Interference Suppression Using Principal Subspace Modification in Multichannel Wiener Filter and Its Application to Speech Recognition

Gibak Kim

It has been shown that the principal subspace-based multichannel Wiener filter (MWF) provides better performance than the conventional MWF for suppressing interference in the case of a single target source. It can efficiently estimate the target speech component in the principal subspace which estimates the acoustic transfer function up to a scaling factor. However, as the input signal-to-interference ratio (SIR) becomes lower, larger errors are incurred in the estimation of the acoustic transfer function by the principal subspace method, degrading the performance in interference suppression. In order to alleviate this problem, a principal subspace modification method was proposed in previous work. The principal subspace modification reduces the estimation error of the acoustic transfer function vector at low SIRs. In this work, a frequency-band dependent interpolation technique is further employed for the principal subspace modification. The speech recognition test is also conducted using the Sphinx-4 system and demonstrates the practical usefulness of the proposed method as a front processing for the speech recognizer in a distant-talking and interferer-present environment.

Keywords: Multichannel Wiener filter, interference suppression, subspace, microphone array, speech recognition.

I. Introduction

When the target signal and interfering noise arrive at multiple microphones from different directions, the interferer can be suppressed by beamforming or multichannel filtering techniques based on the spatial diversity. Fixed beamformers have data-independent filter coefficients and usually provide distortionless response to the direction of the target signal while suppressing the power of interference coming from other directions. The residual noise in the output of the fixed beamformer can be further suppressed by an adaptive beamformer, for example, generalized sidelobe canceller, or single-channel postfiltering followed by the beamformer [1].

Multichannel Wiener filters (MWFs) have been shown to provide better performance than the beamforming techniques since they are less sensitive to the direction of arrival (DOA) estimation error and deviations from the assumed microphone characteristics, for example, gain, phase, and position [2]-[4]. For more efficient interference suppression with the MWF, a subspace-based approach has also been developed which removes noise subspace and estimates the target speech component from the remaining signal subspace [3], [5], [6]. If the subspace decomposition is performed in the frequency domain, the principal subspace vector estimates the acoustic transfer function vector up to a scaling factor, and better performance can be obtained by the principal subspace-based MWF [5].

However, as the input signal-to-interference ratio (SIR) becomes lower, the principal subspace vector deviates from the acoustic transfer function vector, which decreases the

Manuscript received Jan. 20, 2010; revised Sept. 6, 2010; accepted Sept. 29, 2010.

Gibak Kim (phone: +82 53 850 6651, email: imkgb27@gmail.com) was with the Department of Electrical Engineering, University of Texas at Dallas, USA, and is now with the School of Electromic Engineering, Daegu University, Daegu, Rep. of Korea.

doi:10.4218/etrij.10.0110.0045

performance of interference suppression. In Fig. 2, the deviation of the principal subspace vector from the acoustic transfer function will be illustrated in terms of the angle between two vectors (the principal subspace vector and the acoustic transfer function vector). In previous work, a principal subspace vector modification was proposed using the steering vector of the target speech signal for better performance at low SIRs [7]. The principal subspace vector was replaced by the linear interpolation of the original subspace vector and the steering vector of the target speech signal. The modified principal subspace estimates the acoustic transfer function more accurately and yields better performance in terms of SIR gain and mel-frequency cepstral coefficient (MFCC) distortion. In this paper, further improvements were provided by employing a frequency-band dependent interpolation for the principal subspace modification. The automatic speech recognition test was also conducted to support the effectiveness and the potential of the proposed method as a front processing for a distant-talking speech recognition system in the presence of a strong interferer.

The rest of this paper is organized as follows. Section II reviews the principal subspace-based MWF. Section III proposes the principal subspace modification method employed by a frequency-band dependent/independent interpolation. Section IV presents the simulation results with three different interferer scenarios evaluated in terms of SIR gain, MFCC distortion, and word error rate (WER) in automatic speech recognition.

II. Principal Subspace-Based MWF

If a single target signal $S(f)$ arrives at M microphones with M-dimensional acoustic transfer function $\mathbf{H}(f)$ from the source to the microphones and is corrupted by additive interfering noise, the multichannel signal model in the frequency domain is given by

$$\mathbf{Y}(f) = S(f)\begin{bmatrix} H_1(f) \\ H_2(f) \\ \vdots \\ H_M(f) \end{bmatrix} + \begin{bmatrix} N_1(f) \\ N_2(f) \\ \vdots \\ N_M(f) \end{bmatrix} = \begin{bmatrix} X_1(f) \\ X_2(f) \\ \vdots \\ X_M(f) \end{bmatrix} + \begin{bmatrix} N_1(f) \\ N_2(f) \\ \vdots \\ N_M(f) \end{bmatrix}$$
$$= S(f)\mathbf{H}(f) + \mathbf{N}(f) = \mathbf{X}(f) + \mathbf{N}(f), \tag{1}$$

where $\mathbf{Y}(f)$, $\mathbf{X}(f)$, and $\mathbf{N}(f)$ are the M-dimensional signals which denote the observed signal, target component, and additive noise (interferer) component, respectively. The filtered output $Z(f)$ can be written as

$$Z(f) = \mathbf{W}^H(f)\mathbf{Y}(f), \tag{2}$$

with a multichannel interference suppression filter $\mathbf{W}(f)$. Hereafter the frequency index (f) is omitted for the sake of brevity. If we assume that the target speech and interference are uncorrelated and estimate the target speech component in the first microphone signal in the minimum mean square error (MMSE) sense, the frequency domain MWF is given by [3], [5] as

$$\mathbf{W} \simeq \mathbf{R}_\mathrm{Y}^{-1}(\mathbf{R}_\mathrm{Y} - \mathbf{R}_\mathrm{N})\mathbf{e}_1, \tag{3}$$

where $\mathbf{R}_\mathrm{Y}=E\{\mathbf{YY}^H\}$, $\mathbf{R}_\mathrm{N}= E\{\mathbf{NN}^H\}$, and $\mathbf{e}_1=[1\ 0\ \ldots\ 0]^T$. In the conventional MWF, the interfering noise correlation matrix $\mathbf{R}_\mathrm{N}$ is recursively estimated with a forgetting factor during interference-only periods and kept fixed during target-present periods with the help of a target signal detector while the noisy speech correlation matrix $\mathbf{R}_\mathrm{Y}$ is updated during all the periods as

$$\mathbf{R}_\mathrm{N}(t) = \begin{cases} \alpha_N \mathbf{R}_\mathrm{N}(t-1) + (1-\alpha_N)\mathbf{Y}(t)\mathbf{Y}^H(t), & \text{noise-only}, \\ \mathbf{R}_\mathrm{N}(t-1), & \text{target-present}, \end{cases} \tag{4}$$

$$\mathbf{R}_\mathrm{Y}(t) = \alpha_Y \mathbf{R}_\mathrm{Y}(t-1) + (1-\alpha_Y)\mathbf{Y}(t)\mathbf{Y}^H(t), \tag{5}$$

where α_N and α_Y are smoothing constants.

