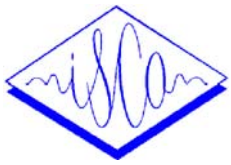

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Improved Speech Enhancement by Applying Time-shift Property of DFT on Hankel Matrices for Signal Subspace Decomposition

Gwo-hwa Ju[1,2] , *Lin-shan Lee*[1]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taipei,
[2]Chunghwa Telecommunication Laboratories, Taoyuan,
Taiwan, Republic of China
jgh@cht.com.tw

Abstract

In previous studies, the signal subspace technique for speech enhancement was extended and a perceptually constrained generalized singular value decomposition (PCGSVD)-based algorithm [1] was developed which properly integrated the auditory masking effect and the GSVD algorithm. Both objective measures and subjective tests verified that this approach can offer better performance than the GSVD-based approach and the conventional spectral subtraction (SS) algorithm. But very high computational complexity is required in the PCGSVD-based method when performing the matrices decomposition via the GSVD algoruthm. In this paper, we properly utilize the time-shift property of DFT and the special structure of Hankel matrices to perform similar functions previously offered by GSVD, and a perceptually constrained minimum variance estimation algorithm is developed. By replacing GSVD algorithm with DFT, the computation complexity is significantly reduced, almost the same as the conventional SS algorithm. Experiments showed that comparable performance to that of the PCGSVD-based approach can be achieved, regardless of whether the additive noise is stationary or not, specially when it is non-white.

1. Introduction

In wired or wireless communication systems, voice quality and intelligibility are important for either human-to-human communications or human-to-machine interactions. For this reason, noise reduction techniques have been employed to improve the quality and intelligibility of the noise-corrupted speech and to improve the performance of the speech recognition systems. The object of speech enhancement aims at diminishing the additive noise from the speech signal, whether it is white or not, stationary or non-stationary. For the case of additive noise, the spectral subtraction (SS) algorithm has been very popular [2,3] for speech enhancement purposes. A weakness of the SS algorithm is that it may produce some unnatural residual noise, the so-called *musical noise*, especially when the signal-to-noise ratio (SNR) of the input speech is low (e.g., less than 10dB), usually due to the inevitable random tone peaks generated in the spectrum of the enhanced speech signal. Other well-known signal-subspace algorithms for speech enhancement, such as the generalized singular value decomposition (GSVD)-based approach or KLT-based method, have received high interests in recent years [1]. In the GSVD-based approach the vector space of each input speech frame was decomposed into non-overlapped signal and noise subspace, and the estimated clean speech is then reconstruct from the signal subspace only. Though these enhancement approaches can effectively alleviate the problem of *musical noise*, some artificial noise is still perceivable under lower SNR conditions.

Our previous studies pointed out this residual noise can be effectively alleviated by considering the masking functions of human auditory system [1], i.e., the residual noise won't be perceivable if it is below the auditory masking thresholds of human ear. Experimental results showed that the previously proposed perceptually constrained GSVD (PCGSVD)-based approach can effectively improve the quality and intelligibility of the estimated speech, regardless of whether the additive noise is stationary or not, especially when it is non-white. However, the high computational complexity of such transformation-based signal subspace approach makes it difficult to be practically applied in the real-world environments. In this paper, we adopt the concept of the PCGSVD-based approach as proposed previously to develop a perceptually constrained minimum variance estimation (PCMVE) algorithm for speech enhancement, in which the time-shift property of DFT is applied with the special structure of Hankel matrices to replace the GSVD operation, such that the required computation load can be as low as the conventional SS algorithm. Experiments showed this proposed approach can offer comparable performance as that of the PCGSVD-based approach, and provides better functions than the conventional SS algorithm.

2. Brief summary of the auditory masking thresholds (AMTs) evaluation

The procedures for evaluating the AMTs for human perception are well known, as briefly summarized here [1,3]. The perceptible frequency range for human auditory system can be modeled by 25 critical bands. We first add up the magnitude square of the corresponding DFT components of the input clean speech signal in each critical band, convolve the critical band energy sequence with a spreading function 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 frame. The AMTs are finally obtained by renormalizing the above 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.

