSVD-based approaches. The rectangular window was used for the framer as in Section II-A due to its better performance. The frame size M was 32 ms (512 samples) with 50% frame overlap. The value of J, the number of AMTs in Section III-A, was the same as the frame size M. The value for the scaling factor α as discussed in Section III-F, was set to 1.0 for the GSVD-MVE- and GSVD-LSE-based approaches, and to 0.6, 1.15, 0.9, and 0.9, respectively, for the White, Volvo, Babble, and Factory noise cases for the PCGSVD-MVE- and PCGSVD-LSE-based approaches [α was roughly estimated (via the spectral entropy method mentioned in Section III-F) from the first two nonspeech frames of a typical utterance in each respective case]. The row and column sizes (L and K) of the two Hankel-form sample matrices H_Y and $H_{\hat{N}}$ were 473 and 40, respectively. The column dimension of 40 for the Hankel-form matrices was sufficient for having the matrix H_Y to be rank deficient under noise-free condition, and it was adequate for us to choose a boundary for the signal and noise subspaces.

For comparison, we also implemented the conventional spectral subtraction algorithm in the power spectral domain (abbreviated as PSS) [2], a modified version of PSS using the auditory masking effect in the enhancement process (abbreviated as PSS-AMT) [6], and a perceptual subspace approach previously proposed, in which a transformation from frequency domain to eigen domain was used to obtain a perceptual upper bound for the residual noise to be applied with the subspace concept (abbreviated as PKLT), which was shown to offer competitive performance as the conventional KLT-based subspace approach in subjective listening tests [19]. The VAD algorithm recommended by ETSI-EAFE for frame-dropping (referred to as ETSI-EAFE VAD here) [26] was employed in all these algorithms under concern for detecting the nonspeech frames and estimating the noise statistics. Compared with other popular VAD algorithms, this algorithm was reported to have low false alarm rate [27]. We also forced the first two frames of each utterance as silence frames for estimating the noise statistics initially.

Equation (22) formulates the processes of the PSS and the PSS-AMT algorithms in frequency domain

$$\left|\widehat{D}(w)\right|^2 = \left\{ \begin{array}{c} \left(\begin{array}{c} |Y(w)|^2 - \varphi(w)\cdot\xi\left[\left|\widehat{N}(w)\right|^2\right] \\ if \frac{|Y(w)|^2}{\xi\left[\left|\widehat{N}(w)\right|^2\right]} > [\varphi(w)+\rho(w)] \end{array} \right) \\ \rho(w)\cdot\xi\left[\left|\widehat{N}(w)\right|^2\right], \text{ otherwise} \end{array} \right\}$$

$$\angle\widehat{D}(w) = \angle Y(w), \quad 1 \le w \le J \tag{22}$$

where w, $w = 1, 2 \cdots J$, is the frequency index, the symbol "$\angle$" denotes the phase, the value of J, or the DFT size, was set to 512, $|Y(w)|^2$ and $|\widehat{D}(w)|^2$ are, respectively, the power spectrum of the noisy speech and enhanced speech, $\xi[|\widehat{N}(w)|^2]$ is the smoothed power spectrum of the estimated noise, and $\varphi(w)$ and $\rho(w)$ are the weighting factor and flooring coefficient respectively. For PSS, $\varphi(w)$ was set to 1.6 and $\rho(w)$ to 0.15, whereas for PSS-AMT, $\varphi(w)$ and $\rho(w)$ are both functions of AMTs. The input noisy speech signal $\boldsymbol{y}_I$ was segmented into overlapped frames via Hamming window and transformed to the frequency domain. The frame size and shift for both the PSS and PSS-AMT were 32 ms (512 samples) and 16 ms (256 samples), respectively. Each detected silence frame from the ETSI-EAFE VAD algorithm was employed to update the noise statistics in frequency domain

$$\begin{aligned} \xi\left[\left|\widehat{N}(w)\right|^2\right] = \varepsilon\cdot\xi\left[\left|\widehat{N}(w)\right|^2\right]_{\text{old}} \\ +(1-\varepsilon)\cdot|Y(w)|^2, \qquad 1 \le w \le J \end{aligned} \tag{23}$$

where $\xi[|\widehat{N}(w)|^2]_{\text{old}}$ is the previously estimated version, and ε is the smoothing factor. For both PSS and PSS-AMT, ε was set to 0.9. This updating process was carried out whenever a new silence frame was detected. For PKLT, the frame size and shift were 32 ms (512 samples) and 2 ms (32 samples), respectively, for computation load reduction. This setting was verified to behave almost as well as the 16 ms (256 samples) of frame updating rate in preliminary informal tests. Various objective and subjective measures were used to evaluate the different approaches as given next.

A. Segmental Signal-to-Noise Ratio

As it is well known, the segmental SNR (SegSNR) measure is more accurate in indicating the speech distortion than the global SNR. The SegSNR is measured by computing the SNR (in decibels) for each of the frames and averaging these SNR values over the entire utterance [28]. To emphasize the processing effect of the estimated speech signal, we manually exclude the nonspeech segments of the test speech in the following SegSNR and SegLSD evaluations. The SegSNR measure used here had a segment length of 256. The results are illustrated in Fig. 7, in which the various abbreviations "*Noisy*" (original noisy speech), "*PSS*" (power spectral subtraction), "*PSS-AMT*" (modified version of *PSS* with AMTs), "*PKLT*" (the previously proposed perceptual subspace approach [19]), and "*GSVD-LSE*," "*GSVD-MVE*," "*PCGSVD-LSE*," and "*PCGSVD-MVE*" (as proposed in this paper) are used to denote different tests. From Fig. 7(a) for the White noise case, we see for almost every case of SNR, *GSVD-MVE* obtained the best results. However, for the colored noise cases [Fig. 7(b)–(d)], *PCGSVD-MVE* outperformed *GSVD-MVE* and other approaches in many cases, especially when the input SNR was low (e.g., <10 dB). We also see that *GSVD-LSE* was not as good as that of *GSVD-MVE* in most cases, so was that of *PCGSVD-LSE* as compared with *PCGSVD-MVE*. This can be understood from the closed-form solution of *GSVD-LSE* (9), in which all the information (including clean speech and corrupting noise source) were kept in the signal subspace of H_Y and $H_{\hat{N}}$, and similarly for *PCGSVD-LSE* versus

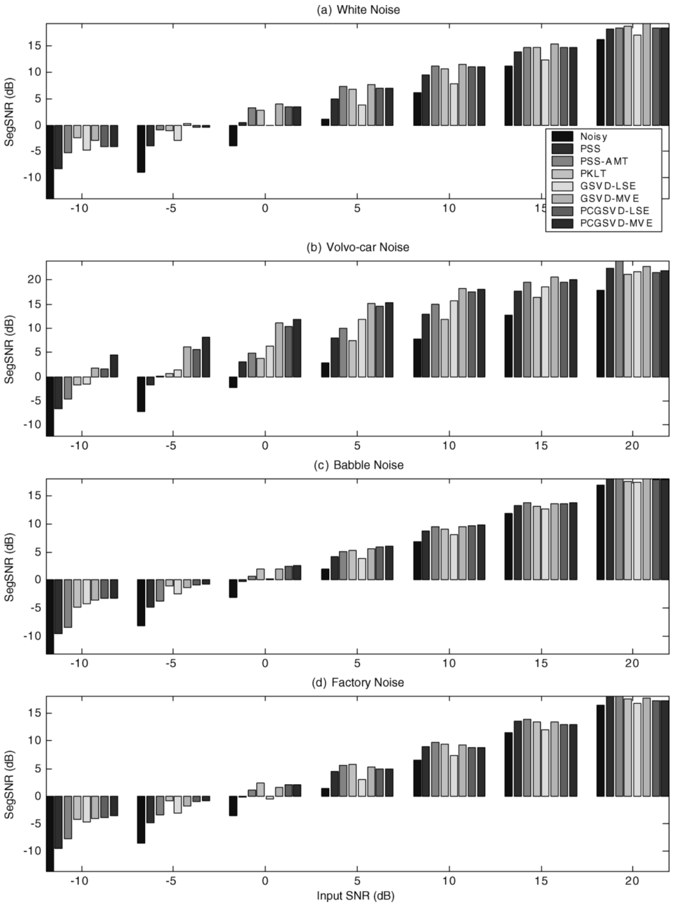


Fig. 7. Segmental SNR measures for (a) White noise, (b) Volvo-car noise, (c) Babble noise, and (d) Factory noise at different SNR values.

