

Perceptually-Constrained Generalized Singular Value Decomposition-Based Approach for Enhancing Speech Corrupted by Colored Noise

Gwo-hwa Ju^{1,2}, Lin-shan Lee¹

¹Graduate Institute of Communication Engineering, National Taiwan University, Taipei,

²Chunghwa Telecommunication Laboratories, Taoyuan,

Taiwan, Republic of China

jgh@cht.com.tw

Abstract

In a previous work, we have successfully integrated the transformation-based signal subspace technique with the generalized singular value decomposition (GSVD) algorithm to develop an improved speech enhancement framework [1]. In this paper, we further incorporate the perceptual masking effect of the psychoacoustics model as extra constraints of the previously proposed GSVD-based algorithm to obtain improved sound feature, and furthermore make sure the undesired residual noise to be nearly unperceivable. Both subjective listening tests and spectrogram-plot comparison showed that the closed-form solution developed here can offer significantly better speech quality than either the conventional spectral subtraction algorithm or the previously proposed GSVD-based technique, regardless of whether the additive noise is white or not.

1. Introduction

Speech enhancement techniques have long been employed to improve the quality and intelligibility of distorted or corrupted speech and to improve the speech recognition accuracy. For speech corrupted by additive noise, the spectral subtraction algorithm in power spectral domain (PSS) has been very popular [2]. A weakness of PSS algorithm is that it may produce some unnatural residual noise, the so-called *musical noise*, especially when the signal-to-noise ratio (SNR) is low (e.g., less than 10dB), usually due to the inevitable random tone peaks generated in the spectrum of the enhanced speech signals. Previous studies indicated that this residual noise can be effectively alleviated by incorporating the frequency-domain masking-based psychoacoustics model [3,4,5], i.e., noise won't be perceived by the human auditory system if it is below some auditory masking thresholds.

An improved signal subspace approach for speech enhancement using a generalized singular value decomposition (GSVD) algorithm was previously developed [1]. Experimental results showed that this approach provided very good speech quality, but some *musical noise* is still perceivable under lower SNR conditions. In this paper, we further incorporate the human auditory masking mechanism with the previously proposed GSVD-based signal subspace techniques mentioned above to establish an improved framework for speech enhancement, referred to as the perceptually-constrained generalized singular value decomposition (PCGSVD)-based approach in this paper. Because the GSVD-based technique operates in the domain of generalized singular vectors, while the popular psychoacoustics model for human auditory masking effect is well defined in the frequency domain, a transformation from the frequency domain to the domain of eigenvectors recently proposed [5] were extended to

a transformation from the frequency domain to the domain of generalized singular vectors, and therefore the desired incorporation of the masking-based psychoacoustics model and the GSVD-based speech enhancement approach is achievable. Experimental results based on subjective listening tests and spectrogram-plot comparison showed this proposed algorithm can effectively reduce the residual noise and produce improved speech quality, regardless of whether the noise is white or not; specially when the SNR is low.

The rest of the paper is organized as follows. In Section 2, the evaluation of the auditory masking thresholds and the GSVD algorithm are very briefly summarized for development purposes. The proposed PCGSVD-based approach for speech enhancement is then described in Section 3, with the details of the core algorithm presented in Section 4. Experimental results and discussions are given in Section 5, and some conclusions finally offered in Section 6.

2. Brief summary of the auditory masking threshold evaluation and the GSVD algorithm

In this section, we very briefly summarize the two fundamental frameworks needed in our proposed approach.

2.1. Evaluation of the auditory masking thresholds [3,4]

The procedures for evaluating the masking thresholds for human perception are well known, as briefly summarized here.

In the human auditory system, the frequency range roughly from 20Hz to 20000Hz can be modeled by 25 critical bands. The evaluation of the masking thresholds is therefore as follows. We first add up the magnitude square of the corresponding FFT components in each critical band, we then convolve the obtained critical band energy sequence with a spreading function in order to consider the cross correlation between 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 frames. The auditory masking thresholds are finally obtained by renormalizing the above bark-domain sequence to compensate for the gain modification of the convolutional process, and make sure they are not below the absolute masking thresholds of human hearing.

