NOISE EIGENVALUE MODIFICATION METHODS FOR SPATIAL SUBSPACE BASED MULTI-CHANNEL SPEECH ENHANCEMENT

Gibak Kim and Nam Ik Cho

School of Electrical Engineering
Seoul National University, Seoul 151-744 Korea
email: kgb@ispl.snu.ac.kr, nicho@snu.ac.kr

ABSTRACT

In this paper, frequency domain multi-channel filtering schemes are proposed for speech enhancement, based on the subspace decomposition of spatial spectral matrices. For better estimation of noise statistics, which is important for most speech enhancers, we propose noise eigenvalue modification methods for the correction of noise spatial spectral matrix. These methods are based on the rank-1 property of the speech spatial spectral matrix for single desired speech source. Simulation results show that the proposed methods yield better performance compared to the conventional multi-channel Wiener filtering.

***Index Terms*—** Microphone array, multi-channel filtering, speech enhancement

1. INTRODUCTION

Noise reduction is one of the most important elements in speech communication and recognition systems. Particularly in the case of distant or hands-free speech acquisition, the system performance is severely degraded due to ambient noises. Hence, there have been many researches on noise suppression, and diverse techniques for a single or multiple microphones have been developed for several decades. In multi-microphone system, noise signals can be reduced by beamforming techniques when the speech and noise signals arrive from different directions. Beamforming based spatial filtering techniques yield more noise reduction and less distortortion compared to single microphone techniques. Recently, the multi-channel Wiener filtering has been developed, which is shown to have better performance than the standard beamforming techniques [1–3].

The estimate of noise statistics is essential in the multi-channel Wiener filter as in most noise reduction filters. The second-order statistics of the noise signal can be estimated from the observed data during noise-only period and also used in subsequent speech-present period under the assumption that the second-order statistics of noise is slowly time varying compared to the speech signal [2,3]. However this assumption results in large amount of errors in the case of highly non-stationary noise such as a competing speech noise. For improving the performance of multi-channel noise reduction, we propose three methods that estimate the noise statistics in speech-present period by correcting the noise statistics estimated from noise-only period. These methods are based on the rank-1 property of the speech spatial spectral matrix for single desired speech signal. In summary, the overall filtering scheme is to perform multi-channel Wiener filtering in the frequency domain by decomposing the spatial spectral matrices, where the decomposed noise eigenvalues are modified.

This paper is organized as follows. Section 2 describes the signal model and briefly reviews the frequency domain multi-channel Wiener filtering for noise reduction. We propose three methods for correcting the noise spatial spectral matrix by modifying the noise eigenvalues in Section 3. Simulation results and performance evaluation are shown in Section 4.

2. FREQUENCY DOMAIN MULTI-CHANNEL WIENER FILTERING

Let us consider an M-channel signal model where a speech source is convolved with M room acoustic transfer functions to every microphone, and each microphone signal is corrupted by additive noise. Then, the signal model can be expressed as

$$y_i[k] = h_i[k] * s[k] + n_i[k] = x_i[k] + n_i[k] \quad i = 1, \ldots, M \tag{1}$$

where $y_i[k]$ denotes the observed signal at the ith microphone at time k, $x_i[k]$ and $n_i[k]$ are speech and additive noise component respectively, $s[k]$ is the desired speech source, and $h_i[k]$ is the acoustic transfer function from the speech source to the ith microphone. Assuming infinite filter lengths, (1) is represented in the frequency domain as

$$\begin{aligned}\mathbf{Y}(f) &= \begin{bmatrix} Y_1(f) \\ Y_2(f) \\ \vdots \\ Y_M(f) \end{bmatrix} = S(f) \begin{bmatrix} H_1(f) \\ H_2(f) \\ \vdots \\ H_M(f) \end{bmatrix} + \begin{bmatrix} N_1(f) \\ N_2(f) \\ \vdots \\ N_M(f) \end{bmatrix} \\ &= S(f)\mathbf{H}(f) + \mathbf{N}(f) = \mathbf{X}(f) + \mathbf{N}(f) \end{aligned} \tag{2}$$