3. The proposed PCMVE-based speech enhancement approach

Let the i^{th} sample of the input noisy speech signal $\boldsymbol{y_I}$ can be expressed by samples of the clean speech $\boldsymbol{d_C}$ and the noise $\boldsymbol{n_I}$:

$$y_i = d_i + n_i, \quad i=0,1,2\ldots, \tag{1}$$

and the goal here is to estimate $\boldsymbol{d_C}$ from $\boldsymbol{y_I}$. The framework of the proposed PCMVE-based speech enhancement approach includes four phases as described below.

3.1. Phase (I): Framer, non-speech detector and buffer

The input speech $\boldsymbol{y_I}$ is first segmented into overlapped frames via a window function of length M, and then the speech enhancement process is repeated for each frame. A frame-synchronous voice activity detection (VAD) algorithm is used to identify and accumulate the non-speech frames of the input signal, and hence the noise statistics can be estimated for further processing. In this paper we employed the ETSI-AFE silence detection algorithm [4] for VAD, which contains two stages. The whole spectrum, sub-bands spectrum, and spectral variance are evaluated in the first stage in a frame-by-frame basis, followed by a decision stage based on a speech likelihood functions, and a final decision is then made in the second stage retrospectively to the earliest frame in the buffer.

3.2. Phase (II): Construction of Hankel-form matrices

Two series of Hankel-form matrices of order LxK, $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}} \in \mathbf{R}^{\mathrm{LxK}}$, are then constructed [1], $\boldsymbol{H_Y}$ for a noisy speech frame of $\boldsymbol{y_I}$ and $\boldsymbol{H}_{\hat{N}}$ for the latest buffered non-speech frame $\hat{\boldsymbol{n}}_{\boldsymbol{I}}$, where L+K-1 equals the frame size M and in general K is much smaller than L. Under noise free situation, the value of K is chosen such that the matrix $\boldsymbol{H_Y}$ is rank deficient, i.e. the rank of $\boldsymbol{H_Y}$ is smaller than K. This rank deficiency condition makes it easier to divide the signal and noise subspace later on when the additive noise is presented. From equation (1), it is clear that the matrix $\boldsymbol{H_Y}$ can be represented as the summation of two Hankel-form matrices $\boldsymbol{H_D}$ and $\boldsymbol{H_N}$, $\boldsymbol{H_Y} = \boldsymbol{H_D} + \boldsymbol{H_N}$, which are respectively constructed from the clean speech frame and the real noise frame. Both $\boldsymbol{H_D}$ and $\boldsymbol{H_N}$ are of course unknown, yet $\boldsymbol{H_N}$ can be approximated by $\boldsymbol{H}_{\hat{N}}$, which is constructed above with the latest buffered non-speech frame of the input signal. In phase (III) below we utilize the time-shift property of DFT and the concept of perceptually constrained signal/noise subspace to estimate the clean speech signal from the given matrices $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$.

3.3. Phase (III): Applying DFT to the Hankel-form matrices $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$

The DFT of an M-point sample sequence y_m, m=0,1,..,M-1, is

$$Y_k = \sum_{m=0}^{\mathrm{M}-1} y_m \cdot W_{\mathrm{M}}^{-mk}, \quad 0 \le k \le \mathrm{M}-1, \tag{2}$$

where $W_{\mathrm{M}} \equiv e^{j2\pi/\mathrm{M}}$, and $\boldsymbol{Y}_k$, k=0,1,..,M-1, is the obtained k^{th} frequency component. The time-shift property of DFT is useful here,

$$y_{m-i} \xleftrightarrow{\text{DFT}} Y_k \cdot W_{\mathrm{M}}^{-ik}, \; 0 \le m,k \le \mathrm{M}-1. \tag{3}$$