2.2. The GSVD algorithm [1,6]

The GSVD algorithm can simultaneously transform two matrices with the same column dimension into two nonnegative diagonal matrices. For two real $L \times K$ matrices H_Y , $H_N \in \mathbf{R}^{L \times K}$, we can find a nonsingular matrix $X \in \mathbf{R}^{K \times K}$ and two orthogonal matrices $U, V \in \mathbf{R}^{L \times L}$, such that

$$\begin{aligned} \mathbf{U}^T \mathbf{H}_Y (\mathbf{X}^T)^{-1} &= \mathbf{C} = \text{diag}[c_1, c_2, \dots, c_K], c_1 \geq c_2 \geq \dots \geq c_K, \\ \mathbf{V}^T \mathbf{H}_N (\mathbf{X}^T)^{-1} &= \mathbf{S} = \text{diag}[s_1, s_2, \dots, s_K], s_K \geq s_{K-1} \geq \dots \geq s_1, \end{aligned} \quad (1)$$

where the diagonal elements of the matrix \mathbf{C} are arranged in descending order, while those of \mathbf{S} are in ascending order. There is a constraint for the matrices \mathbf{C} and \mathbf{S} , $\mathbf{C}^T \mathbf{C} + \mathbf{S}^T \mathbf{S} = \mathbf{I}_K$, where \mathbf{I}_K is the K-dimensional identity matrix. The values $c_l/s_l, \dots, c_K/s_K$ and the columns of the matrix \mathbf{X} are respectively referred to as the generalized singular values and the generalized singular vectors of the matrices \mathbf{H}_Y and \mathbf{H}_N .

3. The proposed PCGSVD-based speech enhancement approach

For additive noise, the noisy speech samples y_n , $n=0,1,2,\dots$, and the desired clean speech samples d_n , $n=0,1,2,\dots$ are related by

$$y_n = d_n + n_n, \quad (2)$$

where n_n is the noise samples. The goal of speech enhancement is to estimate d_n from y_n . Fig. 1 depicts the block diagram of the proposed PCGSVD-based speech enhancement framework, which includes five separated units as briefly described below.

3.1. Unit (I): Non-speech detector and buffer

A simple silence detection algorithm and a noise buffer are employed here to identify and accumulate the non-speech parts of the input signals, based on which the noise statistics can be estimated.

3.2. Unit (II): Framing and matrix construction

The noisy speech samples y_n are segmented into overlapped frames each with M samples, while the estimated noise frames are constructed from the latest buffered noise samples $\hat{n}_i, i=0,\dots,M-1$, which are obtained in the detected non-speech parts from Unit (I). Two series of Hankel-form matrices of order $L \times K$, $\mathbf{H}_Y, \mathbf{H}_{\hat{N}} \in \mathbf{R}^{L \times K}$, as illustrated in Fig. 2, are then constructed on for each of these frames, \mathbf{H}_Y for the noisy speech frames for y_n and $\mathbf{H}_{\hat{N}}$ for the buffered noise \hat{n}_i respectively, where $L+K=M+1$ and in general K is much smaller than L. Under noise free situation, the value of K is chosen such that the matrix \mathbf{H}_Y is rank deficient, i.e. the rank of \mathbf{H}_Y is smaller than K. This rank deficiency condition makes it easier to choose a boundary to divide the signal subspace and the noise subspace later on when the noise is presented.

From equation (2), it is clear that the matrix \mathbf{H}_Y can be represented as the summation of two Hankel-form matrices \mathbf{H}_D and \mathbf{H}_N , $\mathbf{H}_Y = \mathbf{H}_D + \mathbf{H}_N$, respectively constructed from the clean speech frames and the real noise frames. Both \mathbf{H}_D and \mathbf{H}_N are unknown, but \mathbf{H}_N can be approximated by $\mathbf{H}_{\hat{N}}^{EST} = \alpha \cdot \mathbf{H}_{\hat{N}}$, where the scaling factor α describes the degree of distortion for the enhanced speech and $\mathbf{H}_{\hat{N}}$ is constructed above with the buffered noise. The goal here is to estimate the matrix \mathbf{H}_D from the given matrices \mathbf{H}_Y and $\mathbf{H}_{\hat{N}}^{EST}$.