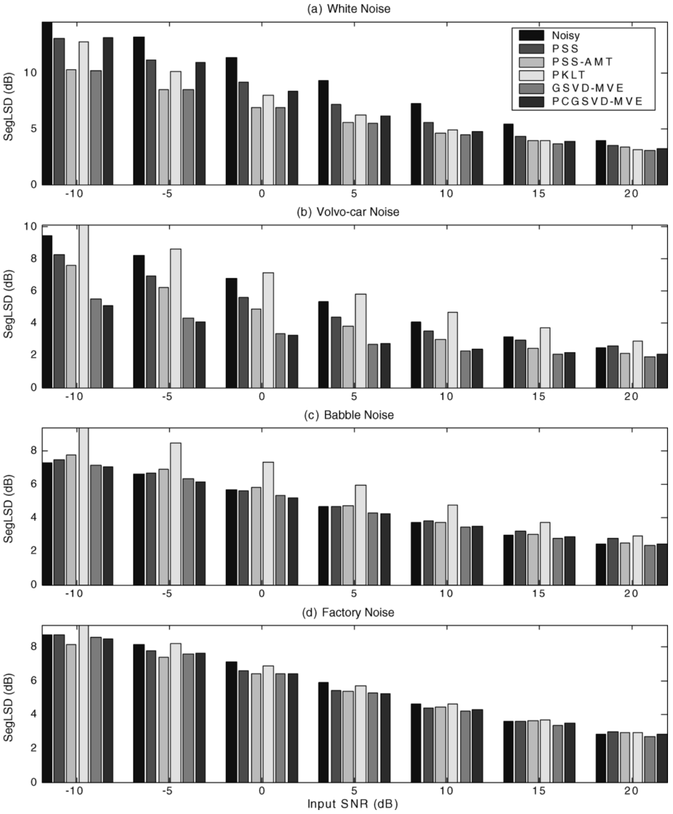


Fig. 8. Segmental log spectral distance measures for (a) White noise, (b) Volvo-car noise, (c) Babble noise, and (d) Factory noise at different SNR values.

PCGSVD-MVE. This observation remains true for all the following subjective measures and objective listening tests. Therefore, we excluded the results of *GSVD-LSE* and *PCGSVD-LSE* evaluations in the following experiments.

B. Segmental Log Spectral Distance Measures

The segmental log spectral distance (SegLSD) measure is formulated as follows:

$$\text{SegLSD} \equiv \frac{20}{TJ}\sum_{t=1}^{T}\left[\sum_{w=1}^{J}\left|\log_{10}|D(w,t)| - \log_{10}\left|\widehat{D}(w,t)\right|\right|\right] \tag{24}$$

where $D(w,t)$ and $\widehat{D}(w,t)$ are, respectively, the wth spectral component, $w = 1, 2 \cdots J$, of the tth frame (totally T nonsilence frames) of clean and enhanced utterances. J was set to 512. Fig. 8 depicts the SegLSD measures. From Fig. 8(a) for the White noise case, *GSVD-MVE* offered the smallest SegLSD, whereas *PSS-AMT* offered satisfactory results as well. However, for the nonwhite noise cases in Fig. 8(b)–(d), *PCGSVD-MVE* actually behaved the best on average. This result indicates the proposed PCGSVD-based approach algorithm generated relatively less spectral distortion in frequency domain if the additive noise did not exist across the entire spectrum.

C. English Phoneme/Digit Recognition Accuracy

The third measure we adopted was the English phonemic accuracy obtained in the free-phoneme decoding (without lexicon and language model) for the 30 test utterances mentioned previously in this section, which can be used to test the intelligibility of the estimated speech. The acoustic model used here consisted of 48 left-to-right continuous hidden Markov models (CHMMs) for the 48 context-independent phoneme units [29], trained from 4-h TIMIT speech corpus and most of the CHMMs included five states with two nonemitting and each emitting state consisted of eight Gaussian mixtures. The total number of Gaussian mixtures was about 1100. The dimension of acoustic feature vectors was 39; including 12 mel-frequency cepstrum coefficients (MFCCs) and the normalized log energy, plus the first and second derivatives. The frame size for obtaining the acoustic features was 30 ms (480 samples) with 10 ms (160 samples) of shift. The HTK [30] was used for feature extraction, acoustic model training, and recognizer in this experiment. Without dropping the silence frames in the recognition phase, the baseline phoneme accuracy for clean speech for the 30 test sentences was 54.48%.

Fig. 9 shows the accuracy results of the phoneme recognition. In Fig. 9(a), all of the six approaches improved the recognition accuracy for the White noise case. For the colored and/or nonstationary noise situations in Fig. 9(b)–(d); however, again *PCGSVD-MVE* outperformed the other enhancement algorithms in most cases, especially for the Volvo-car and Factory noise cases. Extra tests verified that the recognition

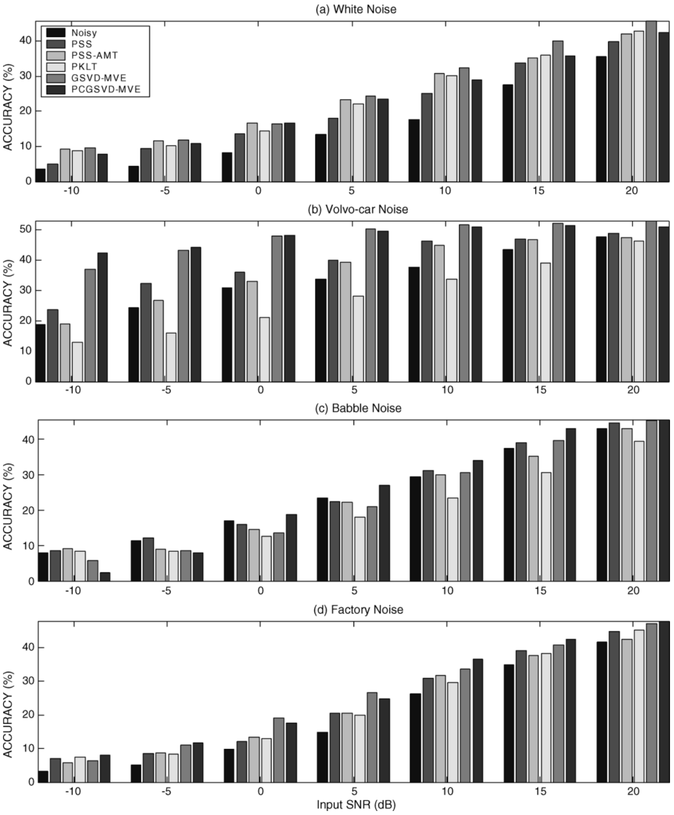


Fig. 9. TIMIT phoneme recognition accuracies for (a) White noise, (b) Volvo-car noise, (c) babble noise, and (d) factory noise at different SNR values.

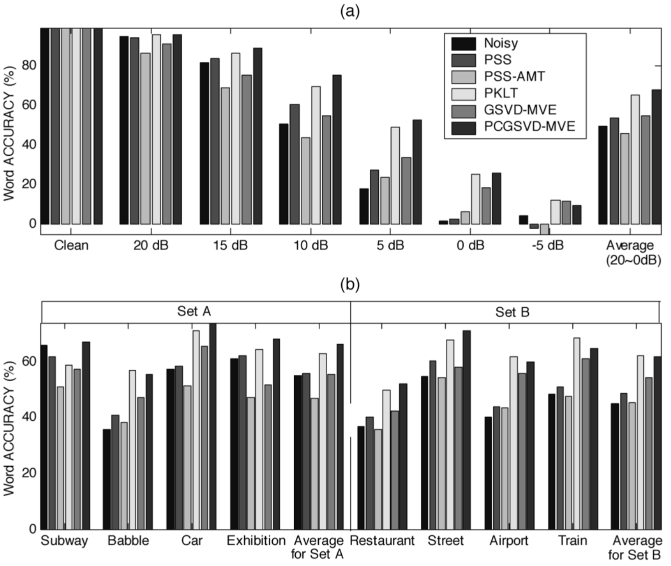


Fig. 10. AURORA2 word accuracies for clean condition training and test sets A and B for (a) different SNR values but averaged over all noise types. (b) Different types of noise but averaged over all SNR values.