where $Y_i(f)$, $H_i(f)$, $S(f)$, $N_i(f)$, $X_i(f)$ are frequency domain representations of $y_i[k]$, $h_i[k]$, $s[k]$, $n_i[k]$, $x_i[k]$, respectively. For the multi-channel filter $\mathbf{W}(f)$, the output signal $Z(f)$ can be written as

$$Z(f) = \mathbf{W}^H(f)\mathbf{Y}(f). \tag{3}$$

If we estimate the speech component from the 1st microphone signal in the minimum mean square error (MMSE) sense, the frequency domain multi-channel Wiener filter can be expressed in terms of spatial spectral matrices as

$$\begin{aligned}\mathbf{W}(f) &= \mathbf{R}_{\mathbf{YY}}^{-1}(f)\mathbf{R}_{\mathbf{XX}}(f)\mathbf{e}_1 \\ &= \mathbf{R}_{\mathbf{YY}}^{-1}(f)\left(\mathbf{R}_{\mathbf{YY}}(f) - \mathbf{R}_{\mathbf{NN}}(f)\right)\mathbf{e}_1 \end{aligned} \tag{4}$$

$$\mathbf{R}_{\mathbf{YY}}(f) = E\left\{\mathbf{Y}(f)\mathbf{Y}^H(f)\right\} \tag{5}$$

$$\mathbf{R}_{\mathbf{XX}}(f) = E\left\{\mathbf{X}(f)\mathbf{X}^H(f)\right\} \tag{6}$$

$$\mathbf{R}_{\mathbf{NN}}(f) = E\left\{\mathbf{N}(f)\mathbf{N}^H(f)\right\} \tag{7}$$

with $\mathbf{e}_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T$. For the practical implementation, the speech spatial spectral matrix $\mathbf{R_{XX}}(f)$ is approximated as

$$\mathbf{R_{XX}}(f) \simeq \mathbf{R_{YY}}(f) - \mathbf{R_{NN}}(f) \tag{8}$$

under the statistical independence assumption between speech and noise. In conventional algorithms, the noise spatial spectral matrix $\mathbf{R_{NN}}(f)$ is estimated and updated during noise-only period. The update stops at speech-present period, and the most recent noise estimate is used until the next noise-only period. This scheme is based on the assumption that the noise is stationary or slowly time varying.

3. SUBSPACE BASED MULTI-CHANNEL WIENER FILTERING IN THE FREQUENCY DOMAIN

3.1. Spatial Subspace Decomposition

The spatial subspace is obtained by subspace decomposition of multi-channel input in frequency domain. The spatial spectral matrices $\mathbf{R_{YY}}(f)$ and $\mathbf{R_{NN}}(f)$ can be jointly diagonalized by solving the generalized eigenvalue problem [4] as

$$\mathbf{R_{YY}}(f)\mathbf{Q}(f) = \mathbf{R_{NN}}(f)\mathbf{Q}(f)\mathbf{\Lambda}(f), \tag{9}$$

$$\begin{cases} \mathbf{Q}(f)^H\mathbf{R_{YY}}(f)\mathbf{Q}(f) = \mathbf{\Lambda_Y}(f) \\ \mathbf{Q}(f)^H\mathbf{R_{NN}}(f)\mathbf{Q}(f) = \mathbf{\Lambda_N}(f) \end{cases} \tag{10}$$

where $\mathbf{\Lambda}(f), \mathbf{\Lambda_Y}(f), \mathbf{\Lambda_N}(f)$ are diagonal matrices as

$$\mathbf{\Lambda}(f) = \mathrm{diag}\left\{\lambda_1(f)\ \ \lambda_2(f)\ \cdots\ \lambda_M(f)\right\} \tag{11}$$