3.3. Unit (III): GSVD algorithm

The GSVD algorithm summarized in Section 2.2 is used here to obtain the matrices $\mathbf{U}, \mathbf{V}, \mathbf{X}$ and to simultaneously transform

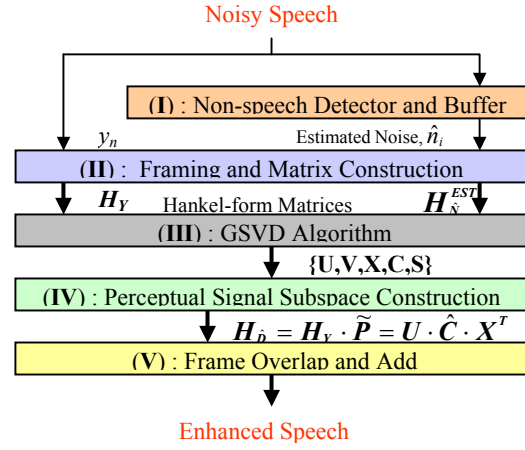


Figure 1: Block diagram of the proposed PCGSVD-based speech enhancement framework

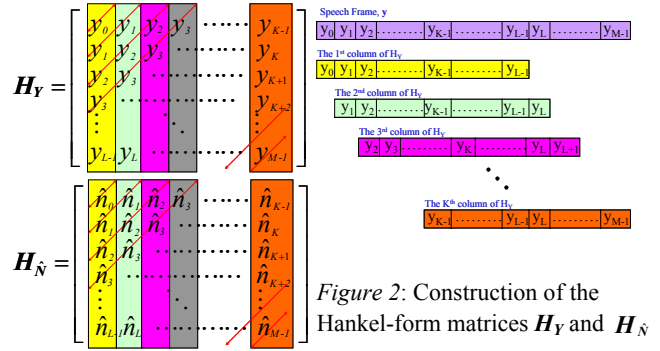


Figure 2: Construction of the Hankel-form matrices \mathbf{H}_Y and $\mathbf{H}_{\hat{N}}$

the two Hankel-form matrices \mathbf{H}_Y and $\mathbf{H}_{\hat{N}}^{EST}$ into nonnegative diagonal matrices \mathbf{C} and \mathbf{S} .

3.4. Unit (IV): Perceptual signal subspace construction

This is the core of the proposed approach, to integrate the auditory masking effect described in Section 2.1 with the previously proposed GSVD-based algorithm using the minimum variance (MV) estimation principle [1,6]. A transformation matrix $\tilde{\mathbf{P}} \in \mathbf{R}^{K \times K}$ is obtained here such that $\mathbf{H}_{\hat{D}} = \mathbf{H}_Y \cdot \tilde{\mathbf{P}}$ is the best estimate of the desired Hankel-form matrix \mathbf{H}_D for the clean speech under some constraints defined by the auditory masking thresholds and the MV estimator. The details of this unit will be presented in the next section.

3.5. Unit (V): Frame overlap and add

After the enhanced speech frames were obtained from the estimated matrix $\mathbf{H}_{\hat{D}}$, we concatenate them frame-by-frame with the overlap-add method to give the finally enhanced speech signals.

4. Construction of the perceptual signal subspace

Here the detailed process of the Unit (IV) mention in section 3.4 will be presented. First, the diagonal elements of the matrix \mathbf{C} in equation (1) can be properly split into two non-overlapped and contiguous parts, referred to as the signal subspace and the noise subspace for speech enhancement purposes here. Then the concept of incorporating the MV estimation and the

auditory masking effect here is to find a transformation matrix $\tilde{\mathbf{P}}$ which can transform the given matrix \mathbf{H}_Y (for noisy speech) to the best estimate of the unknown desired matrix \mathbf{H}_D (for clean speech), or minimize the distance between the two matrices $\mathbf{H}_Y \cdot \tilde{\mathbf{P}}$ and \mathbf{H}_D , under the constraints that the energies of the residual noise component projected onto the generalized singular vectors are below the transformed auditory masking thresholds in the signal subspace, and are zero in the noise subspace. This is formulated in the mathematical expressions below.