By incorporating the subspace decomposition in the frequency domain, the spatial subspaces can be taken into consideration [5], [6]. The subspace decomposition can be performed by the joint diagonalization [8] of $\mathbf{R}_\mathrm{Y}$ and $\mathbf{R}_\mathrm{N}$ as

$$\begin{aligned} \mathbf{Q}^H \mathbf{R}_\mathrm{Y} \mathbf{Q} &= \mathbf{\Lambda}_\mathrm{Y}, \\ \mathbf{Q}^H \mathbf{R}_\mathrm{N} \mathbf{Q} &= \mathbf{\Lambda}_\mathrm{N}, \end{aligned} \tag{6}$$

where $\mathbf{\Lambda}_\mathrm{Y}$ and $\mathbf{\Lambda}_\mathrm{N}$ are diagonal matrices as

$$\mathbf{\Lambda}_\mathrm{Y} = \mathrm{diag}\{\lambda_{Y,1}\ \lambda_{Y,2}\ \cdots\ \lambda_{Y,M}\}, \tag{7}$$

$$\mathbf{\Lambda}_\mathrm{N} = \mathrm{diag}\{\lambda_{N,1}\ \lambda_{N,2}\ \cdots\ \lambda_{N,M}\}, \tag{8}$$

and $\mathbf{Q}$ is an invertible, but not necessarily orthogonal matrix. The correlation matrices can be expressed by the subspace matrix $\bar{\mathbf{Q}}$ as

$$\begin{aligned} \mathbf{R}_\mathrm{Y} &= \bar{\mathbf{Q}}\mathbf{\Lambda}_\mathrm{Y}\bar{\mathbf{Q}}^H, \\ \mathbf{R}_\mathrm{N} &= \bar{\mathbf{Q}}\mathbf{\Lambda}_\mathrm{N}\bar{\mathbf{Q}}^H, \end{aligned} \tag{9}$$

where $\bar{\mathbf{Q}} = \mathbf{Q}^{-H}$. By substituting (9) into (3), the frequency domain MWF is obtained as

$$\mathbf{W} = \mathbf{Q}(\mathbf{I} - \mathbf{\Lambda}_\mathrm{Y}^{-1}\mathbf{\Lambda}_\mathrm{N})\bar{\mathbf{Q}}^H\mathbf{e}_1. \tag{10}$$

When each of the frequency domain multichannel target speech components is the multiplication of the corresponding acoustic transfer function and the single target speech source as shown in (1), the correlation matrix of the target speech component can be written as

$$\mathbf{R}_\mathrm{X} = E\{\mathbf{XX}^H\} = E\{SS^*\}\mathbf{HH}^H, \tag{11}$$

where the rank of $\mathbf{R_X}$ is equal to 1. From (9) and the rank-1 property of $\mathbf{R_X}$, the estimate of the target speech correlation matrix is given by

$$\begin{aligned}\mathbf{R_X} &\simeq \bar{\mathbf{Q}}(\mathbf{\Lambda_Y} - \mathbf{\Lambda_N})\bar{\mathbf{Q}}^H \\ &\simeq (\lambda_{Y,1} - \lambda_{N,1})\bar{\mathbf{q}}_1\bar{\mathbf{q}}_1^H,\end{aligned} \tag{12}$$

where the M-dimensional principal subspace vector $\bar{\mathbf{q}}_1$ is the first column vector of $\bar{\mathbf{Q}}$. From (11) and (12), we note that $\bar{\mathbf{q}}_1$ is the estimate of the acoustic transfer function vector $\mathbf{H}$ up to a scaling factor [5]. In summary, if we replace $\bar{\mathbf{Q}}$ with $\begin{bmatrix}\bar{\mathbf{q}}_1 & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix}$ in (9) and (10), the principal subspace-based MWF can be expressed as

$$\mathbf{W} = \lambda_{N,1}\mathbf{R}_\mathbf{N}^{-1}\bar{\mathbf{q}}_1\left(\frac{\lambda_{Y,1} - \lambda_{N,1}}{\lambda_{Y,1}}\right)\bar{\mathbf{q}}_1^H\mathbf{e}_1, \tag{13}$$

$$\begin{aligned}\lambda_{Y,1} &= (\bar{\mathbf{q}}_1^H\mathbf{R}_\mathbf{Y}^{-1}\bar{\mathbf{q}}_1)^{-1}, \\ \lambda_{N,1} &= (\bar{\mathbf{q}}_1^H\mathbf{R}_\mathbf{N}^{-1}\bar{\mathbf{q}}_1)^{-1}.\end{aligned} \tag{14}$$

III. Principal Subspace Modification

The target speech and interference are assumed to be uncorrelated for obtaining the estimate of $\mathbf{R_X}$ in (12). In practical situations, this assumption is valid at rather high SIRs where the absolute value of the cross correlation $|E\{X_iN_j^*\}|$ is much smaller than $|E\{X_iX_j^*\}|$. However, in the case of low SIRs, the cross correlation cannot be ignored any more, and a large error occurs in the estimate of $\mathbf{R_X}$. Consequently, $\bar{\mathbf{q}}_1$ deviates from $\mathbf{H}$ (see Fig. 2), and the performance of the principal subspace-based MWF is degraded.

To obtain better performance with the principal subspace-based MWF at low SIRs, a principal subspace modification method was proposed using the information on the direction of the target signal (specifically the steering vector) [7]. We extend previous work to include a subspace modification with frequency-band dependent coefficients and evaluations of noise-corrupted speech by an automatic speech recognizer to show the usefulness of the proposed algorithm as a front processing for the automatic speech recognizer in a distant-talking with interfering noise. We further test the robustness of the proposed algorithm in the presence of DOA error.

Though the best way to correct the principal subspace is to replace it with the acoustic transfer function vector $\mathbf{H}$, the measurement of $\mathbf{H}$ should be done in noise-free condition, which is not often practical. Instead, in this paper, we use the steering vector of a target signal to modify the principal subspace. The steering vector can be easily obtained by assuming knowledge of the direction of target signal (typically in a hands-free communication system designed for personal use, for example, PDA, telematics, and smartphone) or estimating the direction by any wideband DOA estimation method [1]. In a far-field, the steering vector is equivalent to the multichannel acoustic transfer function in an ideal case when there is no microphone mismatch and the reverberation time is 0 ms. However, the error between the steering vector and the multichannel transfer function gets larger as the reverberation time increases even if there is no microphone mismatch. Therefore, we assume a moderately reverberant environment (reverberation time ≤ 300 ms). The steering vector of the target signal is denoted as

$$\mathbf{v}_s = [1 \;\; e^{j\phi_2} \;\; \cdots \;\; e^{j\phi_M}]^T, \tag{15}$$

where ϕ_i represents the phase of the i-th microphone signal with respect to the first microphone. The phases can be obtained from the direction (angle) of the target signal with respect to the microphones, signal frequency, and the configuration of microphone array [9].