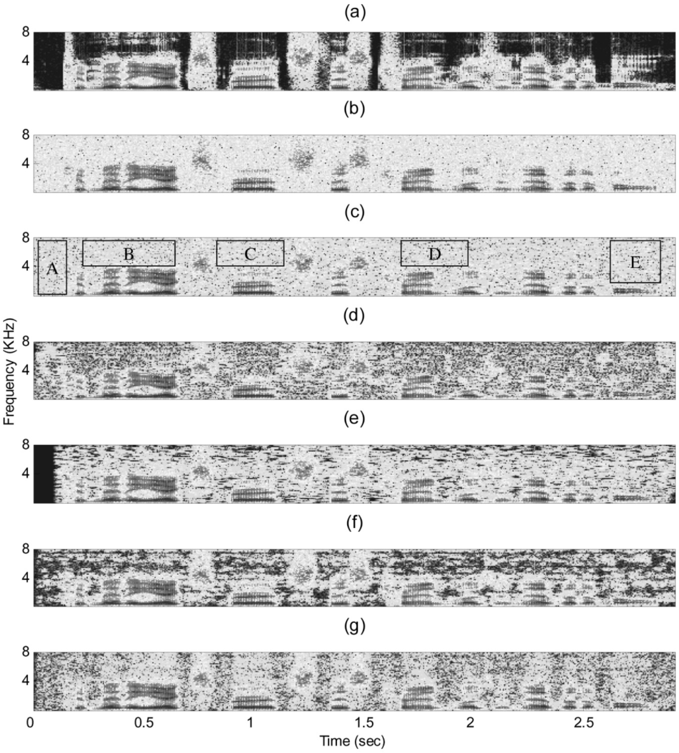


Fig. 11. Spectrogram plots for a typical utterance corrupted by White noise at 10 dB SNR. (a) Clean. (b) Noisy (10-dB White). (c) PSS. (d) PSS-AMT. (e) PKLT. (f) GSVD-MVE. (g) PCGSVD-MVE.

performance could be further improved if the training speech corpus for the acoustic model can be similarly processed *a priori*, but not reported here to save the space.

We also carried out the English digit recognition under AURORA2 testing environment [31]. We adopted the first 200 shortest utterances out of the total of 1001 for test sets A and B of AURORA2 corpus under the clean training condition. We excluded test set C here because it includes channel distortion which is not handled here. The experimental setup was identical to those mentioned previously, except here the sampling rate for the AURORA2 task was 8 kHz and the CHMMs consisted of 11 digit units (0–9 plus OH) and a silence model, each digit HMM contained 16 emitting states and each state comprised two Gaussian mixtures. The value for the scaling factor α for obtaining the matrix $H_{\hat{N}}$ mentioned in Section II-B was fixed as 0.9 for PCGSVD-based approach (because most of the noise sources in the test sets are nonstationary, the average of the α values obtained as mentioned above for all different types of noise in test sets A and B was specified). Fig. 10(a)-(b) reveals the recognition results for different SNR values (but averaged over all noise types) and different types of noise (but averaged over all SNR values), respectively. Similar to that of recognition accuracy measures on TIMIT for colored noise cases in Fig. 9(b)–(d), *PCGSVD-MVE* outperformed the other enhancement algorithms almost in every case. Compared with *Noisy*, on average 24.61% and 30.66% word error rate reduction were achieved by the *PCGSVD-MVE* approach for test sets A and B, respectively. The above results indicate the proposed *PCGSVD-MVE* could be useful as a noise removal preprocessor for a speech recognition system.

D. Spectrogram and Time Domain Waveform Plots

Figs. 11 and 12 are, respectively, the time-frequency domain spectrogram plots and the time domain waveforms for the various versions of a test utterance, "*To many experts, this trend was inevitable.*", produced by a male speaker, corrupted by the White noise at 10-dB SNR. From Fig. 11(c), we can see in

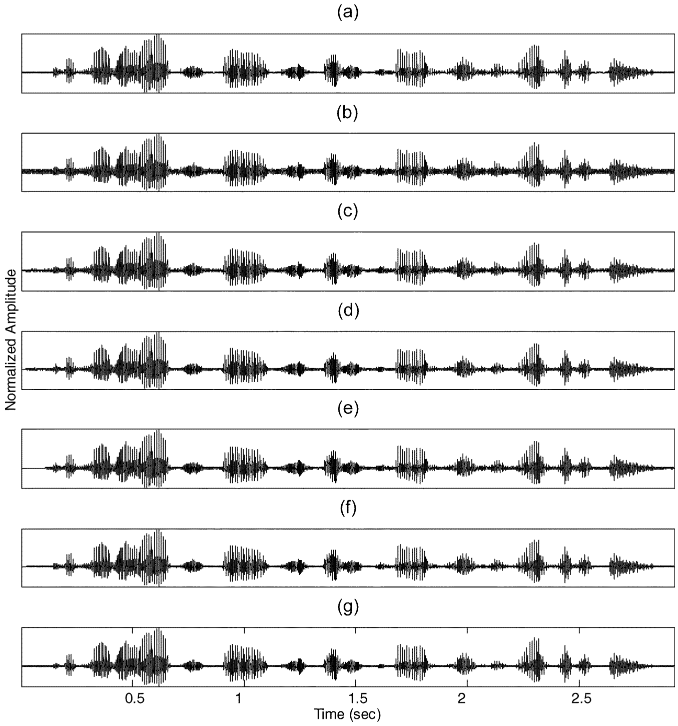


Fig. 12. Waveforms for a typical utterance contaminated by White noise at 10-dB SNR. (a) Clean. (b) Noisy (10-dB White). (c) PSS. (d) PSS-AMT. (e) PKLT. (f) GSVD-MVE. (g) PCGSVD-MVE.

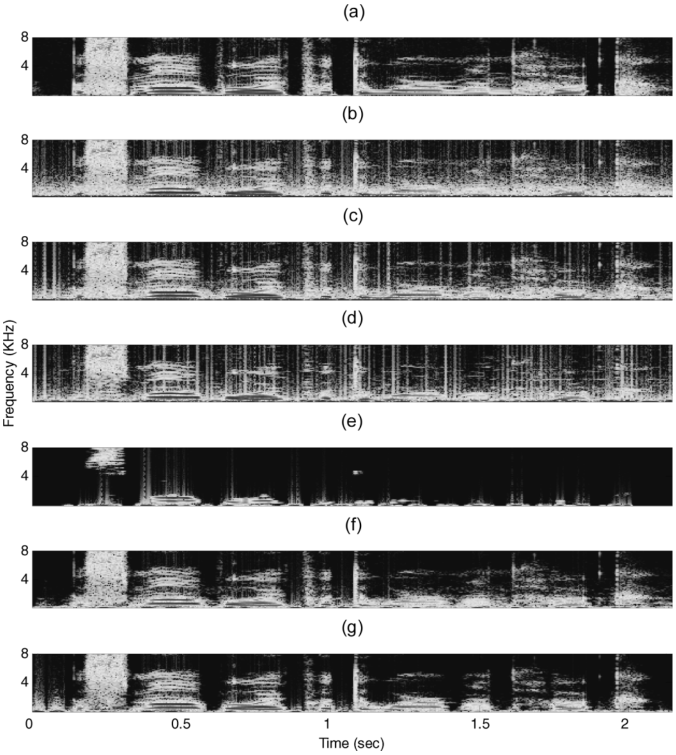


Fig. 13. Spectrogram plots for a typical utterance corrupted by Volvo-car noise at −10-dB SNR. (a) Clean. (b) Noisy (−10 dB Volvo). (c) PSS. (d) PSS-AMT. (e) PKLT. (f) GSVD-MVE. (g) PCGSVD-MVE.