After padding M-L zeros at the end of each column of the matrix $\boldsymbol{H_Y}$ obtained above, the zero-padded augmented matrix of $\boldsymbol{H_Y}$ is denoted as $\boldsymbol{H_{Y,aug}}$. We then perform an M-sample DFT on each column of the matrix $\boldsymbol{H_{Y,aug}}$ and the above time-shift property leads to the following result:

$$\begin{aligned}
\boldsymbol{W}_1^T \cdot \boldsymbol{H_{Y,aug}} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W_{\mathrm{M}}^{-1} & \cdots & W_{\mathrm{M}}^{-(\mathrm{M}-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & W_{\mathrm{M}}^{-(\mathrm{M}-1)} & \cdots & W_{\mathrm{M}}^{-(\mathrm{M}-1)^2} \end{bmatrix}_{\mathrm{MxM}} \cdot \begin{bmatrix} y_0 & y_1 & \cdots & y_{\mathrm{K}-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{\mathrm{L}-1} & y_{\mathrm{L}} & \cdots & y_{\mathrm{M}-1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}_{\mathrm{MxK}} \\
&\cong \begin{bmatrix} Y_0 & Y_0 & \cdots & Y_0 \\ Y_1 & Y_1 \cdot W_{\mathrm{M}}^{-1} & \cdots & Y_1 \cdot W_{\mathrm{M}}^{-(\mathrm{K}-1)} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{\mathrm{M}-1} & Y_{\mathrm{M}-1} \cdot W_{\mathrm{M}}^{-(\mathrm{M}-1)} & \cdots & Y_{\mathrm{M}-1} \cdot W_{\mathrm{M}}^{-(\mathrm{M}-1)(\mathrm{K}-1)} \end{bmatrix}_{\mathrm{MxK}} \\
&= \begin{bmatrix} Y_0 & 0 & \cdots & 0 \\ 0 & Y_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Y_{\mathrm{M}-1} \end{bmatrix}_{\mathrm{MxM}} \cdot \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & W_{\mathrm{M}}^{-1} & \cdots & W_{\mathrm{M}}^{-(\mathrm{K}-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & W_{\mathrm{M}}^{-(\mathrm{M}-1)} & \cdots & W_{\mathrm{M}}^{-(\mathrm{M}-1)(\mathrm{K}-1)} \end{bmatrix}_{\mathrm{MxK}} \equiv \boldsymbol{\Lambda}_Y \cdot \boldsymbol{W}_2,
\end{aligned} \tag{4}$$

where $\boldsymbol{W}_1 \in \mathbf{Z}^{\mathrm{MxM}}$ and $\boldsymbol{W}_2 \in \mathbf{Z}^{\mathrm{MxK}}$ are two DFT transformation matrices, and $\boldsymbol{\Lambda}_Y \in \mathbf{Z}^{\mathrm{MxM}}$ is a diagonal matrix whose diagonal elements are the frequency components of the input noisy speech frame. Similar decomposition procedure can be performed on the matrix $\boldsymbol{H}_{\hat{N}}$. Therefore we can simultaneously diagonalize the augmented versions of the matrices $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$ with the transformation matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$:

$$\begin{aligned}
\boldsymbol{W}_1^T \cdot \boldsymbol{H_{Y,aug}} \cdot \boldsymbol{W}_2^{-R} &= diag(Y_0,...,Y_{\mathrm{M}-1}) = \boldsymbol{\Lambda}_Y \\
\boldsymbol{W}_1^T \cdot \boldsymbol{H}_{\hat{N},aug} \cdot \boldsymbol{W}_2^{-R} &= diag(\hat{N}_0,...,\hat{N}_{\mathrm{M}-1}) = \boldsymbol{\Lambda}_{\hat{N}},
\end{aligned} \tag{5}$$

where the matrix $\boldsymbol{\Lambda}_Y$ is as defined in equation (4), the diagonal elements $\hat{N}_k$, k=0,1,..,M-1, of the matrix $\boldsymbol{\Lambda}_{\hat{N}} \in \mathbf{Z}^{\mathrm{MxM}}$ are the frequency components of the latest buffed noise signal, $\boldsymbol{W}_2^{-R} \in \mathbf{Z}^{\mathrm{KxM}}$ is the right pseudo-inverse of the matrix $\boldsymbol{W}_2$, and $\boldsymbol{H}_{\hat{N},aug}$ is the zero-padded augmented version of $\boldsymbol{H}_{\hat{N}}$. The result of equation (5) is almost identical to the GSVD algorithm and thus the following enhancement procedures are nearly the same as the PCGSVD-based approach [1].