$$\mathbf{\Lambda_Y}(f) = \mathrm{diag}\left\{\lambda_{Y,1}(f)\ \ \lambda_{Y,2}(f)\ \cdots\ \lambda_{Y,M}(f)\right\} \tag{12}$$

$$\mathbf{\Lambda_N}(f) = \mathrm{diag}\left\{\lambda_{N,1}(f)\ \ \lambda_{N,2}(f)\ \cdots\ \lambda_{N,M}(f)\right\} \tag{13}$$

with $\mathbf{\Lambda}(f) = \mathbf{\Lambda_Y}(f)\mathbf{\Lambda_N}^{-1}(f)$, $\lambda_i(f) = \frac{\lambda_{Y,i}(f)}{\lambda_{N,i}(f)}$, $\lambda_1(f) > \lambda_2(f) > \cdots > \lambda_M(f)$ and $\mathbf{Q}(f)$ is an invertible, but not necessarily orthogonal matrix. Then the spatial spectral matrices are expressed as

$$\begin{cases} \mathbf{R_{YY}}(f) = \bar{\mathbf{Q}}(f)\mathbf{\Lambda_Y}(f)\bar{\mathbf{Q}}^H(f) \\ \mathbf{R_{NN}}(f) = \bar{\mathbf{Q}}(f)\mathbf{\Lambda_N}(f)\bar{\mathbf{Q}}^H(f) \end{cases} \tag{14}$$

with $\bar{\mathbf{Q}}(f) = \mathbf{Q}^{-H}(f)$. By substituting (14) into (4) the frequency domain multi-channel Wiener filter is obtained as

$$\mathbf{W}(f) = \mathbf{Q}(f)\left(\mathbf{I} - \mathbf{\Lambda_Y}^{-1}(f)\mathbf{\Lambda_N}(f)\right)\bar{\mathbf{Q}}^H(f)\mathbf{e}_1. \tag{15}$$

3.2. Noise Eigenvalue Modification

We propose several methods for the modification of noise eigenvalues, based on the inherent rank-1 property of the speech spatial spectral matrix for single desired speech source, which can alleviate the performance degradation resulting from the error of noise spatial spectral matrix estimate. In principle, when the multi-channel speech components are generated by multiplication of each acoustic transfer function and a single speech source in the frequency domain as in (2), the speech spatial spectral matrix can be written as

$$\mathbf{R_{XX}}(f) = E\left\{\mathbf{X}(f)\mathbf{X}^H(f)\right\} = E\left\{S(f)S^*(f)\right\}\mathbf{H}^H(f)\mathbf{H}(f) \tag{16}$$

and the rank of speech spatial spectral matrix $\mathbf{R_{XX}}(f)$ is equal to 1. From (8) and (14), the estimate of the speech spatial spectral matrix is rewritten as

$$\mathbf{R_{XX}}(f) \simeq \bar{\mathbf{Q}}(f)\left(\mathbf{\Lambda_Y}(f) - \mathbf{\Lambda_N}(f)\right)\bar{\mathbf{Q}}^H(f). \tag{17}$$

Since $\bar{\mathbf{Q}}(f)$ is a full rank matrix, the rank of $(\mathbf{\Lambda_Y}(f) - \mathbf{\Lambda_N}(f))$ should be equal to 1 in order to make the speech estimate matrix to be of rank-1. However, the rank of $(\mathbf{\Lambda_Y}(f) - \mathbf{\Lambda_N}(f))$ deviates from 1 due to the error in noise spatial spectral matrix estimate especially when the noise spatial spectral matrix is estimated during noise-only period and kept unchanged in speech-present period. As stated previously, the main purpose of this paper is to propose methods that updates noise spatial spectral matrix in the frequency domain, even in the speech-present period. To be precise, let $\tilde{\mathbf{R}}_{\mathbf{NN}}(f)$ and $\tilde{\mathbf{\Lambda}}_{\mathbf{N}}(f)$ be the noise spatial spectral matrix and the noise eigenvalue matrix estimated during the noise-only period. Also, let us denote the modified noise eigenvalue matrix and the modified speech spatial spectral matrix as $\hat{\mathbf{\Lambda}}_{\mathbf{N}}(f)$ and $\hat{\mathbf{R}}_{\mathbf{XX}}(f)$, respectively. Three different approaches are proposed in this paper:

3.2.1. Least Squares Estimate (Method 1)

The first method is to find the least square estimate, i.e., to find $\hat{\mathbf{\Lambda}}_{\mathbf{N}}(f)$ from

$$\min_{\mathrm{rank}\left\{\hat{\mathbf{R}}_{\mathbf{XX}}(f)\right\}=1} \left\| \mathbf{\Lambda_Y}(f)\tilde{\mathbf{\Lambda}}_{\mathbf{N}}^{-1}(f) - \mathbf{\Lambda_Y}(f)\hat{\mathbf{\Lambda}}_{\mathbf{N}}^{-1}(f) \right\|_F^2. \tag{18}$$

Then the least squares estimate is described as

Method 1:

$$\hat{\mathbf{\Lambda}}_{\mathbf{N}}(f) = \mathrm{diag}\left\{\tilde{\lambda}_{N,1}(f)\ \ \lambda_{Y,2}(f)\ \ \lambda_{Y,3}(f)\ \cdots\ \lambda_{Y,M}(f)\right\}. \tag{19}$$

This estimate retains the signal spatial subspace with the largest signal to noise ratio and removes the other noise spatial subspace. In this case we get the frequency domain multi-channel Wiener filter derived in [5] as

$$\mathbf{W}(f) = \mathbf{q}_1(f)\left(1 - \frac{1}{\lambda_1(f)}\right)\bar{\mathbf{q}}_1^H(f)\mathbf{e}_1 \tag{20}$$

where $\mathbf{q}_1(f)$ and $\bar{\mathbf{q}}_1(f)$ are respectively the column of $\mathbf{Q}(f)$ and $\bar{\mathbf{Q}}(f)$ corresponding to the largest generalized eigenvalue $\lambda_1(f)$.

3.2.2. Time Invariant Spatial Coherence (Method 2 and 3)

The other two methods are derived from the time invariant spatial coherence assumption. If we assume a homogeneous noise field, i.e., $P_{N_iN_i}(f) = P_N(f), \forall i = 1, \cdots, M$ with $P_N(f)$ the power spectral density of noise and $P_{N_iN_j}(f) = E\left\{N_i(f)N_j^*(f)\right\}$, the noise spatial spectral matrix can be expressed in terms of the spatial coherence matrix as

$$\mathbf{R_{NN}}(f) = P_N(f)\mathbf{\Gamma_N}(f) \tag{21}$$

with spatial coherence matrix

$$\mathbf{\Gamma_N}(f) = \begin{bmatrix} 1 & \Gamma_{N_1N_2}(f) & \cdots & \Gamma_{N_1N_M}(f) \\ \Gamma_{N_2N_1}(f) & 1 & \cdots & \Gamma_{N_2N_M}(f) \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{N_MN_1}(f) & \Gamma_{N_MN_2}(f) & \cdots & 1 \end{bmatrix}. \tag{22}$$

The elements of the spatial coherence matrix are the complex coherence functions between two microphone signals as

$$\Gamma_{N_iN_j}(f) = \frac{P_{N_iN_j}(f)}{\sqrt{P_{N_iN_i}(f)P_{N_jN_j}(f)}}. \tag{23}$$