$$\begin{aligned} \tilde{\mathbf{P}} &= \arg \min_{\mathbf{P} \in \mathbf{R}^{K \times K}} \|\mathbf{H}_Y \cdot \mathbf{P} - \mathbf{H}_D\|_F^2, \\ \text{subject to } &\left\{ \begin{array}{l} \left| \mathbf{v}_i^T \cdot \mathbf{H}_{\hat{N}}^{EST} \cdot \mathbf{P} \right|^2 < \beta_i \cdot \gamma_i \cdot \|\mathbf{X}_i\|^2, 1 \leq i \leq N \\ \left| \mathbf{v}_i^T \cdot \mathbf{H}_{\hat{N}}^{EST} \cdot \mathbf{P} \right|^2 = 0, N+1 \leq i \leq K \end{array} \right\}, \quad (3) \end{aligned}$$

where $\|\mathbf{A}\|_F$ is the Frobenius norm of a matrix \mathbf{A} , \mathbf{v}_i is the i^{th} column vector of the matrix \mathbf{V} obtained in equation (1). β_i , $i=1, \dots, N$, is a constant and for simplification here, we set all of the β_i to unity. N is the dimension of the signal subspace for the matrix \mathbf{H}_Y and $\mathbf{H}_{\hat{N}}^{EST}$, obtainable from equation (1) such that $c_N > s_N$ and $s_{N+1} \geq c_{N+1}$, $1 \leq N \leq K$, $K-N$ is the dimension of the noise subspace, and $\|\mathbf{X}_i\|$ is the l_2 norm of the i^{th} column of the matrix \mathbf{X} given in equation (1). γ_i is the i^{th} auditory masking threshold but transformed to the domain of generalized singular vectors which can be evaluated by the procedures summarized below [1,5].

Because the clean speech samples d_n are unknown, the periodogram average (Bartlett's method) of the clean speech can be approximated as follows. (assume the clean speech signals are short-time wide sense stationary random process)

$$\hat{\Phi}_{BT} \cong \frac{1}{MK} \sum_{i=1}^N \mathbf{w}_i \cdot c_i^2 \cdot \left(1 - \frac{s_i^2}{c_i^2}\right)^2, \quad (4)$$

where c_i and s_i are the i^{th} diagonal element of the matrix \mathbf{C} and \mathbf{S} respectively in equation (1), \mathbf{w}_i is the vector for the magnitude square of the J-point DFT (J is the number of masking thresholds obtained by the procedure summarized in Section 2.1) of the i^{th} column vector of the matrix \mathbf{X} obtained from equation (1). Note that the summation in equation (4) is over $i=[1, 2, \dots, N]$, or the dimensions of the signal subspace of the input speech frames. The frequency domain masking thresholds, Θ_γ , can then be evaluated from $\hat{\Phi}_{BT}$ in equation (4) based on the procedures described in Section 2.1. With Θ_γ , the masking thresholds γ_i , $i=1, 2, \dots, N$, in each of the dimensions of the generalized singular vectors are:

$$\gamma_i = \left| \gamma_i' \right|, i = 1, 2, \dots, N, \quad (5)$$

where γ_i' is the i^{th} element of the vector $\mathbf{G}' \in \mathbf{R}^{J \times 1}$ given below.

$$\mathbf{G}' = \frac{1}{J} \mathbf{G}^T \cdot \Theta_\gamma, \quad (6)$$

where the i^{th} column vector \mathbf{g}_i , $1 \leq i \leq K$, of the transformation matrix $\mathbf{G} \in \mathbf{R}^{J \times K}$ is the magnitude squared J-point DFT of the i^{th} column vector of the inverse matrix \mathbf{X}^{-1} . With the formulation

Table 1. Percentage of listening preference comparison of **PCGSVD-processed speech** with noisy, PSS-processed, and GSVD-processed speech for different kinds of noise source, respectively

| Noise Type | SNR(dB)/ SNRSeg | Compared with Noisy's | Compared with PSS's | Compared with GSVD's |
|-------------|--------------------|--------------------------|------------------------|-------------------------|
| White | 10.0 /2.68 | 87.8% | 76.0% | 68.0% |
| Factory | 10.0 /3.08 | 71.0% | 66.6% | 72.7% |
| F16 cockpit | 10.0 /2.84 | 57.1% | 73.6% | 66.6% |

as summarized above, it can be shown that with the GSVD algorithm mentioned in Section 2.2, the estimation for the Hankel-form matrix \mathbf{H}_D for the desired clean speech d_n has a close-form solution as given in equations (7) and (8):

$$\mathbf{H}_{\hat{D}} = \mathbf{H}_Y \cdot \tilde{\mathbf{P}} = \mathbf{U} \cdot \hat{\mathbf{C}} \cdot \mathbf{X}^T, \quad (7)$$