3.4. Phase (IV): PCMVE-based signal-subspace construction in frequency domain

We first split the diagonal elements of the matrices $\boldsymbol{\Lambda}_Y$ into two non-overlapped sets; the principal set $\boldsymbol{\Omega}_1$ (for those components of Y_k and $\hat{N}_k$, $0 \le k \le \mathrm{M}-1$, such that $|Y_k| \ge |\hat{N}_k|$; associated with the signal subspace of $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$) and the minor set $\boldsymbol{\Omega}_2$ (for those Y_k and $\hat{N}_k$ do not belong to $\boldsymbol{\Omega}_1$; associated with the noise subspace of $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$). The dimension of the signal subspace is N, 1≤N≤M. Therefore the signal subspace of $\boldsymbol{H_{Y,aug}}$ and $\boldsymbol{H}_{\hat{N},aug}$ is constituted from the N orthogonal columns of the matrix $\boldsymbol{W}_1$ associating with the index of Y_k, k=0,1,..,M-1, belonging to the principal set $\boldsymbol{\Omega}_1$, and the rest M-N orthogonal columns of which span the noise subspace of $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$. Now we can apply the PCMVE algorithm to weight the features of the principal set $\boldsymbol{\Omega}_1$ and null those of the minor set $\boldsymbol{\Omega}_2$. The concept of PCMVE is to find a transformation matrix $\hat{\boldsymbol{P}} \in \mathbf{R}^{\mathrm{KxK}}$ such that the Frobenius distance of the two matrices $\boldsymbol{H_{Y,aug}} \cdot \hat{\boldsymbol{P}}$ and $\boldsymbol{H_{D,aug}}$ (zero-

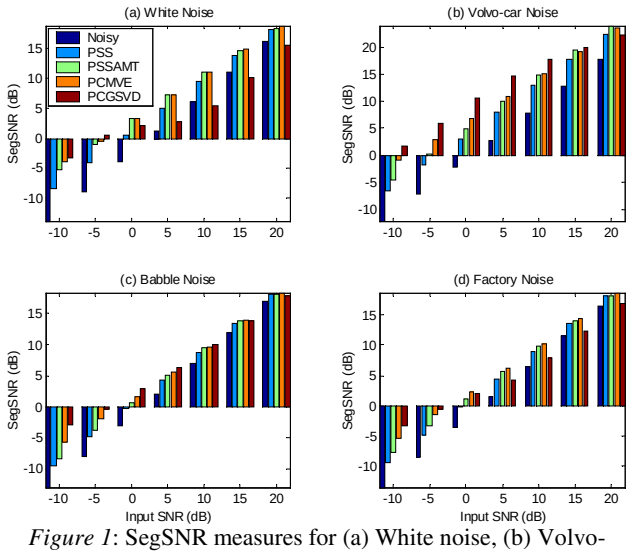

Figure 1: SegSNR measures for (a) White noise, (b) Volvo-car noise, (c) Babble noise, and (d) Factory noise

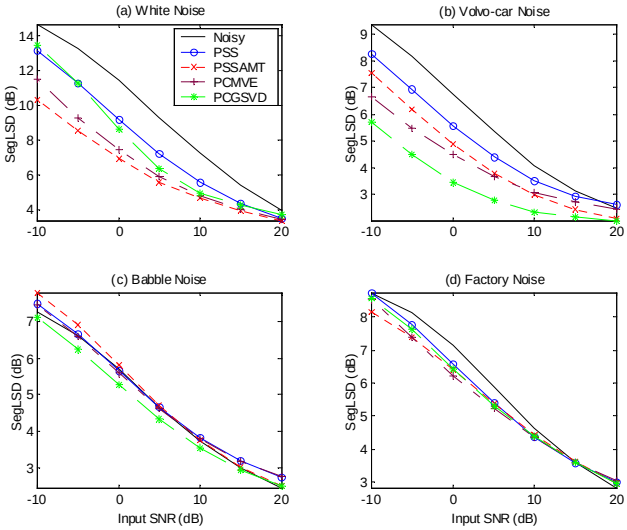

Figure 2: SegLSD measures for (a) White noise, (b) Volvo-car noise, (c) Babble noise, and (d) Factory noise