the spectrogram of the *PSS* processed speech some undesired random tone peaks present in the nonspeech regions (e.g., region A) and low-energy, noise-like speech segments (e.g., segments B, C, D, and E), compared with the clean speech version in Fig. 11(a), which are perceivable *musical noise*. This phenomenon was improved in the version of *PSS-AMT*, *PKLT*, and *GSVD-MVE* processed utterances as shown in Fig. 11(d)–(f), respectively, although it was found in parallel informal listening tests that the residual noise was still quite perceivable. It is further evident in Fig. 11(g) that almost the same detailed information of the speech spectrum were recovered by *PCGSVD-MVE* as compared to that in Fig. 11(f) by *GSVD-MVE*, but with much less random tone peaks present in the silence segments and low-energy speech regions. Though some residual noise still occurred in the *PCGSVD-MVE* processed speech for the White noise case, this is due to the de-emphasis of the estimated noise constructed Hankel-form matrix $H_{\hat{N}}$ by the parameter α (0.6 in this case). However, by the parallel informally subjective listing tests, in which many subjects agreed that the *musical noise* was less perceivable for the utterances processed by *PCGSVD-MVE* than those by *GSVD-MVE* and other approaches of interest. Another test utterance, "*The small boy put the worm on the hook.*", produced by a female speaker was repeated in the same tests, except the additive noise signal was the Volvo-car noise at −10-dB SNR, and the results are in Figs. 13 and 14. Again, *PCGSVD-MVE* kept most of the original speech information and can eliminate most of the residual noise existing in the spectrogram of the enhanced speech of *PSS* and *PSS-AMT* processed speech. Furthermore, from Figs. 13 and 14, we can see that the high-frequency components in the spectrogram of *PKLT* processed utterance are seriously attenuated, and the waveform is quite different from the original speech, which means under highly noisy environments, *PKLT* not only eliminates the noise signal but hurts the signal itself as well.

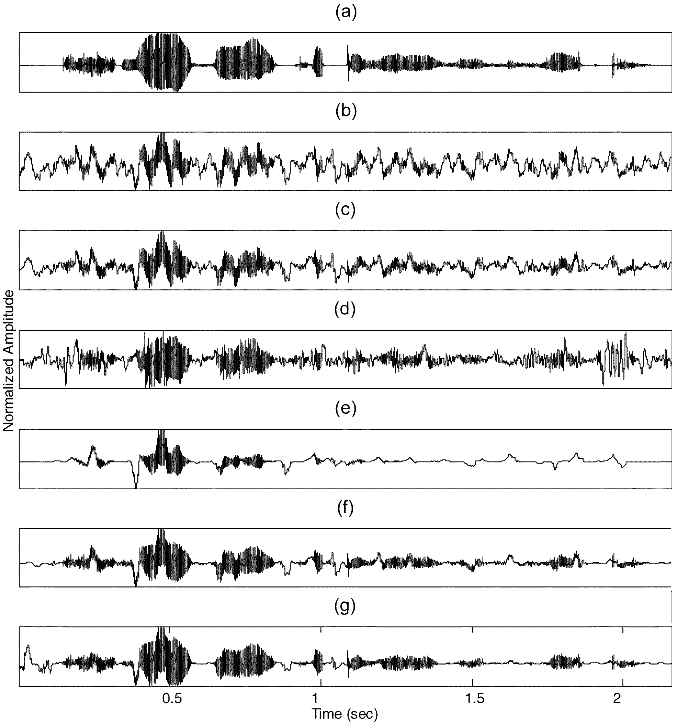


Fig. 14. Waveforms for a typical utterance contaminated by Volvo-car noise at −10-dB SNR. (a) Clean. (b) Noisy (−10-dB Volvo). (c) PSS. (d) PSS-AMT. (e) PKLT. (f) GSVD-MVE. (g) PCGSVD-MVE.

TABLE I
PERCENTAGE (%) OF SUBJECTS WHO PREFERRED THE UTTERANCES PROCESSED BY *PCGSVD-MVE* AS COMPARED TO *NOISY* AND THOSE PROCESSED BY *PSS*, *PSS-AMT*, *PKLT*, AND *GSVD-MVE*, RESPECTIVELY, FOR DIFFERENT NOISY CONDITIONS

$Noise\ Type$ $SNR/SegSNR$	*Noisy*	*PSS*	*PSS* *-AMT*	*PKLT*	*GSVD* *-MVE*
White $10.00/6.13$	93.33	76.67	93.33	80.00	100.00
White $0.00/-3.86$	80.00	96.66	76.67	70.00	93.33
Volvo-car $0.00/-2.85$	73.33	90.00	50.00	76.67	83.33
Volvo-car $-10.00/-12.85$	89.28	89.28	82.14	53.57	85.71
Babble $5.00/1.70$	71.43	82.14	78.57	67.85	67.85
Factory $5.00/1.49$	78.57	53.57	78.57	67.85	71.43
Average	80.99	81.39	76.55	69.32	83.61

E. Subjective Listening Tests

Two sets of subjective listening tests were performed and reported here.

1) Listening Preference Comparison: As shown in the first two columns of Table I, the White, Volvo-car, Babble, and Factory noises were artificially added to the test speech at some SNR values (10 to -10 dB) with corresponding input SegSNR values. For each case of noise type and SNR values, five test utterance pairs were generated. For each pair, one was processed by the proposed *PCGSVD-MVE* approach and the other was one of the *Noisy* or *PSS*, *PSS-AMT*, *PKLT*, *GSVD-MVE* processed utterances. The first two rows for the White noise at SNR values of 10 and 0 dB targeted at the moderate and highly noisy conditions, respectively. The next two rows of the Volvo-car noise at SNR values of 0 and -10 dB were for the situation of car traveling in the city at a speed of 40 to 50 km/h, and on the highway at a speed of about 100 km/h, respectively. The other two cases for the Babble and Factory noises at 5-dB SNR were also close to the real-world situation. A total of 26 subjects, 8 females and 18 males between 20 and 45 years of age, participated in the test. Each subject was asked to evaluate four different sets of utterances (totally 20 test utterance pairs) with ordinary headphones and chose the one they preferred without knowing which one was which, and on average each utterance pair was evaluated by 17 subjects. From the evaluation results in Table I, it is clear that the *PCGSVD-MVE* approach proposed in this paper outperformed the other approaches for almost all cases of noise types and SNR values. On average, 69.32% to 83.61% of subjects preferred the *PCGSVD-MVE* processed speech than other approaches under concern.

2) MOS Comparison: Mean opinion score (MOS) rating is the most widely used measure for subjective quality tests, in which the subjects rate the test speech from 5 to 1 scales for "Excellent," "Good," "Fair," "Poor," and "Unsatisfactory," respectively, [28]. The same group of 26 subjects as mentioned above participated in this MOS rating, using exactly the same noisy and enhanced test utterances as those in Section IV-E1, plus two clean utterances for the purpose of referenced MOS rating. Fig. 15 depicted the rating results. The averaged MOS rating for the two clean utterances was 4.39. It is clear that the proposed *PCGSVD-MVE* approach received the highest MOS rating, whether the noise is white or not, stationary, or nonstationary.