Considering the steering vector as a reference for the acoustic transfer function vector, we modify the principal subspace toward the reference according to the amount of the deviation between the principal subspace vector $\bar{\mathbf{q}}_1$ and the steering vector $\mathbf{v}_s$. First, the angle between $\bar{\mathbf{q}}_1$ and $\mathbf{v}_s$ is calculated to measure the closeness of the two vectors. The angle between two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is a measure for closeness and can be defined as

$$\angle(\mathbf{v}_1, \mathbf{v}_2) = \cos^{-1}\left(\frac{\left|\mathbf{v}_1^H\mathbf{v}_2\right|}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right), \tag{16}$$

where ||·|| denotes the vector norm, and the range of the angle is [0, π/2]. Before calculating the angle between $\bar{\mathbf{q}}_1$ and $\mathbf{v}_s$, each element of $\bar{\mathbf{q}}_1$ is divided by its absolute value as

$$\bar{\bar{\mathbf{q}}}_1 \triangleq \left[\frac{\bar{q}_1}{|\bar{q}_1|} \;\; \frac{\bar{q}_2}{|\bar{q}_2|} \;\; \cdots \;\; \frac{\bar{q}_M}{|\bar{q}_M|}\right]^T, \tag{17}$$

where $\bar{\mathbf{q}}_1 = [\bar{q}_1 \; \bar{q}_2 \cdots \bar{q}_M]^T$. By calculating the angle between $\bar{\bar{\mathbf{q}}}_1$ and $\mathbf{v}_s$ instead of the angle between $\bar{\mathbf{q}}_1$ and $\mathbf{v}_s$, we alleviate the error caused by the microphone gain mismatch. We adopt a simple way to modify the principal subspace using linear interpolation between $\bar{\bar{\mathbf{q}}}_1$ and $\mathbf{v}_s$ as

$$\bar{\mathbf{q}}_1' = (1-\alpha)\frac{\bar{\bar{\mathbf{q}}}_1}{\|\bar{\bar{\mathbf{q}}}_1\|} + \alpha\frac{\mathbf{v}_s}{\|\mathbf{v}_s\|}, \tag{18}$$

$$\alpha = \frac{\angle(\mathbf{v}_s, \bar{\bar{\mathbf{q}}}_1)}{\pi/2}. \tag{19}$$

When the principal subspace is close to the steering vector, the principal subspace is barely modified since α is close to 0. On the contrary, the principal subspace with a large angle against the steering vector is modified toward the steering vector. After the interpolation, each element of $\bar{\bar{\mathbf{q}}}_1'$ is multiplied by each absolute value of the element of $\bar{\mathbf{q}}_1$ as

$$\bar{\mathbf{q}}_1' = \bar{\bar{\mathbf{q}}}_1' \bullet \left| \bar{\mathbf{q}}_1 \right|, \tag{20}$$

where $\bullet$ denotes the elementwise product. After the modification of the principal subspace vector, the MWF is calculated by replacing the original principal subspace vector $\bar{\mathbf{q}}_1$ with $\bar{\mathbf{q}}_1'$ in (13) and (14).

In (19), the two vectors, $\bar{\bar{\mathbf{q}}}_1$ and $\mathbf{v}_s$ are functions of frequency, and the angle between the two vectors is also a function of frequency, accordingly. For example, when considering two steering vectors for two different directions in a microphone array, the angles between the two steering vectors are proportional to the input frequency. Note that the angles between the two vectors in (19) tend to be larger at high frequencies compared to angles at low frequencies (see Fig. 2). We now propose a frequency-band dependent interpolation in (18) to consider the frequency dependent angle between the two vectors as

$$\alpha = \begin{cases} \left(\dfrac{\angle(\mathbf{v}_s, \bar{\bar{\mathbf{q}}}_1)}{\pi/2} \right)^{\frac{1}{2}}, & f < 1 \text{ kHz}, \\ \dfrac{\angle(\mathbf{v}_s, \bar{\bar{\mathbf{q}}}_1)}{\pi/2}, & 1 \text{ kHz} \le f < 4 \text{ kHz}, \\ \left(\dfrac{\angle(\mathbf{v}_s, \bar{\bar{\mathbf{q}}}_1)}{\pi/2} \right)^{2}, & f \ge 4 \text{ kHz}. \end{cases} \tag{21}$$

As proposed in (21), at a low frequency-band, the interpolation coefficient α is boosted by applying square-root function. On the contrary, α^2 is used as an interpolation coefficient to reduce the effect of the angle between $\bar{\bar{\mathbf{q}}}_1$ and $\mathbf{v}_s$ at high frequency-band ($f \ge 4$ kHz).

IV. Simulation Results

1. Simulation Data

In this simulation, we tested the algorithm in the presence of competing speech. For the competing speech interference, we prepared two news clips recorded at a 16 kHz sampling rate for 16 seconds. For the target signal, we used connected digits taken from the TIDIGITS database.

The TIDIGITS is a well-known database containing spoken digits from 0 through 9, with a single utterance including a sequence of digits [10]. The audio files in the TIDIGITS have a sampling rate of 20 kHz and were resampled to 16 kHz. From the competing speech recording, randomly cut segments were used for corrupting different target signals.

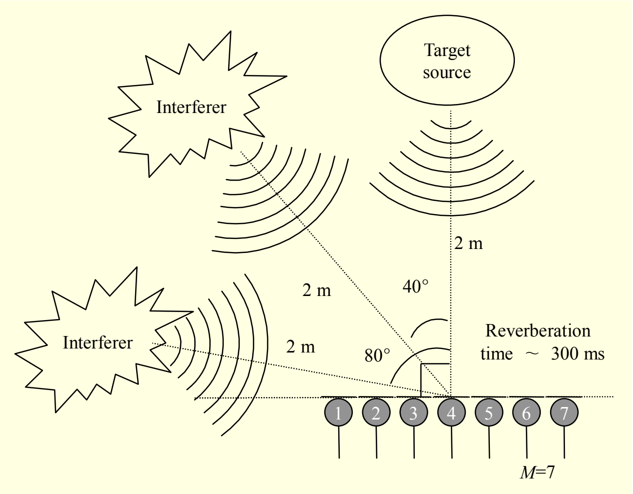


Fig. 1. Simulation environment: target source and interfering signals arriving at a linear microphone array.

The multichannel signals were created by the convolution of a sound source with acoustic impulse responses. The impulse responses were obtained from the RWCP Sound Scene Database, measured 2 m away from the center of the microphone array in real environments. The reverberation time was around 300 ms [11]. The microphone array is a linear type and has 7 microphones with 5.66 cm uniform intervals. Multichannel target signals were created by convolving the target signals (utterances of connected digits) with the impulse response measured in front of the microphone array. In the same way, the multichannel interfering noise signals were prepared coming at the angle of 40° and 80° to the direction of the target speech (see Fig. 1). Three different interference scenarios were tested: 1) "Angle 40": single competing speech interference coming at the angle of 40°; 2) "Angle 80": single competing speech interference coming at the angle of 80°; and 3) "Angle 40+Angle 80": two different competing speech interferences coming simultaneously at the angle of 40° and 80°, respectively. The multichannel interfering noise corrupted the target signal at a wide range of SIR (power ratio of signal-to-interfering noise in the time domain) levels ranging from –5 dB to 30 dB in steps of 5 dB. In this simulation, the direction of the target signal was assumed to be known.

The interference suppression procedure for the test data is as follows. The multichannel noisy signal was first segmented into 32 ms (512 samples for 16 kHz sampling) frames with 50% overlap between adjacent frames. Each frame was multiplied by Hann window and applied with 512 point fast Fourier transform (FFT). To reconstruct the time-domain signal, an inverse FFT was applied to the filtered frequency-domain signal, and the overlap-and-add technique was subsequently used.