The spatial coherence matrix depends mainly on the microphone array configuration, the position of sound source, and the acoustic environment, not the spectral characteristics of signal. Generally the spatial coherence matrix is slowly time varying compared to spectral characteristics. Therefore it is more realistic to assume that the spatial coherence matrix is short-term stationary than to assume that noise spatial spectral matrix is short-term stationary. Thus, let us assume that the noise spatial coherence matrix $\mathbf{\Gamma_N}(f)$ is short-term stationary and let the noise spatial spectral matrix during speech-present period as

$$\tilde{\mathbf{R}}_{\mathbf{NN}}(f) = \tilde{P}_N(f)\mathbf{\Gamma_N}(f) \tag{24}$$

where $\tilde{P}_N(f)$ is the power spectral densigy of noise during noise-only period. Then the noise spatial spectral matrix and noise eigenvalue matrix can be written as

$$\mathbf{R}_{\mathbf{NN}}(f) = \alpha(f)\tilde{\mathbf{R}}_{\mathbf{NN}}(f) \tag{25}$$

$$\mathbf{\Lambda_N}(f) = \alpha(f)\tilde{\mathbf{\Lambda}}_{\mathbf{N}}(f) \tag{26}$$

with $\alpha(f) = P_N(f)/\tilde{P}_N(f)$.

From (8),(14), and (26), the estimate of the speech matrix is rewritten as

$$\mathbf{R_{XX}}(f) \simeq \bar{\mathbf{Q}}(f)\left(\mathbf{\Lambda_Y}(f) - \alpha(f)\tilde{\mathbf{\Lambda}}_{\mathbf{N}}(f)\right)\bar{\mathbf{Q}}^H(f) \tag{27}$$

with $\hat{\lambda}_{N,i}(f) = \alpha(f)\tilde{\lambda}_{N,i}(f)$. If the rank of speech spatial spectral matrix is one, the whole power of speech signal lies in the speech spatial subspace which is the column vector of $\bar{\mathbf{Q}}(f)$ corresponding to the largest generalized eigenvalue. Since there is no speech power in the noise subspace, the generalized eigenvalues ($\lambda_i(f), i = 2, \cdots, M$: the ratio of the speech+noise eigenvalue to the noise eigenvalue) in the noise subspace are equal to 1. In practice, the noise subspace generalized eigenvalues deviate from 1 and we can estimate $\alpha(f)$ by reducing this deviation in the least squares sense as

$$\min\left[\sum_{i=2}^{M}\left(\frac{\lambda_{Y,i}(f)}{\alpha(f)\tilde{\lambda}_{N,i}(f)} - 1\right)^2\right]. \tag{28}$$

Then the noise eigenvalue modification is obtained as

Method 2:

$$\hat{\mathbf{\Lambda}}_{\mathbf{N}}(f) = \mathrm{diag}\left\{\alpha(f)\tilde{\lambda}_{N,1}(f)\ \ \lambda_{Y,2}(f)\ \ \lambda_{Y,3}(f)\ \cdots\ \lambda_{Y,M}(f)\right\} \tag{29}$$

$$\alpha(f) = \frac{\sum_{i=2}^{M}\left(\frac{\lambda_{Y,i}(f)}{\tilde{\lambda}_{N,i}(f)}\right)^2}{\sum_{i=2}^{M}\frac{\lambda_{Y,i}(f)}{\tilde{\lambda}_{N,i}(f)}} \tag{30}$$

where 2nd $\sim M$th noise eigenvalues are set to the signal eigenvalues to make the speech spatial spectral matrix rank one. Another estimate of noise modification factor $\alpha(f)$ is obtained by setting the second largest generalized eigenvalue to one as

Method 3:

$$\frac{\lambda_{Y,2}(f)}{\hat{\lambda}_{N,2}(f)} = 1\ \rightarrow\ \alpha(f) = \frac{\lambda_{Y,2}(f)}{\tilde{\lambda}_{N,2}(f)}. \tag{31}$$