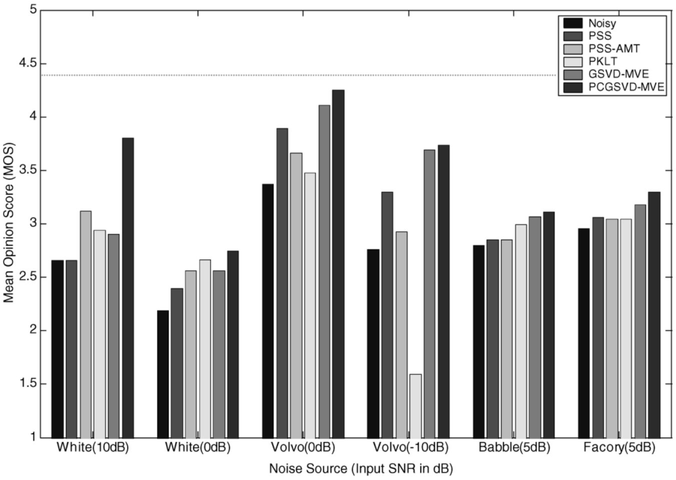


Fig. 15. Mean opinion score rating results for different types of noisy environment.

F. Discussions

For the objective measures in Sections IV-A–IV-C, *GSVD-MVE* roughly offered the best performance in the White noise case. This is apparently because *GSVD-MVE* actually optimized the estimated speech in the signal processing sense. *PCGSVD-MVE* behaved not so well in those objective evaluations for the White noise situation, in that the spectral components of noise-like speech usually have lower energy, but the White noise distribution is flat. For such wide-band noise cases, in order to mask the residual frequency components based on the auditory masking effect, *PCGSVD-MVE* tended to attenuate the high-frequency components (which very often cannot be masked by speech signal under adverse conditions) to make sure the residual noise is nearly unperceivable, which may cause some distortion in the estimated speech. Though *PSS-AMT* also utilized the human auditory effect; however, it only applied the AMTs to figure out the bounded weighting factor and flooring coefficient in the spectral subtraction algorithm in order to achieve a good tradeoff between the perceivable residual noise and the signal distortion [6]; therefore, *PSS-AMT* may introduce less distortion in the high frequency parts than *PCGSVD-MVE* in the White noise case. To reduce the signal distortion of the PCGSVD-processed speech for such broad-band noise situations, we may de-emphasize the Hankel-form matrix $H_{\hat{N}}$ by setting the scaling factor α, as specified in Sections II-B and III-F, less than unity. Moreover, from the subjective listening tests in Section IV-E, the proposed *PCGSVD-MVE* approach indeed offered the best speech quality for this broad-band noise case. In fact, during the tests most of the subjects reported that the *musical noise* induced by

PCGSVD-MVE was less perceivable than that by *GSVD-MVE*, *PSS*, *PSS-AMT*, and *PKLT*.

As for the colored noise cases, it is evident from both the objective and subjective evaluations, *PCGSVD-MVE* actually offered the best performance on average for Volvo-car, Babble, and Factory noises, whether the SNR was high or low. In other words, such real-world noise may be narrow-band and low-passed, so often masked by the voiced speech (but usually not true for the unvoiced speech), and therefore *PCGSVD-MVE* may behave better than *GSVD-MVE* and the other approaches for human perception. An especially worth mentioning case is that for the Volvo-car noise at 0 dB SNR, the MOS rating for *PCGSVD-MVE* in Fig. 15 was even almost as good as that of the clean speech. This is because the Volvo-car noise is low-passed and narrow-band, easily masked by the voiced segments of input speech. Besides, from the evaluation results for the listening preference comparison as in Table I, it is interested to point out that many subjects disliked *GSVD-MVE* processed utterances more than those by *PSS-AMT* and *PKLT*, especially when the additive noise was white. Because the artificially generated residual noises were quite different for various enhancement algorithms, subjects may preferred one kind of residual noise than the others in the listening preference comparison, and hence we obtained the remarkable result as in Table I.

We also evaluated the performance of SegSNR and SegLSD without silence detection error (i.e., using manually labeled silence frames). Experimental results showed that the difference was in general insignificant for stationary types of noise, whether white or not. On the other hand, for nonstationary noise cases, correct detection of noise segments could improve the performance.

Finally, we also estimated the computational complexity of the different enhancement algorithms discussed here. The primary CPU load of *PSS* is the DFT and inverse DFT operations. For *PSS-AMT*, extra computations are for the AMTs, which are relatively limited compared with the DFT operation. *PKLT* requires performing the KLT on the autocorrelation matrix of each speech frame with complexity proportional to $\mathcal{O}(q^3)$, where q is the dimension of the autocorrelation matrix of the speech frame. However, there existed a recursive algorithm to approximate the KLT which can roughly reduce one order computation load of the KLT algorithm [32]. The GSVD algorithm, on the other hand, requires roughly $3LK^2+17K^3$ operations per frame [33], where L and K are, respectively, the row and column sizes of the Hankel-form matrices H_Y and $H_{\hat{N}}$ as in Section II-B, which is roughly more than one order of computational complexity than that of *PSS*. Accordingly, *PCGSVD-MVE* needs to figure out the AMTs in frequency domain, and the AMTs have to be transformed to the generalized singular domain and thus extra operations are needed. Besides, some special techniques for reducing the computations of the GSVD algorithm have been developed as well [34], in which the complexity of one GSVD-update can be reduced to $23.5K^2$ and thus the real-time implementation of the GSVD- and PCGSVD-based approaches on the commercial communication products is achievable.

G. Summary of the Performance Evaluation

All the above results are briefly summarized here. Compared with the noisy speech not enhanced, the proposed GSVD- and PCGSVD-based approaches achieved, respectively, for different types of noise sources on average 6.62 and 6.67 dB improvements in SegSNR evaluations, 1.60 and 1.31 dB reductions in SegLSD measures, and 7.59% and 7.81% absolute phoneme recognition accuracy improvements for the TIMIT test utterances and 4.83% and 13.96% absolute word accuracy improvements for the AURORA2 digit recognition task. In subjective listening tests, on average 69.32% to 83.61% of subjects preferred the *PCGSVD-MVE* processed speech than other approaches being considered. In particular, the good performance for the proposed approach with respect to colored noise is quite clear.

V. Conclusion

In this paper, we first proposed a GSVD-based speech enhancement approach, which was extended from the concept of the truncated QSVD-based approach. Based on this algorithm, a new PCGSVD-based approach by integrating the auditory masking thresholds transformed onto the generalized singular domain has been presented. Closed-form solutions for both the GSVD- and PCGSVD-based enhancement approaches were obtained. Objective measures including time and frequency domain evaluations and speech recognition accuracy measures were used as compared to three other transformation-based speech enhancement algorithms under different noisy environments. The results indicated that the new proposed PCGSVD-based approach can effectively alleviate the perceivable residual noise introduced by the enhancement processes, retain the features of the original speech, and improve the accuracy of the speech recognition system, whether the additive noise is stationary or nonstationary, especially when the noise is nonwhite.

Appendix

The proof of (15) and (16) [referred to here as the solution for the constrained optimization problem (10) of *PCGSVD-MVE*] is given here. It will be shown later on that (6) and (7) [referred to here as the solution for the unconstrained optimization problem (5) of *GSVD-MVE*] turn out to be a special case of (15) and (16). For the constrained optimization problem of (10), an object function $\mathcal{F}(P)$ together with the Kuhn–Tucker conditions can be introduced to convert this constrained optimization problem into an unconstrained one [35], in which two Lagrange multiplier vectors $\mu \in \mathcal{R}^{m\times 1}$ and $\lambda \in \mathcal{R}^{(K-m)\times 1}$ are introduced such that

$$\begin{aligned}\mathcal{F}(P) = &\operatorname{tr}\left[(H_Y P - H_D)^T (H_Y P - H_D)\right]\\ &+ \sum_{i=1}^{m} \mu_i \left\{\operatorname{tr}\left[P^T (H_{\hat{N}})^T \boldsymbol{v}_i \boldsymbol{v}_i^T H_{\hat{N}} P\right] - \gamma_i \|\check{\boldsymbol{x}}_i\|^2\right\}\\ &+ \sum_{j=m+1}^{K} \lambda_j \operatorname{tr}\left[P^T (H_{\hat{N}})^T \boldsymbol{v}_j \boldsymbol{v}_j^T H_{\hat{N}} P\right]\end{aligned} \tag{25}$$