2. Speech Recognition Test

The proposed algorithm was evaluated using the Sphinx-4 automatic speech recognizer [12]. For the recognition test, we used the TIDIGITS models included as part of the distribution of Sphinx-4. The acoustic model uses continuous density three-state HMMs with 8 Gaussian components per state. The cepstral analysis was done yielding 13 MFCCs including log energy as feature vectors. In addition, delta-MFCCs and delta-delta-MFCCs were used to obtain a 39-dimensional feature vector for each frame. Five hundred audio files were taken from 50 speakers (25 men and 25 women) for the test database, which were not included in the training of the model. For the selection of the test audio files from each speaker, we sorted the audio files of each speaker in the order of the file size and chose the largest 10 files for the test. The numbers of digits in the test database range from 5 to 7, but most of the utterances contained 7 digits. To assess the performance of the speech recognizer, a common metric, WER was computed as

$$\mathrm{WER}=\frac{S+D+I}{N_{\mathrm{ref}}}, \tag{22}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N_{ref} is the number of words in the reference.

The performance was also evaluated by SIR gain and MFCC distortion. The SIR gain is a common measure to evaluate an interference suppression algorithm. It is computed by the difference between the output SIR and input SIR as

$$\text{SIR gain}=10\log_{10}\frac{\sum_{t=1}^{T}|z_x(t)|^2}{\sum_{t=1}^{T}|z_n(t)|^2}-10\log_{10}\frac{\sum_{t=1}^{T}|x_1(t)|^2}{\sum_{t=1}^{T}|n_1(t)|^2}, \tag{23}$$

where T is the length of time domain signal, and $z_x(t)$ and $z_n(t)$ are the target and interfering noise component of output signals, that is, time domain signals of $Z_X(f)$ $(=\mathbf{W}^H(f)\mathbf{X}(f))$ and $Z_N(f)$ $(=\mathbf{W}^H(f)\mathbf{N}(f))$, respectively. The true target and interfering noise component observed at the first microphone (labeled "1" in Fig. 1) are $x_1(t)$ and $n_1(t)$, respectively. The MFCC distortion was evaluated to assess the spectral distortion between the (interference suppressed) output signal and the clean target signal (multichannel) and computed by

$$\text{MFCC distortion}=\frac{1}{K}\sum_{k=1}^{K}\sqrt{\sum_i[c_{x_1}^i(k)-c_z^i(k)]^2}, \tag{24}$$

where $c_{x_1}^i(k)$ and $c_z^i(k)$ are the i-th MFCC of the true target component at the first microphone and the output signal at k-th frame, respectively. The number of frames in a test audio file is represented by K. For the MFCC distortion measurement, the frequency region of 130 Hz to 6,800 Hz was filtered by 40 mel-scale filters, used in the feature extraction of the automatic speech recognizer (Sphinx-4).

3. Tested Algorithms

For the purpose of comparison, three other methods were also evaluated as well as the proposed algorithms.

- MWF_PS: The original principal subspace-based MWF (baseline algorithm, as per (13) and (14)).
- MWF_PS_MOD: The principal subspace-based MWF applied by the principal subspace modification with interpolation coefficient in (19).
- MWF_PS_MOD_2: The principal subspace-based MWF applied by the principal subspace modification with frequency-band dependent interpolation coefficients in (21).
- MWF_PS_SV: The principal subspace-based MWF where the principal subspace vector is replaced with the normalized steering vector of the target signal. The elements in the vector were further multiplied by the absolute values of the principal subspace vector. This method is equivalent to (18) with α=1.
- MVDR: The minimum variance distortionless response beamformer given [9] by

$$\mathbf{W}_{\mathrm{MVDR}}=\frac{\mathbf{R}_{\mathbf{N}}^{-1}\mathbf{v}_s}{\mathbf{v}_s^H\mathbf{R}_{\mathbf{N}}^{-1}\mathbf{v}_s}. \tag{25}$$

4. Estimation of Noise Correlation Matrix $\mathbf{R}_{\mathbf{N}}$

The noise correlation matrix $\mathbf{R}_{\mathbf{N}}$ is usually updated in the noise-only periods and kept unchanged during the target speech-present periods. For the purpose of the initialization of $\mathbf{R}_{\mathbf{N}}$, 320 ms silence was appended ahead of every test audio file. This was done because some test audio files do not have enough leading silence period for the initialization of $\mathbf{R}_{\mathbf{N}}$ (the test audio files have leading silence ranging from 48 ms to 800 ms before utterance starts). The 320 ms silence was used only for the $\mathbf{R}_{\mathbf{N}}$ initialization and not included for the evaluation of SIR gain, MFCC distortion, and speech recognizer. After the initialization, $\mathbf{R}_{\mathbf{N}}$ can be updated during noise-only periods using the target signal detector as per (4). However, the $\mathbf{R}_{\mathbf{N}}$ update is another topic of which performance is highly dependent on the target signal detector. In this work, the initialized $\mathbf{R}_{\mathbf{N}}$ was used as the estimate of $\mathbf{R}_{\mathbf{N}}$ and fixed throughout each audio file without further update.[1] The smoothing coefficient for the $\mathbf{R}_{\mathbf{Y}}$ update α_Y was set to 0.9.

[1] Performance with RN update was also evaluated by hand-labeling the target signal presence. However, there were very small differences between with and without noise updating because interfering signals were spatially invariant (not moving), and most improvement with MWF was achieved by the spatial filtering.

5. Results and Discussion

Figure 2 illustrates the angles between the acoustic transfer function vector **H** and the principal subspace vector as a function of input SIR before and after modifications in the presence of competing speech for five frequency bands. Angles were calculated for all frequency bins and for all input SIRs. Then, angles falling in each frequency band and in each SIR bin were averaged. The input SIR is not a global SIR but indicates the local SIR of a time-frequency unit in the short-time frequency analysis. In this figure, **H** was approximated by the principal eigenvector of the target speech correlation matrix $\mathbf{R}_X$ which was estimated using the oracle multichannel clean target signals. Before modifications, the angles between the two vectors increase at lower input SIR, which implies that the principal subspace vector $\overline{\mathbf{q}}_1$ deviates from **H**. The modified principal subspace vector $\overline{\mathbf{q}}_1'$ becomes closer to **H** at lower SIRs. At higher SIRs, little benefit is observed, and $\overline{\mathbf{q}}_1$ is even closer to **H** in the high frequencies (4,000 Hz to 8,000 Hz).

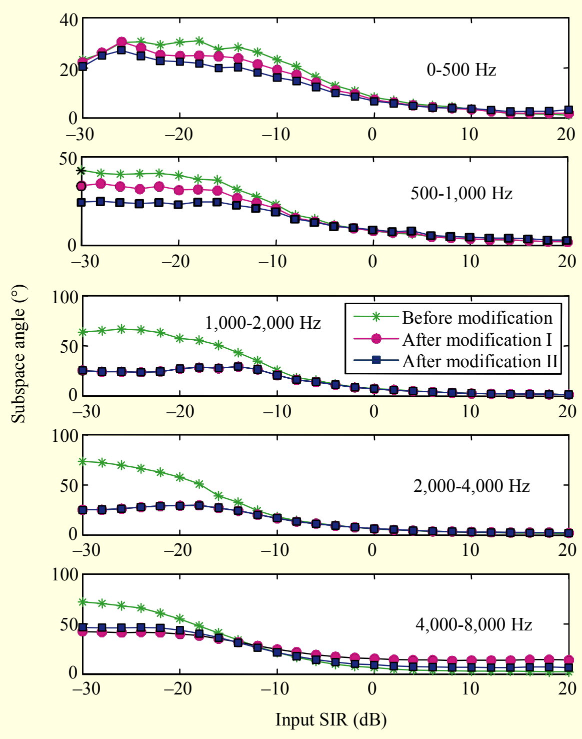


Fig. 2. Angles between the acoustic transfer function vector and the principal subspace vector for five frequency bands before and after the modification in the presence of competing speech. “Modification I” and “modification II” denote the principal subspace modification with frequency-band independent (as per (19)) and dependent (as per (21)) interpolation coefficients, respectively.