The second largest generalized eigenvalue is the largest power ratio of signal to noise component in the noise spatial subspace, and (31) removes the speech component in all the noise subspace, which tends to overestimate the noise power. The proposed three methods can be represented in a unified notation as

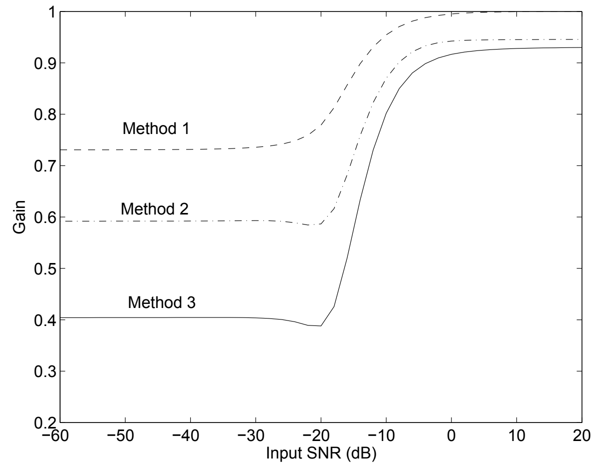

Fig. 1. Filter gains in the signal subspace: competing speech noise.

$$\hat{\mathbf{\Lambda}}_{\mathbf{N}}(f) = \mathrm{diag}\left\{\alpha(f)\tilde{\lambda}_{N,1}(f)\ \ \lambda_{Yy,2}(f)\ \ \lambda_{Y,3}(f)\ \cdots\ \lambda_{Y,M}(f)\right\}$$

$$\alpha(f) = \begin{cases} 1 & : \mathrm{Method1} \\ \frac{\sum_{i=2}^{M}\left(\frac{\lambda_{Y,i}(f)}{\tilde{\lambda}_{N,i}(f)}\right)^2}{\sum_{i=2}^{M}\frac{\lambda_{Y,i}(f)}{\tilde{\lambda}_{N,i}(f)}} & : \mathrm{Method2} \\ \frac{\lambda_{Y,2}(f)}{\tilde{\lambda}_{N,2}(f)} & : \mathrm{Method3.} \end{cases} \tag{32}$$

and the filter vector $\mathbf{W}(f)$ can be written with the principal eigenvector and the principal spatial subspace filter gain $G(f)$ as

$$\mathbf{W}(f) = \mathbf{q}_1(f)G(f)\bar{q}_1^*(f) \tag{33}$$

$$G(f) = 1 - \frac{\alpha(f)\tilde{\lambda}_{N,1}(f)}{\lambda_{Y,1}(f)} \tag{34}$$

where $\bar{q}_1^*(f)$ is the complex conjugate of the first element of $\bar{\mathbf{q}}_1(f)$.

4. SIMULATION RESULTS

4.1. Simulation data

In the simulations, we generate multi-channel noisy signals by adding noise to the speech. The reverberant multi-channel signals are generated by the convolution of dry source (sound data measured in an anechoic room) with acoustic impulse responses from the RWCP Sound Scene Database [6]. The impulse responses are measured at several positions which are 2m distance from the microphone array with reverberation time of 300 ms. The microphone array is a linear type and has 7 transducers located at 2.83cm uniform intervals. In this simulation, speech signal is convolved with the impulse response measured at the fore side of the microphone array and added by a competing speech noise signal coming at the angle of 40°.

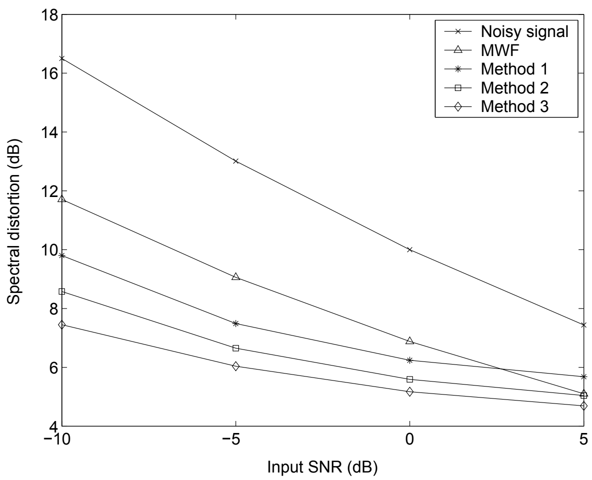

Fig. 2. Spectral distortion as a function of input SNR : competing speech noise.