padded augmented matrix of $\boldsymbol{H_D}$) is minimized, under the constrains that the averaged energies of the N frequency components (corresponding to those belong to $\boldsymbol{\Omega}_1$) of the residual noise signal (referred to as the matrix $\boldsymbol{H}_{\hat{N},aug} \cdot \hat{\boldsymbol{P}}$) do not exceed the corresponding AMTs, and are zero for the rest K-N frequency components. Hence there are N inequality constrains and K-N equality ones for the PCMVE-based algorithm, as illustrated below:

$$\hat{\boldsymbol{P}} = \arg\min_{\hat{\boldsymbol{P}} \in \mathbf{R}^{\mathrm{KxK}}} \left\| \boldsymbol{H}_{Y,aug} \cdot \hat{\boldsymbol{P}} - \boldsymbol{H}_{D,aug} \right\|_{\mathrm{F}}^2, \text{ subject to} \tag{6}$$

$$\left\{ \begin{array}{l} \left\| \boldsymbol{w}_{1,k}^T \cdot \boldsymbol{H}_{\hat{N},aug} \cdot \hat{\boldsymbol{P}} \right\|^2 \le \gamma_k \cdot \left\| \boldsymbol{w}_{2,k} \right\|^2 = \mathrm{K} \cdot \gamma_k, \ \forall\ \boldsymbol{w}_{1,k} \in \text{sig. subsp} \\ \left\| \boldsymbol{w}_{1,k}^T \cdot \boldsymbol{H}_{\hat{N},aug} \cdot \hat{\boldsymbol{P}} \right\|^2 = 0, \ \ \forall\ \boldsymbol{w}_{1,k} \in \text{noise subsp}, \quad 0 \le k \le \mathrm{M}-1 \end{array} \right\},$$

where $\|\boldsymbol{A}\|_{\mathrm{F}}^2 \equiv \sum_{i=0}^{\mathrm{L-1}} \sum_{j=0}^{\mathrm{K-1}} \left| \boldsymbol{A}_{i,j} \right|^2$ is the Frobenius norm of a matrix $\boldsymbol{A} \in \mathbf{Z}^{\mathrm{LxK}}$, the orthogonal vectors $\boldsymbol{w}_{1,k}$ and $\boldsymbol{w}_{2,k}$ are respectively the k^{th} column vector, k=0,1,..,M-1, of the matrices $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, and γ_k is the k^{th} AMT of the clean speech signal as evaluated in section 2. Though the clean speech signal cannot be known a priori, they can be estimated from the input signal (e.g., via the conventional SS algorithm). With the similar deviation of PCGSVD-based approach, it can be shown that the magnitude responses of the estimated clean speech $\left| \hat{Y}_k \right|, k = 0,1,..,\mathrm{M}-1$, are as follows:

$$\left| \hat{Y}_k \right| = \left\{ \begin{array}{ll} \left| Y_k \right| \cdot \min\left[(1 - \dfrac{\left| \hat{N}_k \right|^2}{\left| Y_k \right|^2}), \ \dfrac{\sqrt{\gamma_k}}{\left| \hat{N}_k \right|} \right], & for \quad Y_k \in \Omega_1 \\ 0, & for \quad Y_k \in \Omega_2 \end{array} \right\}, \tag{7}$$

where the elements $\left| Y_k \right|$ and $\left| \hat{N}_k \right|, k = 0,1,..,\mathrm{M}-1$, are the k^{th} magnitude response of the input speech frame and the estimated noise signal respectively. Unlike the previous PCGSVD-based approach in which the reconstructed Hankel-form matrix $\boldsymbol{H}_{\hat{D}}$ of the estimated clean speech is needed to obtain the enhanced speech frame [1], here directly applying the inverse DFT on the enhanced frequency components plus the original phase information of the input speech can give the enhanced speech frames. We can then concatenate them frame-by-frame with the overlap-add method to give the finally enhanced speech signal.