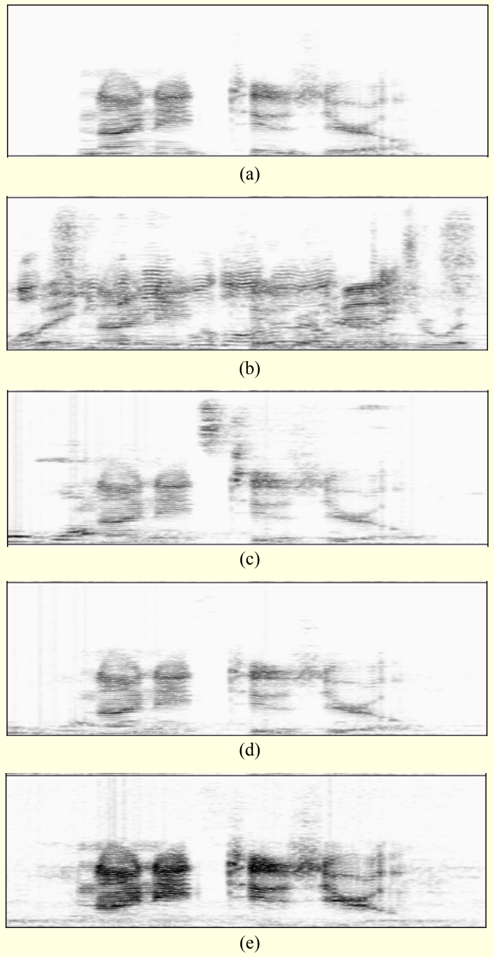


Fig. 3. Wideband spectrograms (a) of the clean target signal, (b) of a corrupted signal (two competing speech interferences coming at the angle of 40° and 80° to the direction of the target speech (“Angle 40+Angle 80”), SIR=0 dB), (c) of a signal processed by MWF_PS, (d) of a signal processed by MWF_PS_MOD_2, and (e) of a signal processed by MVDR. The example sentence is 2.9 seconds long, and the frequency ranges up to 8 kHz.

However, this disadvantage does not much affect the performance of the proposed method since most of the energy of speech signal resides below 4,000 Hz, at least for voiced segments, for example, vowels. Note that angles in Fig. 2 before modification tend to be larger at high frequency bands, which motivates the frequency-band dependent interpolation coefficient (as per (21)) for the principal subspace modification (indicated as “modification II”). At lower frequency bands (0 Hz to 500 Hz and 500 Hz to 1,000 Hz), “modification II” provides more angle reduction at low SIRs compared to “modification I,” which uses frequency-band independent interpolation coefficient (as per (19)). At the highest frequency band (4,000 Hz to 8,000 Hz), “modification II” reduces angles at high SIRs while slightly increasing angles at low SIRs.

Figure 3(d) shows an example spectrogram of an enhanced signal using the proposed algorithm, MWF_PS_MOD_2. For

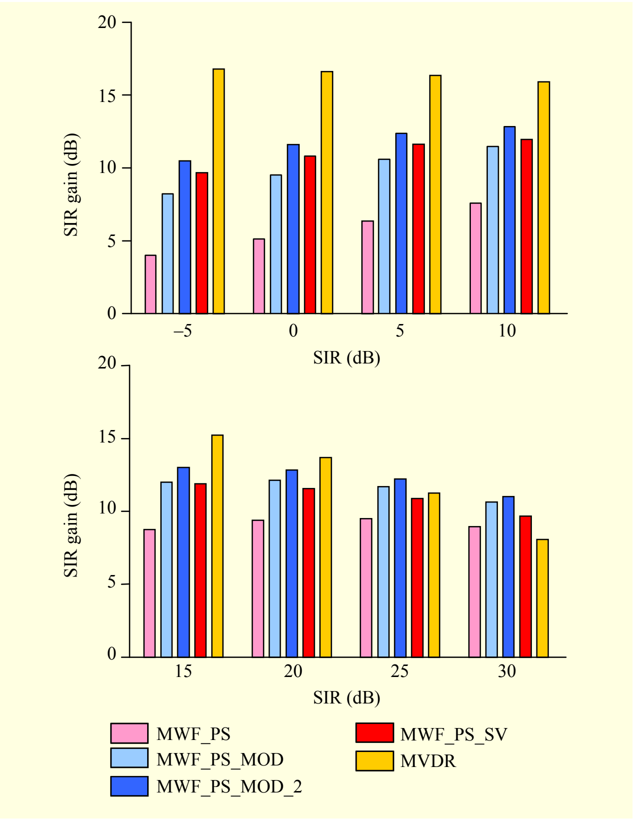


Fig. 4. SIR gains with interfering speech coming at 40°.

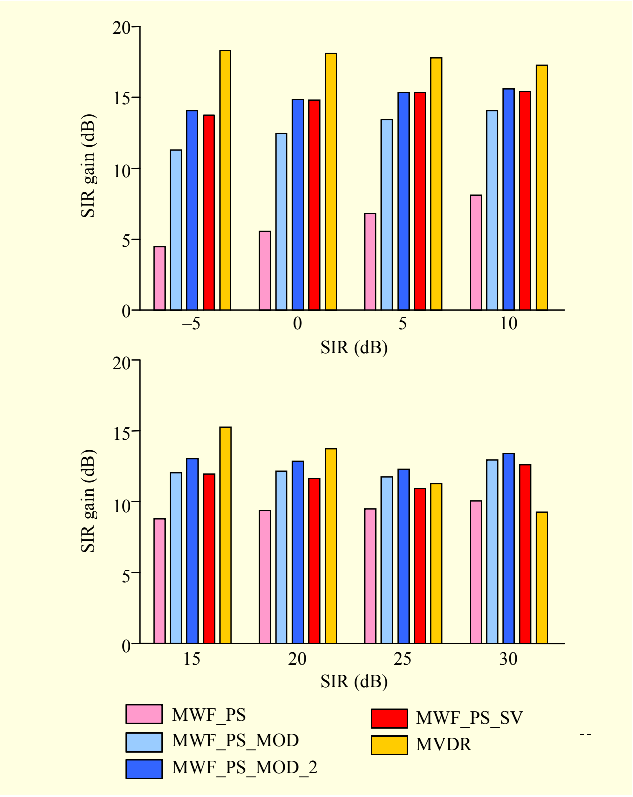


Fig. 5. SIR gains with interfering speech coming at 80°.