where the i^{th} diagonal element \hat{c}_i , $1 \leq i \leq K$, of the nonnegative diagonal matrix $\hat{\mathbf{C}}$ is

$$\hat{c}_i = \begin{cases} \min \left[c_i \cdot \left(1 - \frac{s_i^2}{c_i^2}\right), c_i \cdot \frac{\sqrt{\gamma_i}}{s_i} \right], & 1 \leq i \leq N \\ 0, & N+1 \leq i \leq K \end{cases}, \quad (8)$$

where the matrices \mathbf{U} and \mathbf{X} , the diagonal elements c_i and s_i of the matrices \mathbf{C} and \mathbf{S} are those in equation (1), and the two parts of equation (8) are for the signal subspace and the noise subspace respectively. The estimated matrix $\mathbf{H}_{\hat{D}}$ obtained here may not have the Hankel-structure. We can then simply average the anti-diagonal elements of $\mathbf{H}_{\hat{D}}$ to recover the Hankel-structure and thus the enhanced speech, as described above with the Unit (V).

5. Experimental results

The experimental environment was as follows. The sampling rate of the input speech signals was 16kHz. We randomly selected 30 utterances from TIMIT speech corpus for testing. Three types of noise source, 'White', 'Factory', and 'F16 cockpit', chosen from NOISEX-92 database, were artificially added to the test speech. The Factory noise is non-stationary while F16 cockpit noise is stationary; both of them are non-white. The window function for the framer was rectangular, which was found to produce the best performance than other window shapes. The frame size, M, was 512 samples (32 ms) with a 50% frame overlap. The value of J, the number of masking thresholds as mention in Section 4, was 512. The typical value of scaling factor α , specified in section 3.2, ranged from 0.5 to 2.0. Too small value of α could not provide satisfactory results while too large value would severely distort the quality of the enhanced speech as well. Here we simply set α to be unity. The dimensions of the two Hankel-form matrices, L and K, were 473 and 40 respectively. Under the noise free condition with tests for various speech utterances, the signal subspace dimension, N, for the matrix \mathbf{H}_Y as defined in the previous section, practically ranged between 8 and 25. Therefore, when the noise was present, the column size of 40 for the Hankel-form matrices was adequate to divide the signal and noise subspace of the input noisy speech frames.

Table 1 reveals the listening preference comparison for the utterances processed by the PCGSVD-based approach

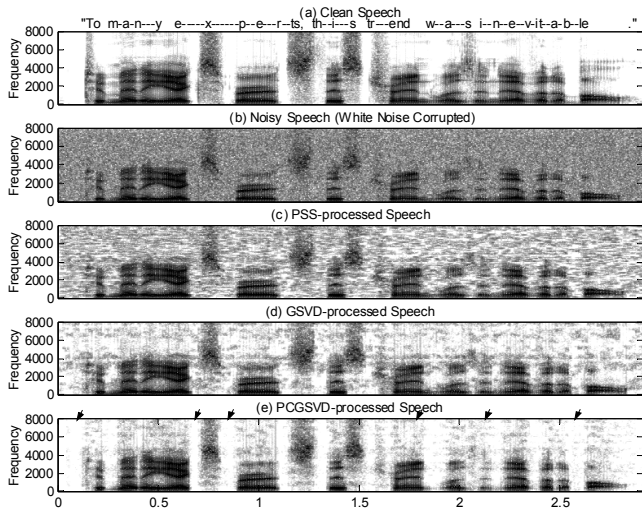


Figure 3: Spectrogram plots for a typical utterance corrupted by white noise

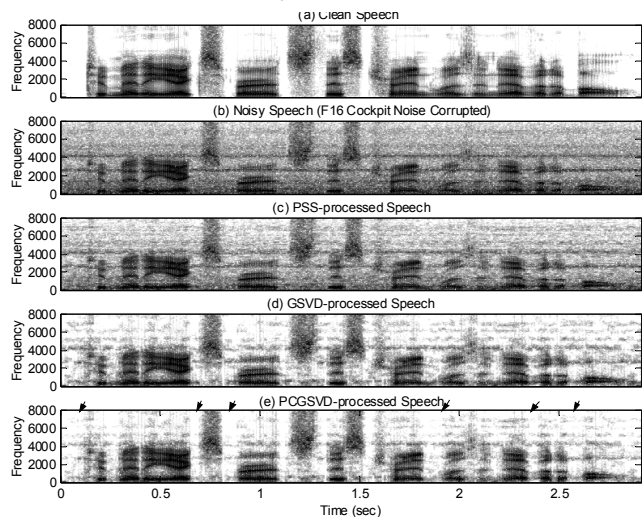


Figure 4: Spectrogram plots for a typical utterance corrupted by F16 cockpit noise