4. Experiments and performance evaluations

The experimental environment was as follows. The sampling rate of the input speech signal was 16kHz. We randomly selected 30 clean utterances, recorded by 2 females and 2 males, from TIMIT speech corpus for testing. Four types of noise source, 'White', 'Volvo-car', 'Babble (speech-like)', and 'Factory', chosen from NOISEX-92 database, were artificially added to the test speech. For each kind of noise source mentioned above, we generated noise-contaminated utterances with input SNR ranged from 20dB to -10dB with 5dB step size for evaluation. The babble and factory noises are non-stationary while the Volvo-car noise is stationary; all of the three are non-white. The frame size, M, was 512 samples (32 ms) with 50% frame overlap, and the row and column size (L and K) of the two Hankel-form matrices $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$ were 473 and 40 respectively. The window function for the framer operation in section 3.1 was Hamming. Under high SNR conditions (e.g. SNR>10dB) for various speech utterances, the dimension of the signal subspace for the matrix $\boldsymbol{H_Y}$ and $\boldsymbol{H}_{\hat{N}}$ practically ranged from 8 to 25. Therefore when the noise was present, the column size of 40 was adequate to decompose into the signal/noise subspace from the vector space of the matrix $\boldsymbol{H_Y}$. The computation complexity of the PCMVE-based approach is almost the same as the SS algorithm ($\cong 4\mathrm{M} \cdot \log_2^{\mathrm{M}}$) and much less than that of the PCGSVD-based method ($\cong 6\mathrm{L} \cdot \mathrm{K}^2 + 16\mathrm{K}^3$; Hint: L+K-1=M).

For comparison purposes, we also implemented the conventional spectral subtraction algorithm in the power spectral domain (PSS) [2], and a modified version of the PSS algorithm by involving the auditory masking thresholds (PSSAMT) of human hearing system [3].

Fig. 1 shows the obtained segmental SNR (SegSNR) measures [5]. For all the data presented below, where 'Noisy' is for noisy speech without any processing, while 'PSS', 'PSSAMT', 'PCMVE', and 'PCGSVD' are for the respective algorithms where 'PCMVE' is the approach proposed here. From Fig. 1(a) for white noise, we can see that in most cases PCMVE was slightly better than PSSAMT but very close, and

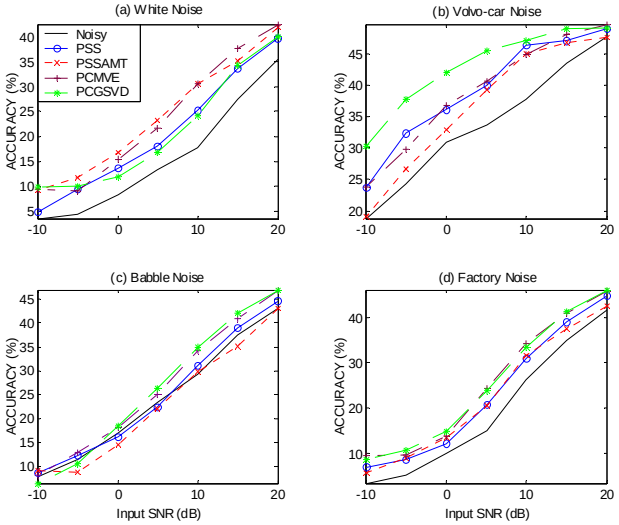

Figure 3: Accuracy measures for (a) White noise, (b) Volvo-car noise, (c) Babble noise, and (d) Factory noise

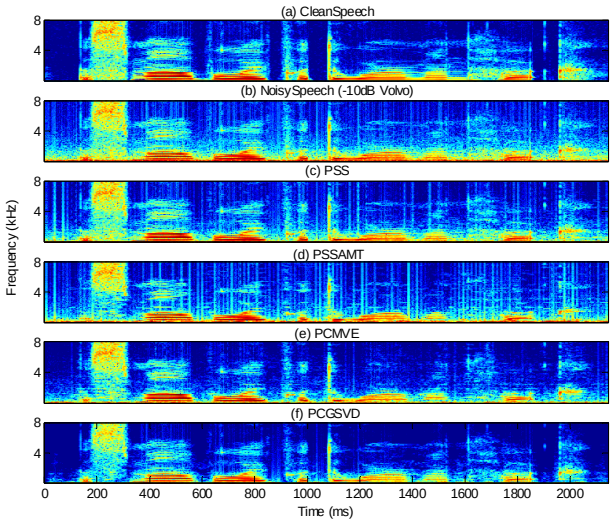

Figure 4: Spectrogram plots for a typical utterance corrupted by Volvo-car noise at -10dB of SNR