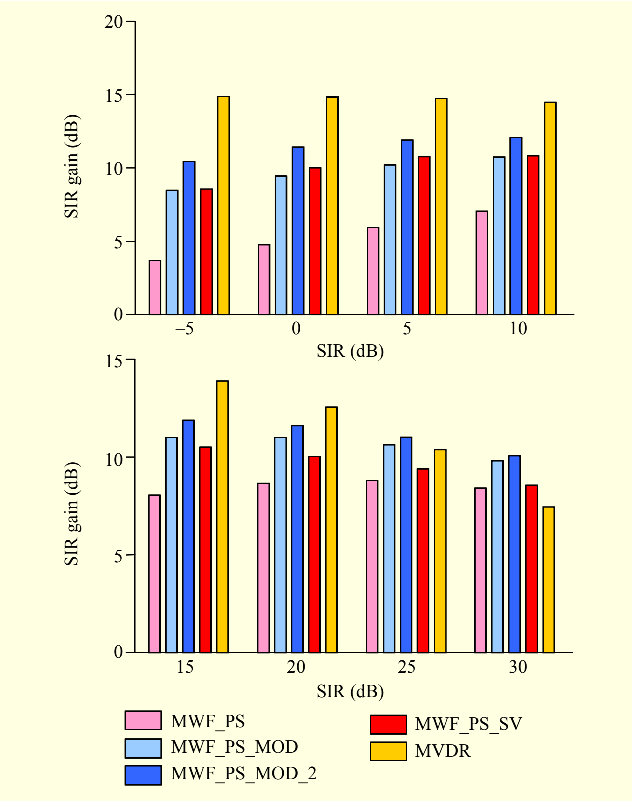


Fig. 6. SIR gains with interfering speech signals coming at 40° and 80°.

the purpose of comparison, Figs. 3(c) and 3(e) show signals processed by MWF_PS and MVDR, respectively. In the spectrogram in Fig. 3(e), MVDR shows higher SIR compared to the spectrogram in Fig. 3(d), but signal attenuation is noticeable at low frequencies due to microphone gain mismatch and results in worse performance in terms of MFCC distortion and WER (see Figs. 6, 9, and 12).

Figures 4 to 12 show the performance achieved by each method when the target speech signal is corrupted in each of three different interference scenarios specified in section IV.1. The recognition test was also done for the clean signal and multichannel clean signal (convolved with the multichannel impulse response as described in section IV.1). The WERs for the two conditions were 0.52% and 0.98%, respectively, which shows that the multichannel target component received at the microphones insignificantly degrades the performance of the speech recognizer. The proposed methods, MWF_PS_MOD and MWF_PS_MOD_2, provide better performances in terms of MFCC distortion and WER of speech recognition than other tested algorithms for all the three interference scenarios. While the MVDR algorithm shows large SIR gain, it fails to reduce MFCC distortion or WER in most conditions. The main reason for the large WER of the MVDR algorithm in spite of the high SIR gain is due to signal distortion (see Fig. 3(e))

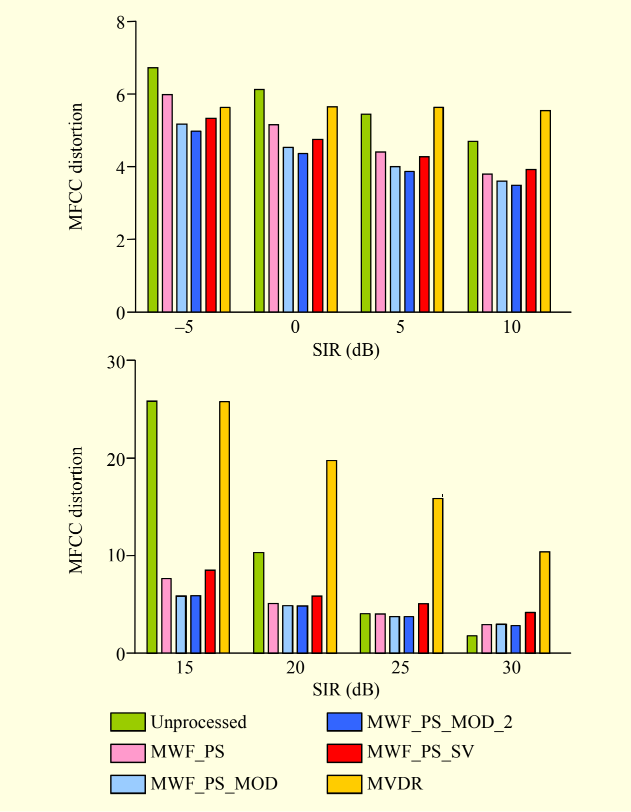


Fig. 7. MFCC distortions with interfering speech coming at 40°.

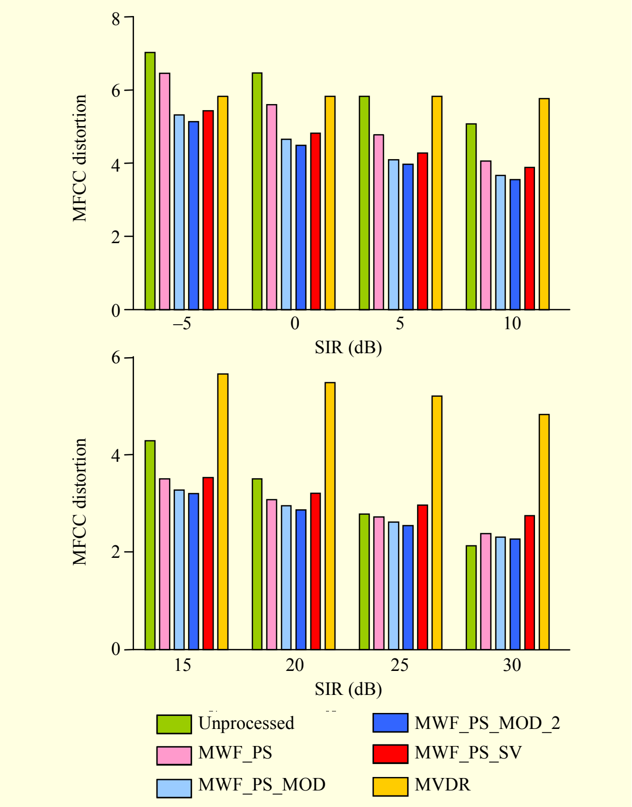


Fig. 9. MFCC distortions with interfering speech signals coming at 40° and 80°.

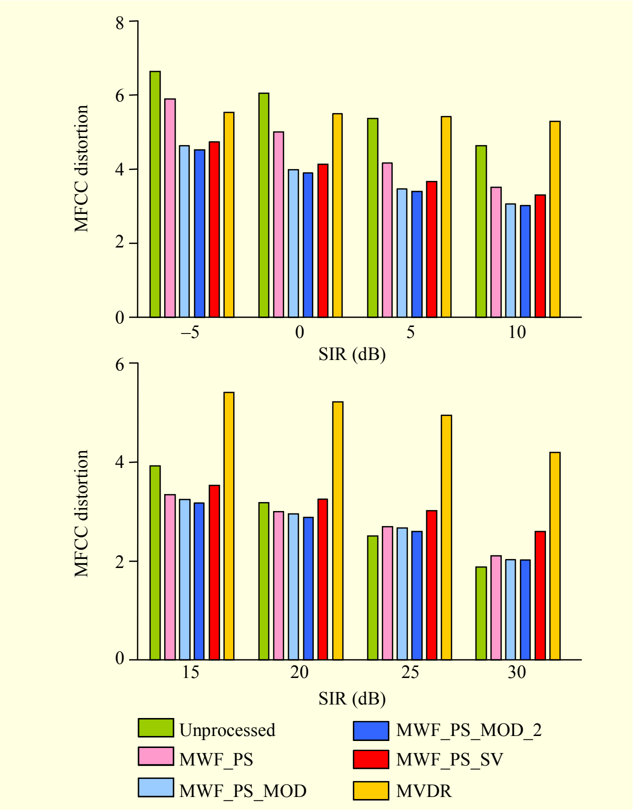


Fig. 8. MFCC distortions with interfering speech coming at 80°.