condition on

$$\sum_{i=1}^{m} \mu_i \left\{\operatorname{tr}\left[P^T (H_{\hat{N}})^T \boldsymbol{v}_i \boldsymbol{v}_i^T H_{\hat{N}} P\right] - \gamma_i \|\check{\boldsymbol{x}}_i\|^2\right\} = 0 \tag{26}$$

where $\mathrm{tr}\{\cdot\}$ is the trace of a matrix, $\mu_i \geq 0$, $i = 1, 2 \cdots m$, and λ_j, $j = m+1, m+2 \cdots K$, are the Lagrange multipliers or the components of the vectors μ and λ, respectively. The condition described in (26) is the *complementarity condition*, i.e., the value of μ_i is zero when the corresponding ith constraint is inactive; otherwise, it should be a positive value. Therefore, the constrained *PCGSVD-MVE* optimization problem defined by (10) can be transformed to the unconstrained *GSVD-MVE* problem defined by (5) by assigning all the Lagrange multipliers (μ_i and λ_j) to zero. We take the gradient operation of $\mathcal{F}(P)$ with respect to P to obtain the estimate result P^*

$$\frac{\partial \mathcal{F}(P)}{\partial P} = 0 \tag{27}$$

such that

$$\begin{aligned} 2H_Y^T H_Y P^* - 2H_Y^T H_D + 2\sum_{i=1}^{m} \mu_i (H_{\hat{N}})^T \boldsymbol{v}_i \boldsymbol{v}_i^T H_{\hat{N}} P^* \\ +2\sum_{j=m+1}^{K} \lambda_j (H_{\hat{N}})^T \boldsymbol{v}_j \boldsymbol{v}_j^T H_{\hat{N}} P^* = 0. \end{aligned} \tag{28}$$

We further define a diagonal matrix $\Lambda \in \mathcal{R}^{K\times K}$ whose diagonal elements are exactly the Lagrange multipliers $\mu_i, i = 1, 2 \cdots m$, and λ_j, $j = m+1, m+2 \cdots K$

$$\Lambda \equiv \begin{bmatrix} \mu_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \mu_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & \mu_m & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \lambda_K \end{bmatrix}. \tag{29}$$

Equation (28) can then be written in matrix form using the matrix Λ defined in (29)

$$H_Y^T H_Y P^* - H_Y^T H_D + (H_{\hat{N}})^T V \Lambda V^T H_{\hat{N}} P^* = 0. \tag{30}$$

The matrices H_Y and $H_{\hat{N}}$ in (30) can be further decomposed via the GSVD algorithm [see (2) and (3)] and leading to the following result:

$$X^{-T} C^T C X^{-1} P^* + X^{-T} S^T \Lambda S X^{-1} P^* = X^{-T} C^T U^T H_D. \tag{31}$$

For further developing purposes, with the GSVD algorithm, we can whiten the matrix $H_{\hat{N}}$ by multiplying it by a transformation matrix $(SX^{-1})^{-1} \in \mathcal{R}^{K\times K}$

$$H_{\hat{N},W} = H_{\hat{N}}(SX^{-1})^{-1} = V(SX^{-1})(SX^{-1})^{-1} = V \tag{32}$$

where $H_{\hat{N},W}$ denotes the whitened version of $H_{\hat{N}}$, and the matrices V, S, and X are given in (2) and (3). The autocorrelation matrix for the whitened version of the estimated noise can be approximated by $(H_{\hat{N},W})^T H_{\hat{N},W} = V^T V = I_K$, the identity autocorrelation matrix indicates that it is completely whitened.

From Section II-B, we have $H_Y = H_D + H_N \cong H_D + H_{\hat{N}}$. In order to whiten the noise component in the noisy speech signal $\boldsymbol{y_I}$ as in (32), we can multiply the matrix H_Y with $(SX^{-1})^{-1}$ accordingly

$$\begin{aligned} H_{Y,W} &\cong (H_D + H_{\hat{N}})(SX^{-1})^{-1} \\ &= H_D (SX^{-1})^{-1} + V \\ &= H_{D,W} + H_{\hat{N},W} \end{aligned} \tag{33}$$

where the extra subscript W for the matrices $H_{Y,W}$ and $H_{D,W}$ again denotes the version of H_Y and H_D with noise component whitened. Again with the GSVD algorithm, the matrix $H_{Y,W}$ can be factorized as follows:

$$\begin{aligned} H_{Y,W} &= H_Y (SX^{-1})^{-1} = UCS^{-1} I_K \\ &= [U_1\ U_2] \begin{bmatrix} C_1 S_1^{-1} & 0 \\ 0 & C_2 S_2^{-1} \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & I_{K-m} \end{bmatrix} \end{aligned} \tag{34}$$

where the matrices $U_1 \in \mathcal{R}^{L\times m}$ and $U_2 \in \mathcal{R}^{L\times(K-m)}$, respectively, consist of the first m orthonormal column vectors and the rest $K-m$ orthonormal columns of the matrix U. The diagonal matrices $C_1 \in \mathcal{R}^{m\times m}$ and $C_2 \in \mathcal{R}^{(K-m)\times(K-m)}$, respectively, comprise the preceding m diagonal elements and the rest $K-m$ diagonal elements of the matrix C, and $S_1 \in \mathcal{R}^{m\times m}$ and $S_2 \in \mathcal{R}^{(K-m)\times(K-m)}$ can be similarly obtained from the matrix S. In a similar way, with the SVD algorithm, the matrix $H_{D,W}$ can be decomposed as follows:

$$\begin{aligned} H_{D,W} &= QDZ^T = [Q_1\ Q_2] \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1^T \\ Z_2^T \end{bmatrix} \\ &= Q_1 D_1 Z_1^T \end{aligned} \tag{35}$$

where each column of the matrices $Q \in \mathcal{R}^{L\times K}$ or $Z \in \mathcal{R}^{K\times K}$ is orthonormal, $D \in \mathcal{R}^{K\times K}$ is a diagonal matrix whose rank is not known *a priori*, but can be reasonably approximated as the dimension of the signal subspace of the matrices H_Y and $H_{\hat{N}}$ (i.e., m), the diagonal matrix $D_1 \in \mathcal{R}^{m\times m}$ consists of the m nonzero diagonal elements of the matrix D, $Q_1 \in \mathcal{R}^{L\times m}$ and $Q_2 \in \mathcal{R}^{L\times(K-m)}$ consist of the first m columns and the rest $K-m$ columns of the matrix Q, respectively, and $Z_1 \in \mathcal{R}^{K\times m}$ (whose columns span the signal subspace of $H_{D,W}$) and $Z_2 \in \mathcal{R}^{K\times(K-m)}$ (whose columns span the noise subspace of $H_{D,W}$) are similarly obtained from the matrix Z. Substituting (34) and (35) into (33) with the relation $ZZ^T = Z_1 Z_1^T + Z_2 Z_2^T = I_K$, we have

$$\begin{aligned} H_{Y,W} &= Q_1 D_1 Z_1^T + V\left(Z_1 Z_1^T + Z_2 Z_2^T\right) \\ &= \left[(Q_1 D_1 + V Z_1)\left(D_1^2 + I_m\right)^{-\frac{1}{2}}\ V Z_2\right] \\ &\quad \begin{bmatrix} \left(D_1^2 + I_m\right)^{\frac{1}{2}} & 0 \\ 0 & I_{K-m} \end{bmatrix} \begin{bmatrix} Z_1^T \\ Z_2^T \end{bmatrix} \\ &= [U_1\ U_2] \begin{bmatrix} C_1 S_1^{-1} & 0 \\ 0 & C_2 S_2^{-1} \end{bmatrix} \begin{bmatrix} I_m & 0 \\ 0 & I_{K-m} \end{bmatrix}. \end{aligned} \tag{36}$$