proposed here, with respect to the original noisy speech and those proposed by PSS [2] and the GSVD-based algorithm proposed previously [1] for utterances respectively corrupted by White, Factory, and F16 cockpit noise sources. The SNR of the original noisy speech, as shown in the first column of Table 1, was 10dB, although different noise sources gave slightly different segmental SNR (SNRSeg) values. A total of 30 subjects joined in the listening tests. Each subject evaluated 10 sets of utterances and each set includes 2 utterances in random order, one of them was the PCGSVD-processed speech and the other randomly selected from the original noisy, PSS-processed, and GSVD-processed speech. They were asked to choose a better version without knowing which one was which. From Table 1, it is clear that the PCGSVD-based approach proposed in this paper outperforms the PSS and GSVD-based algorithm for various noise sources, whether it is white or not, stationary or non-stationary.

The spectrogram plots for a test utterance 『To many experts, this trend was inevitable.』 by a male speaker for (a) clean speech, (b) noisy speech, (c) speech enhanced by the PSS algorithm [2], (d) speech enhanced by the previously proposed

GSVD-based algorithm [1], and (e) speech enhanced by the PCGSVD-based algorithm proposed here in this paper are respectively illustrated in Fig. 3 for additive White noise at 10dB of SNR. In Fig. 3(c), we can see that in the PSS-processed speech many undesired random tone peaks were present in the non-speech region and higher frequency parts of voiced regions, which are perceived as the *musical noise*. This situation was significantly improved by the GSVD-base approach in Fig. 3(d), although it was found in the listening tests that the residual noise was still quite perceivable. With the PCGSVD-based algorithm as proposed here, however, it can be observed in Fig. 3(e) that almost the same detailed information of the speech spectra as those in Fig. 3(d) were retained, but much less random tone peaks were present in both the silence and higher frequency components of voiced portion. This was also verified in the listing tests, in which most of the testers agreed that the *musical noise* is less perceivable in the PCGSVD enhanced speech than in the PSS and GSVD-based approach. Very similar trends were observed in Fig. 4, where exactly the same spectrogram plots as those in Fig. 3 were shown, except the additive noise was the F16 cockpit noise with 5dB of SNR. These results further support the fact that the proposed PCGSVD-based speech enhancement algorithm can effectively alleviate the residual noise introduced by the enhancing process and retain the signal quality, whether the additive noise is white or not.

6. Conclusions

In this paper, we proposed a speech enhancement algorithm by integrating the frequency-domain masking-based psychoacoustics model and the previously proposed GSVD-based signal subspace technique. With these, a close-form solution was obtained to reconstruct the enhanced speech from the signal subspace of the input noisy speech. Both the subjective listening tests and the spectrogram-plot comparison showed that the proposed algorithm could offer significant improvements in speech quality as compared to the well known spectral subtraction algorithm or the previously proposed GSVD algorithm, in particular when SNR is low, whether the additive noise is white or not.

7. References

- [1] Gwo-hwa Ju and Lin-shan Lee, "Speech Enhancement based on Generalized Singular Value Decomposition Approach", in *proc. ICSLP*, pp. 1801–1804, Sep. 2002.
- [2] Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. on ASSP*, Vol. ASSP-27, No. 2, pp. 113-120, April 1979.
- [3] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria", *IEEE J. Select. Areas Comm.*, Vol. 6, pp. 314–323, Feb. 1988.
- [4] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", *IEEE Trans. on SAP*, Vol. 7, pp. 126–137, March 1999.
- [5] Firas Jabloun and B. Champagne, "A Perceptual Signal Subspace Approach for Speech Enhancement in Colored Noise", in *proc. ICASSP*, pp. 569-572, April 2002.
- [6] Jensen S., Hansen P., Hansen S., and Sorensen J., "Reduction of Broad-Band Noise in Speech by Truncated QSVD", *IEEE Trans. on SAP*, Vol. 3, pp. 439-448, Nov. 1995.