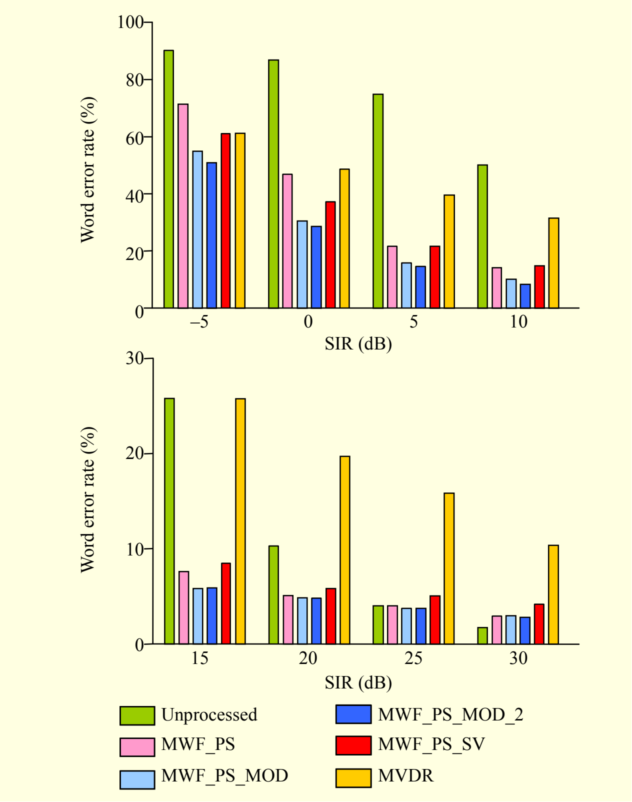


Fig. 10. WERs with interfering speech coming at 40°.

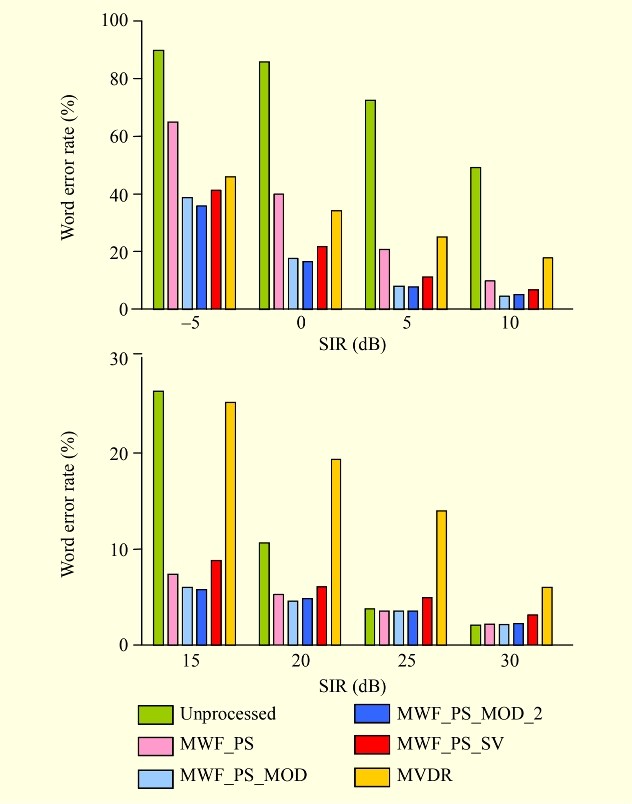


Fig. 11. WERs with interfering speech coming at 80°.

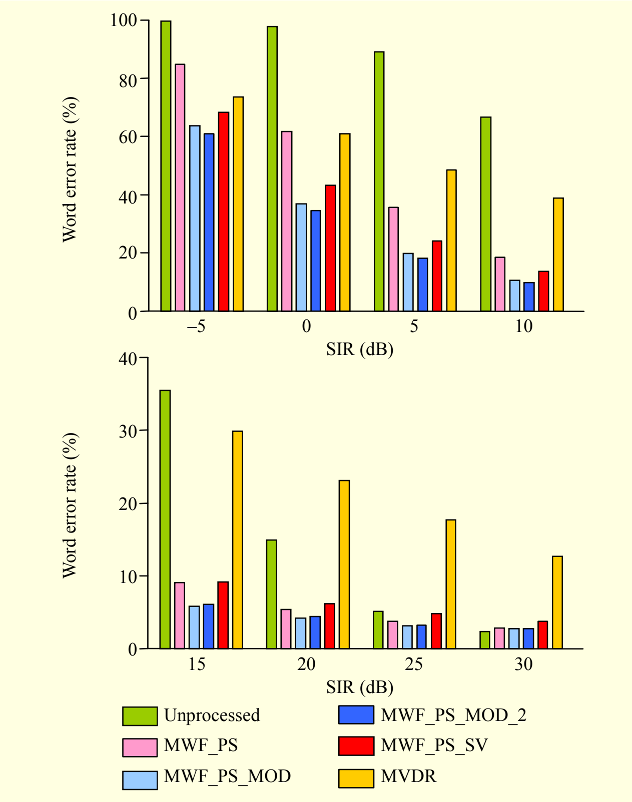


Fig. 12. WERs with interfering speech signals coming at 40° and 80°.

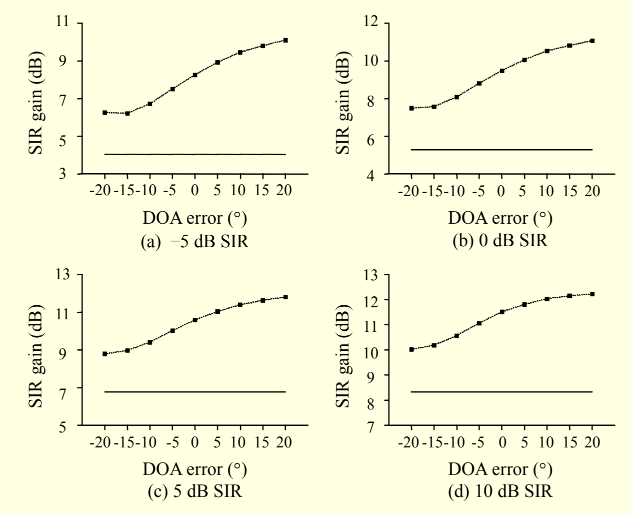
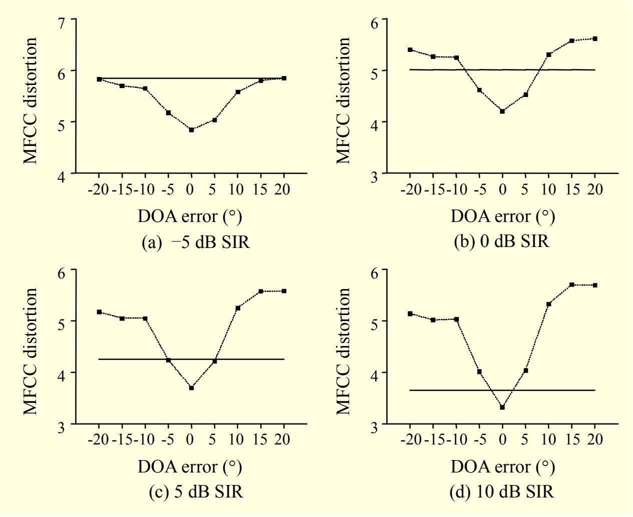


Fig. 13. SIR gains of the proposed method (MWF_PS_MOD) as a function of DOA error in the presence of competing speech at the four different SIR levels with the horizontal solid line showing the SIR gain of the MWF_PS.

caused by lack of account of the microphone gain mismatch, which was considered in the other methods (MWF_PS, MWF_PS_MOD, MWF_PS_MOD_2, and MWF_PS_SV). Larger improvements were observed for the interference scenario “Angle 80” compared to “Angle 40,” which implies more effective interference suppression for a wider angle between the directions of target signal and interference. Among the three interference scenarios, the worst performance was achieved for “Angle 40 + Angle 80” where there were two different interfering noise sources.