both of them superior to PSS and PCGSVD. However, for the Volvo-car noise case in Fig. 1(b), PCMVE and PCGSVD outperformed PSS and PSSAMT, specially when the input SNR was low. This implied the proposed PCMVE could offer some improvement in SegSNR, particularly when the noise is stationary but not white. Nevertheless, from Fig. 1(c) (babble noise) and (d) (factory noise), we see that when the noise was non-stationary, almost in all cases the SegSNR improvement of PCMVE was better than PSS and PSSAMT, and comparable to that of PCGSVD. Fig. 2 is the segmental log spectral distance (SegLSD) measure [5]. For the white noise case in Fig. 2(a), PCMVE was better than PSS and PCGSVD, but slightly worse than PSSAMT. In Fig. 2(b) to (d), however PCMVE outperformed PSS and PSSAMT for non-white noise in most cases.

The third measure is the English phoneme recognition accuracy obtained in the free-phoneme decoding (without lexicon and language model) for the 30 test utterances. The acoustic model consisted of 48 left-to-right continuous hidden Markov models (CHMMs) for 48 context-independent phoneme units, which were trained from the 4-hour TIMIT speech corpus and the total number of Gaussian mixtures was about 1100. The dimension of feature vectors was 39, including 12 MFCCs, normalized log energy, and their first and second derivatives. The front-end, the acoustic model training, and the recognizer were adopted from HTK. The baseline phoneme accuracy for the 30 test sentences was 54.48%. Fig. 3 depicts the recognition results. As we can see in Fig. 3(a), (b), and (d), all of the four enhancement approaches under evaluation can improve the recognition accuracy for the white, Volvo-car, and factory noise cases; but PCMVE and PCGSVD offered the best results for the non-white noises, which was also true for the babble noise case in Fig. 3(c). The recognition performance could be further improved if the training corpus for the acoustic model can be similarly processed a priori.

Fig. 4 is the spectrogram for a test utterance for the case of Volvo-car noise at -10dB. From Fig. 4(c) and (d) we can see that in PSS and PSSAMT processed speech some undesired random tone peaks were present in the non-speech regions and the high frequency part (>2KHz) of the voiced segments, or perceivable as the *musical noise*. This was significantly improved in the PCMVE and PCGSVD processed utterances in Fig. 4(e) and (f) respectively. Moreover, in Fig. 4(e) almost the same detailed information of the speech spectrum were retained by PCMVE as that of Fig. 4(c) by PSS, but with much less random tone peaks in both the silenced segments and the high frequency part of the voiced regions. This was actually verified by informal subjective listing tests, in which many subjects confirmed that the *musical noise* is less perceivable for the utterances processed by PCMVE than those by PSS and PSSAMT, and the former sounded similar to the PCGSVD-processed speech.

5. Conclusions

In this paper, we proposed a new algorithm for speech enhancement to integrate the masking-based psychoacoustics model and the minimum variance estimate technique in frequency domain. This new approach requires much less computational complexity than that of the PCGSVD-based method, very close to that of the conventional SS algorithm, but offers comparable performance to the PCGSVD-based approach, regardless of whether the noise is stationary or not, specially when the noise is non-white.

6. References

[1] 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*, pp. 533–536, Geneva, Switzerland, Sep. 2003.

[2] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No. 2, pp. 113-120, April 1979.

[3] N. Virag, "Single Channel Speech Enhancement based on Masking Properties of the Human Auditory System", *IEEE Trans. on Speech and Audio Processing*, Vol. 7, pp. 126–137, March 1999.

[4] ETSI ES 202 050 V1.1.3 Recomm., "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm, Compression Algorithms", 2003.

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