The association of the matrices Z_1, Z_2, U_1, D_1, Q_1, and others in (36) can be obtained as follows:

$$Z_1 = \begin{bmatrix} I_m \\ 0 \end{bmatrix} \tag{37}$$

$$Z_2 = \begin{bmatrix} 0 \\ I_{K-m} \end{bmatrix} \tag{38}$$

$$\begin{aligned} U_1 &= (Q_1 D_1 + V Z_1)\left(D_1^2 + I_m\right)^{-\frac{1}{2}} \\ &= (Q_1 D_1 + V_1)\left(D_1^2 + I_m\right)^{-\frac{1}{2}} \quad (39)\\ C_1 S_1^{-1} &= \left(D_1^2 + I_K\right)^{\frac{1}{2}} \quad (40)\\ C_2 S_2^{-1} &= I_{K-m} \quad (41)\end{aligned}$$

where the matrix $V_1 \in \mathcal{R}^{L\times m}$ consists of the first m orthonormal column vectors of the matrix V. Moreover, from (39) and (40) it is evident that

$$Q_1 D_1 = U_1\left(D_1^2 + I_K\right)^{\frac{1}{2}} - V_1 = U_1 C_1 S_1^{-1} - V_1. \quad (42)$$

With the derivation from (35)–(42), the matrix H_D can be written as follows:

$$\begin{aligned} H_D &= H_{D,W} S X^{-1} = Q_1 D_1 Z_1^T S X^{-1} \\ &= \left(U_1 C_1 S_1^{-1} - V_1\right) Z_1^T S X^{-1} \\ &= U_1 C_1 S_1^{-1}[I_m\ 0] S X^{-1} - V_1 [I_m\ 0] S X^{-1} \\ &= U_1 [C_1\ 0] X^{-1} - V_1 [S_1\ 0] X^{-1}. \quad (43)\end{aligned}$$

Substituting (43) into (31), we have

$$\begin{aligned} X^{-T} C^T U^T H_D &= X^{-T} C^T U^T U_1 [C_1\ 0] X^{-1} \\ &\quad - X^{-T} C^T U^T \overbrace{V_1 [S_1\ 0] X^{-1}}^{\mathcal{A}_1} \\ &= \overbrace{X^{-T}\begin{bmatrix} C_1^2 & 0 \\ 0 & 0\end{bmatrix} X^{-1}}^{\mathcal{B}_1} - H_Y^T \mathcal{A}_1 \\ &= \mathcal{B}_1 - (H_D + H_N)^T \mathcal{A}_1 \\ &\cong \mathcal{B}_1 - (H_D + H_{\hat{N}})^T \mathcal{A}_1 \\ &\cong \mathcal{B}_1 - (H_{\hat{N}})^T \mathcal{A}_1 \\ &= \mathcal{B}_1 - X^{-T} S^T V^T \mathcal{A}_1 \\ &= X^{-T}\begin{bmatrix} C_1^2 & 0 \\ 0 & 0\end{bmatrix} X^{-1} \\ &\quad - X^{-T}\begin{bmatrix} S_1^2 & 0 \\ 0 & 0\end{bmatrix} X^{-1} \\ &= X^{-T}\begin{bmatrix} C_1^2 - S_1^2 & 0 \\ 0 & 0\end{bmatrix} X^{-1}. \quad (44)\end{aligned}$$

Based on the assumption that the clean speech is uncorrelated with the additive noise, the matrices H_D and $\mathcal{A}_1$ (reconstructed from the signal subspace of the matrix $H_{\hat{N}}$) are thus uncorrelated accordingly. Afterwards we substitute the result of (44) into (31) with the assumption that the equal sign of (31) still applies

$$\begin{aligned} X^{-T} C^T C X^{-1} P^* &+ X^{-T} S^T \Lambda S X^{-1} P^* \\ &= X^{-T}\begin{bmatrix} C_1^2 - S_1^2 & 0 \\ 0 & 0\end{bmatrix} X^{-1} \quad (45)\end{aligned}$$

or

$$C_2 X^{-1} P^* X + S^T \Lambda S X^{-1} P^* X = \begin{bmatrix} C_1^2 - S_1^2 & 0 \\ 0 & 0\end{bmatrix} \quad (46)$$

where the matrices Λ, C_1, and S_1 are those defined in (29) and (34), respectively. We then define a matrix $P' \equiv X^{-1} P^* X$, $P' \in \mathcal{R}^{K\times K}$, and from (46) we know that the matrix P' is diagonal with its last $K-m$ diagonal elements being zero, i.e., the rank of P' is m. Equation (46) can be rearranged as follows:

$$P' \equiv \begin{bmatrix} P_1' & 0 \\ 0 & 0\end{bmatrix} \quad (47)$$

and thus

$$C_1^2 P_1' + S_1^T \Lambda_1 S_1 P_1' = C_1^2 - S_1^2 \quad (48)$$

or

$$P_1' = \left(C_1^2 + S_1^T \Lambda_1 S_1\right)^{-1}\left(C_1^2 - S_1^2\right) \quad (49)$$

where the diagonal matrices $\Lambda_1 \in \mathcal{R}^{m\times m}$ and $P_1' \in \mathcal{R}^{m\times m}$ consist of the first m diagonal elements of the matrices Λ and P', respectively. Thus the diagonal elements of the matrix P', p_i', $i = 1, 2\cdots K$, are as follows:

$$p_i' = \left\{\begin{array}{ll} \frac{c_i^2 - s_i^2}{c_i^2 + s_i^2\mu_i} = \frac{1-\frac{s_i^2}{c_i^2}}{1+\frac{s_i^2}{c_i^2}\mu_i}, & 1 \le i \le m \\ 0, & m+1 \le i \le K \end{array}\right\} \quad (50)$$

where c_i, s_i, and μ_i are those defined in (2), (3), and (25), respectively. Because the last $K-m$ diagonal elements of the matrix P' are zero, the equality part of the constraints of (10) are always guaranteed. For the inequality part of the constraints, the energy of the transformed residual noise components as in (10) can be evaluated as follows:

$$\begin{aligned} &\left\|\boldsymbol{v}_i^T H_{\hat{N}} P^*\right\|^2 \le \overbrace{\gamma_i \|\check{\boldsymbol{x}}_i\|^2}^{\mathcal{G}_i}, \quad 1 \le i \le m \quad (51)\\ &\text{implies} : \boldsymbol{v}_i^T H_{\hat{N}} P^* (P^*)^T (H_{\hat{N}})^T \boldsymbol{v}_i \le \mathcal{G}_i \quad (52)\\ &\boldsymbol{v}_i^T V S X^{-1} X P' X^{-1} X^{-T} (P')^T X^T X^{-T} S^T V^T \boldsymbol{v}_i \le \mathcal{G}_i \\ &\quad (53)\\ &I_{K,i}^T S P' X^{-1} X^{-T} (P')^T S^T I_{K,i} \le \mathcal{G}_i \quad (54)\\ &\begin{bmatrix} 0 & 0 & \cdots & s_i \cdot \frac{1-\frac{s_i^2}{c_i^2}}{1+\frac{s_i^2}{c_i^2}\mu_i} & \cdots & 0 & 0 \end{bmatrix} X^{-1} X^{-T} \\ &\begin{bmatrix} 0 & 0 & \cdots & s_i \cdot \frac{1-\frac{s_i^2}{c_i^2}}{1+\frac{s_i^2}{c_i^2}\mu_i} & \cdots & 0 & 0 \end{bmatrix}^T \le \mathcal{G}_i \quad (55)\end{aligned}$$

which leads to

$$\left(s_i \cdot \frac{1-\frac{s_i^2}{c_i^2}}{1+\frac{s_i^2}{c_i^2}\mu_i}\right)^2 \cdot \|\check{\boldsymbol{x}}_i\|^2 \le \gamma_i \|\check{\boldsymbol{x}}_i\|^2 \quad (56)$$

or

$$\frac{1-\frac{s_i^2}{c_i^2}}{1+\frac{s_i^2}{c_i^2}\mu_i} \le \frac{\sqrt{\gamma_i}}{s_i}, \quad 1 \le i \le m \tag{57}$$

where the vectors $I_{K,i}$ and $\check{\boldsymbol{x}}_i$, $i = 1, 2 \cdots m$, are, respectively, the ith column and row of a K-dimensional identity matrix and the matrix X^{-1}, and $\|\cdot\|^2$ denotes the l_2 norm operation. From (51), with the Kuhn–Tucker conditions, the Lagrange multipliers μ_i, $i = 1, 2 \cdots m$, can be classified according to the following two cases:

$$\begin{aligned} &\text{(i) Constraint-inactivated case} : \mu_i = 0 \\ &\left\|\boldsymbol{v}_i^T H_{\hat{N}} P^*\right\|^2 < \gamma_i \|\check{\boldsymbol{x}}_i\|^2, \quad 1 \le i \le m, \end{aligned} \tag{58}$$

and

$$\begin{aligned} &\text{(ii) Constraint-activated case} : \mu_i > 0 \\ &\left\|\boldsymbol{v}_i^T H_{\hat{N}} P^*\right\|^2 = \gamma_i \|\check{\boldsymbol{x}}_i\|^2, \quad 1 \le i \le m. \end{aligned} \tag{59}$$