At 30 dB SIR, slight performance drops were observed in MFCC distortion and WER (Figs. 7 to 12), which were possibly caused by the fact that the proposed modification techniques slightly increased the angle (estimation error) between the acoustic transfer function vector and the principal subspace vector at high SIR (especially at high frequencies) as shown in Fig. 2. We believe that the estimation error of the interfering noise correlation matrix $\mathbf{R}_N$ (see (4)) is the main cause of the increased angle at high SIR after modification. We also observe that the performance of the speech recognizer is much more related to the spectral distortion (MFCC distortion) than the SIR gain. Since the proposed principal subspace modification technique relies on information about the direction of the target signal, it is reasonable to ask how much the algorithm is affected by the estimation error of DOA. Figures 13 to 15 show the performance as a function of DOA error with 5 degree steps for interference coming at 40°. When the target signal comes from the front of the linear microphone array (see Fig. 1), the negative DOA error indicates the case where the DOA is estimated as if the target signal comes from the left of the target source. As shown in Fig. 13, large SIR gain was obtained as the DOA



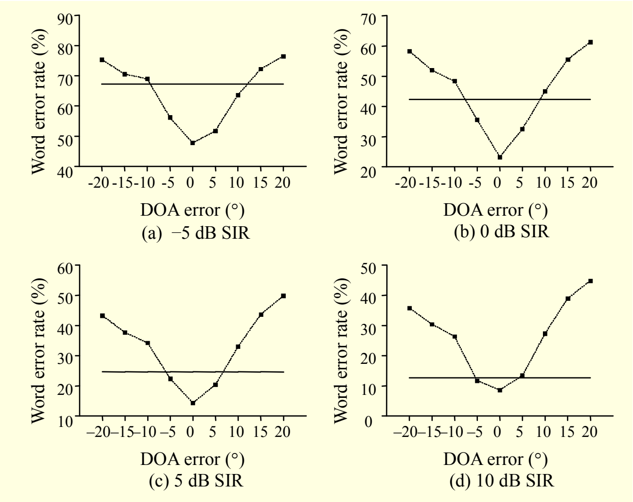
(a) −5 dB SIR

(b) 0 dB SIR

(c) 5 dB SIR

(d) 10 dB SIR

Fig. 14. MFCC distortions of the proposed method (MWF_PS_MOD) as a function of DOA error in the presence of competing speech at the four different SIR levels with the horizontal line showing the MFCC distortion of the MWF_PS.



(a) −5 dB SIR

(b) 0 dB SIR

(c) 5 dB SIR

(d) 10 dB SIR

Fig. 15. WERs of the proposed method (MWF_PS_MOD) as a function of DOA error in the presence of competing speech at the four different SIR levels with the horizontal line showing the WER of the MWF_PS.

error increases in the positive angle. This is because the interfering noise comes from the left of the target source, and larger SIR gain was obtained by increasing the angle between the interfering noise and the subspace vector.

However, it turns out that this large SIR gain was obtained at the cost of spectral distortion, which was shown in the evaluation of the MFCC distortion and WER. The performance degrades as DOA error increases in both cases (negative or positive error) as shown in Figs. 14 and 15.

In the figures, the horizontal lines indicate the performance of the MWF_PS, and as noted, the proposed method tolerates some errors depending on the input SIR, for example, up to 5 degrees in the case of 5 dB.

We also obtained similar results with the MWF_PS_MOD_2 algorithm, but only showed results with the MWF_PS_MOD algorithm in Figs. 13 to 15 for brevity.

V. Conclusion

In this paper, modification techniques for the principal subspace-based MWF have been proposed. The principal subspace vector was modified by the interpolation between the principal subspace vector and the steering vector of the target speech signal. It reduces the estimation error of the acoustic transfer function vector at low SIRs, where the conventional method MWF_PS usually performs poorly.

The speech recognition test was conducted, and the results support the efficiency of the proposed method as a front processing of a distant-talking speech recognition system, especially in the presence of a strong interferer. It was also demonstrated that a frequency-band dependent interpolation provides further improvements compared to the frequency-band dependent linear interpolation for the principal subspace modification.

References

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, Springer-Verlag, 2001.

[2] D. Florencio and H. Malvar, "Multichannel Filtering for Optimum Noise Reduction in Microphone Arrays," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Salt Lake City, UT, USA, May 2001, pp. 197-200.

[3] S. Doclo and M. Moonen, "GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, Sept. 2002, pp. 2230-2244.

[4] A. Spriet, M. Moonen, and J. Wouters, "Robustness Analysis of Multichannel Wiener Filtering and Generalized Sidelobe Cancellation for Multimicrophone Noise Reduction in Hearing Aid Applications," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, July 2005, pp. 487-503.

[5] S. Doclo and M. Moonen, "Combined Frequency-Domain Dereverberation and Noise Reduction Technique for Multi-microphone Speech Enhancement," *Int. Workshop Acoustic Echo Noise Control*, Darmstadt, Germany, Sept. 2001, pp. 31-34.

[6] W. Herbordt, *Sound Capture for Human/Machine Interfaces*, Springer-Verlag, 2005.

[7] G. Kim and N.I. Cho, "Principal Subspace Modification for Multi-channel Wiener Filter in Multi-microphone Noise Reduction," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, 2008, pp. 4909-4912.

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

[9] H. Van Trees, *Detection*, *Estimation and Modulation Theory*, Part IV: Optimum Array Processing, New York: Wiley, 2002.

[10] R.G. Leonard, "A Database for Speaker-Independent Digit Recognition," *Proc. IEEE Int. Conf. Acoustics*, *Speech, Signal Process.*, 1984, pp. 111-114.

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

[12] The CMU Sphinx Group Open Source Speech Recognition Engines. Available: http://cmusphinx.sourceforge.net

Gibak Kim received the BS and MS in electronics engineering and the PhD in electrical engineering from Seoul National University, Seoul, Korea, in 1994, 1996, and 2007, respectively. From 1996 to 2000, he was with the Machine Intelligence Group, Department of Information Technology at LG Electronics Inc., Seoul, Korea. He also worked at Voiceware Ltd. from 2000 to 2003, as a senior research engineer involved in the development of an automatic speech recognizer. From 2007 to 2010, he was a research associate at the University of Texas at Dallas, Richardson, working on the development of noise-reduction algorithms that can improve speech intelligibility. He is currently with School of Electronic Engineering, College of Information and Communication, Daegu University, Daegu, Korea. His general research interests are in speech enhancement, speech recognition, and microphone array signal processing.