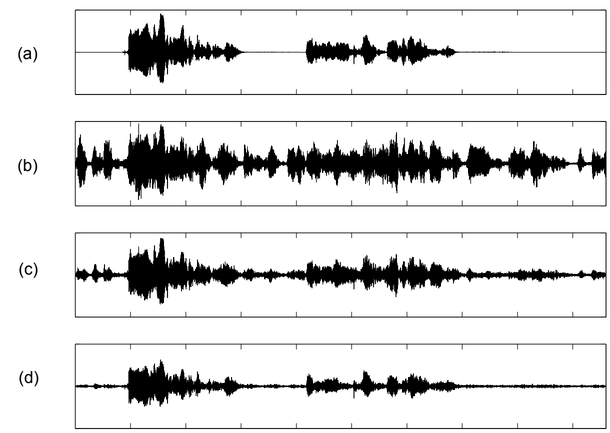

Fig. 3. Signal waveforms. (a) Clean signal. (b) Corrupted by competing speech noise with 0dB input SNR. (c) Enhanced signal by conventional multi-channel Wiener filter. (d) Enhanced signal by proposed method 3.

4.2. Performance Evaluation

The principal spatial subspace filter gains are plotted as a function of input SNR in Fig. 1 for a competing speech noise at 1000Hz. The gains of method 3 are the lowest due to the noise overestimation. The performance is evaluated by the measure of spectral distortion in Fig. 2. The proposed spatial subspace based methods are compared with the conventional frequency domain multi-channel Wiener filter (MWF). The best performance is obtained by the noise eigenvalue modification method (method 3) which removes the speech power in all the noise subspace for the competing speech noise environments. The signal waveforms are described in Fig. 3 for the case of 0dB input SNR.

5. CONCLUSIONS

We have proposed frequency domain multi-channel filtering schemes, which is based on the decomposition of spatial subspace matrices, with the noise eigenvalue modification. In the previous multi-channel Wiener filtering techniques, the noise statistics are estimated during noise-only period and used in subsequent speech-present period, which introduces considerable estimation error and performance degradation especially when the noise is highly nonstationary. We have developed several methods that modify noise eigenvalue matrix in the speech-present period using the rank-1 property of the speech spatial spectral matrix which attempts to correct the noise spatial spectral matrix. The first method is to find least square estimate and the other two methods are based on the time invariant spatial coherence property. Proposed methods are evaluated with the objective measure of spectral distortion, and the simulation result demonstrates that the best performance is obtained by the noise eigenvalue modification method (method 3) which removes the speech power in all the noise subspace.

ACKNOWLEDGEMENT

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea. This work was also partly supported by IT R&D Project funded by Korean Ministry of Information and Communications.

6. REFERENCES

[1] D. Florencio and H. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 197-200, Salt Lake City UT, USA, May 2001.

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

[3] G. Rombouts and M. Moonen, "QRD-based unconstrained optimal filtering for acoustic noise reduction," *Signal Processing*, vol. 83, no. 9, pp. 1889-1904, Sept. 2003.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 3rd edition, 1996.

[5] S. Doclo and M. Moonen, "Combined frequency-domain dereverberation and noise reduction technique for multi-microphone speech enhancement," *International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, pp. 31-34, September 2001.

[6] "RWCP Sound Scene Database in Real Acoustical Environments," Real World Computing Partnership, (c)1998-2001.