For the case of constraint-inactivated, we set μ_i, $i = 1, 2 \cdots m$, to zero and having the result [referred to (50) and (57)]

$$p_i' = 1 - \frac{s_i^2}{c_i^2} \tag{60}$$

where

$$\frac{\sqrt{\gamma_i}}{s_i} > \left(1 - \frac{s_i^2}{c_i^2}\right). \tag{61}$$

For the constraint-activated case, on the other hand, the equal sign of (57) holds, and therefore the ith diagonal element of the matrix P', p_i', $i = 1, 2 \cdots m$, in (50) has the form

$$p_i' = \frac{1-\frac{s_i^2}{c_i^2}}{1+\frac{s_i^2}{c_i^2}\mu_i} = \frac{\sqrt{\gamma_i}}{s_i}. \tag{62}$$

With (50)–(62), the diagonal elements p_i', $i = 1, 2 \cdots K$, of the transformation matrix P' are

$$p_i' = \left\{ \begin{matrix} \min\left[1 - \frac{s_i^2}{c_i^2}, \frac{\sqrt{\gamma_i}}{s_i}\right], & 1 \le i \le m \\ 0, & m+1 \le i \le K \end{matrix} \right\}. \tag{63}$$

Hence, the estimated matrix $H_{\hat{D}}$ for the clean speech frame can be expressed as follows:

$$\begin{aligned} H_{\hat{D}} &= H_Y P^* = UCX^{-1}XP'X^{-1} \\ &= UCP'X^{-1} = UC'X^{-1} \end{aligned} \tag{64}$$

where $C' \equiv CP'$ is a $K \times K$ diagonal matrix with its diagonal elements being those in (16). For the unconstrained optimization problem for *GSVD-MVE* as formulated in (5), it is easy to verify its solution as given in (6) and (7) by assigning all the Lagrange multipliers μ_i, $i = 1, 2 \cdots m$, of (50) to zero. This concludes the proof for (15) and (16) for the solution of the constrained optimization problem of *PCGSVD-MVE*, and (6) and (7) for the solution of the unconstrained optimization problem of *GSVD-MVE*.

ACKNOWLEDGMENT

The authors would like to thank Dr. Y.-S. Lee, Dr. B.-S. Jeng, Dr. S.-W. Swan, and Dr. R.-K. Chen of Chunghwa Telecommunication Laboratories, Taiwan, R.O.C., for their consecutive support and encouragements. They would also like to thank the anonymous reviewers and associated editor for their valuable comments.

REFERENCES

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

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

[3] B.-L. Sim, Y.-C. Tong, J.-S. Chang, and C.-T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[4] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[5] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 497–514, Nov. 1997.

[6] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[7] C.-H. You, S.-N. Koh, and S. Rahardja, "Subspace speech enhancement for audible noise reduction," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2005, pp. 145–148.

[8] E. Zwicker and H. Fastle, *Psychoacoustics*, 2nd ed. New York: Springer-Verlag, 1999.

[9] S. V. Huffel, "Enhanced resolution based on minimum variance estimation and exponential data modeling," *Signal Process.*, vol. 33, pp. 333–355, Sep. 1993.

[10] B. T. Lilly and K. K. Paliwal, "Robust speech recognition using singular value decomposition based speech enhancement," in *Proc. IEEE TENCO-Speech and Image Tech. Comput.Telecommun.*, 1997, pp. 257–260.

[11] M. Klein and P. Kabal, "Signal subspace speech enhancement with perceptual post-filtering," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2002, pp. 537–540.

[12] 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 Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[13] G.-H. Ju and L.-S. Lee, "Speech enhancement based on generalized singular value decomposition approach," in *Proc. ICSLP*, 2002, pp. 1801–1804.

[14] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[15] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[16] Y. Hu and P. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Process. Lett.*, vol. 9, no. 7, pp. 204–206, Jul. 2002.

[17] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.

[18] G.-H. Ju and L.-S. Lee, "Perceptually constrained generalized singular value decomposition-based approach for enhancing speech corrupted by colored noise," in *Proc. Eurospeech*, 2003, pp. 533–536.

[19] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2002, pp. 569–572.

[20] C. C. Paige and M. A. Saunders, "Towards a generalized singular value decomposition," *SIAM J. Numer. Anal.*, vol. 18, pp. 398–405, 1981.

[21] M. Dendrinos, S. Bakamidis, and G. Garayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.*, vol. 10, pp. 45–57, Feb. 1991.

[22] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*. New York: Wiley, 1999.

[23] J.-L. Shen, J.-W. Hung, and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. ICSLP*, 1998, pp. 232–235.

[24] J. S. Garofolo, *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*. Gaithersburg, MD: National Inst. Stand. Technol. (NIST), 1988.

[25] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.

[26] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Extended Advanced Front-End Feature Extraction Algorithm; Compression Algorithms; Back-End Speech Reconstruction Algorithm*, ETSI Std. ES 202 212 V1.1.1 Recommendation, Nov. 2003.

[27] J. Ramirez, J. C. Segura, C. Benitez, Á. Torre, and A. Rubio, "A new adaptive long-term spectral estimation voice activity detector," in *Proc. Eurospeech*, 2003, pp. 3041–3044.

[28] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[29] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[30] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[31] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR2000*, 2000, pp. 181–188.

[32] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 87–95, Feb. 2001.

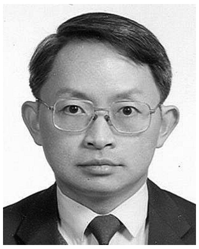
[33] F. T. Luk, "A parallel method for computing the generalized singular value decomposition," *J. Paral. Distrib. Comput.*, vol. 2, pp. 250–260, Aug. 1985.

[34] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[35] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 2000.

Gwo-Hwa Ju received the M.S. degree in electrical engineering and the Ph.D. degree in communication engineering, from National Taiwan University, Taipei, Taiwan, R.O.C., in 1990 and 2006, respectively.

He has been a Researcher with the Multimedia Applications Technology Laboratory, Telecommunication Laboratories, Chunghwa Telecom Corporation, Ltd., Taoyuan, Taiwan, since 1990. In 1992, he was a Visiting Researcher with the Speech Recognition and Synthesis Group, Human Interface Laboratories, NTT Yokosuka, Japan. His research interests include robust speech recognition, speech synthesis, speech coding, digital signal processing, and embedded system design.



Lin-Shan Lee (F'94) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of electrical engineering and computer science at the National Taiwan University, Taipei, Taiwan, R.O.C., since 1982 and holds a joint appointment as a Research Fellow of Academia Sinica, Taipei. His research interests include digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world including text-to-speech system, natural language analyzer, dictation systems, and voice information retrieval system.

Dr. Lee was Guest Editor of a Special Issue on Intelligent Signal Processing in Communications of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS in December 1994 and January 1995. He was Vice President for International Affairs (1996–1997) and the Awards Committee chair (1998–1999) of the IEEE Communications Society. He has been a member of Permanent Council of International Conference on Spoken Language Processing (ICSLP), was the convener of COCOSDA (International Coordinating Committee of Speech Databases and Assessment, 2000–2001), and is currently a member of the Board of International Speech Communication Association (ISCA).