MULTI-MICROPHONE INTERFERENCE SUPPRESSION USING THE PRINCIPAL SUBSPACE MODIFICATION AND ITS APPLICATION TO SPEECH RECOGNITION

Gibak Kim

School of Electronic Engineering, College of Information & Communication, Daegu University, Korea
email: imkgb27@gmail.com

ABSTRACT

It has been shown that the principal subspace based multi-channel Wiener filter (MWF) provides better performance than the conventional MWF for the interference suppression in the case of a single target source. It can efficiently estimate the target component in the principal subspace and the acoustic transfer function up to a scaling factor. However, as input signal-to-interference ratio (SIR) becomes lower, larger errors are incurred in the estimation of the acoustic transfer function by the principal subspace degrading the performance in interference suppression. In order to alleviate this problem, a principal subspace modification method was proposed in previous work, which uses a priori information on the direction of the target signal. In this paper, a frequency-band dependent interpolation technique is further employed for the principal subspace modification and the speech recognition test is conducted to demonstrate the practical usefulness of the proposed methods as a front processing for the speech recognizer in interferer-present environment.

Index Terms—Multi-channel Wiener filter, interference suppression, microphone array, speech recognition

1. INTRODUCTION

When the target signal and interferer arrive at the multiple microphones from different directions, the interferer can be suppressed by beamforming or multi-channel filtering techniques based on the spatial diversity. The MWF has been shown to provide better performance than the beamforming techniques since it is less sensitive to DOA (Direction Of Arrival) estimation error and deviations from the assumed microphone characteristics (e.g., gain, phase, position, etc.) [1–3]. For more efficient interference suppression with the MWF, a subspace based approach has also been developed, which removes noise subspace and estimates the target component from the remaining signal subspace [2,4,5]. 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 [4].

However, as the input SIR becomes lower, the principal subspace vector deviates from the acoustic transfer function vector, which decreases the performance of interference suppression. In previous work, a principal subspace vector modification was proposed using the steering vector of the target signal for better performance at low SIRs [6]. The principal subspace vector was replaced by the linear interpolation of the original subspace vector and the steering vector of the target signal. The modified principal subspace estimates the acoustic transfer function more accurately and yields better performance in terms of SIR gain, MFCC distortion. In this paper, further improvements are provided by employing a frequency-band dependent interpolation for the principal subspace modification. The automatic speech recognition tests are conducted to show the potential of the proposed methods as a front processing for a distant-talking speech recognizer in the presence of interferer.

2. PRINCIPAL SUBSPACE-BASED MWF

2.1. Multi-channel Wiener filter

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

\[
Y(f) = S(f)H(f) + N(f) = X(f) + N(f)
\]

where \( Y(f), X(f), N(f) \) 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

\[
Z(f) = W^H(f)Y(f)
\]

with a multi-channel interference suppression filter \( W^H(f) \). Hereafter the frequency index \( f \) is omitted for the sake of brevity. If we assume that the target and interfering noise signals are uncorrelated and estimate the target component in the first microphone signal in the minimum mean square error (MMSE) sense, the frequency domain MWF is given by [2,4]

\[
W \simeq R_Y^{-1} (R_Y - R_N) e_1
\]

with \( R_Y = E\{YY^H\}, R_N = E\{NN^H\}, \) and \( e_1 = [1 \ 0 \ \cdots \ 0]^T \). In the conventional MWF, the interfering noise correlation matrix, \( R_N \), is recursively estimated with a forgetting factor during noise-only periods and kept fixed during target-present periods with the help of a target signal detector while the noisy signal correlation matrix, \( R_Y \), is updated during all the periods with another forgetting factor. By incorporating the subspace decomposition in the frequency domain, the spatial subspaces can be taken into consideration [4,5]. The subspace decomposition can be performed by the joint diagonalization of \( R_Y \) and \( R_N \) as

\[
\begin{align*}
Q_Y R_Y Q_Y^H &= \Lambda_Y \\
Q_N R_N Q_N^H &= \Lambda_N
\end{align*}
\]

where \( \Lambda_Y, \Lambda_N \) are diagonal matrices as

\[
\Lambda_Y = \text{diag} \{ \lambda_{Y,1} \ \lambda_{Y,2} \ \cdots \ \lambda_{Y,M} \} \\
\Lambda_N = \text{diag} \{ \lambda_{N,1} \ \lambda_{N,2} \ \cdots \ \lambda_{N,M} \}
\]
and $Q$ is an invertible, but not necessarily orthogonal matrix. Then the correlation matrices can be expressed by the subspace matrix $Q$ as

$$
\begin{align}
R_Y &= Q \Lambda_Y Q^H \\
R_N &= Q \Lambda_N Q^H
\end{align}
(7)
$$
with $Q = Q^{-H}$. By substituting (7) into (3) the frequency domain MWF is obtained as

$$
W = Q \left( I - \Lambda_Y^3 \Lambda_N \right) Q^H e_1.
(8)
$$

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

$$
R_X = E \{ X X^H \} = E \{ S S^* \} H H^H
(9)
$$
where the rank of $R_X$ is equal to 1. From (7) and the rank-1 property of $R_X$, the estimate of the target correlation matrix is given by

$$
\hat{R}_X \simeq Q (\Lambda_Y - \Lambda_N) Q^H
(10)
$$
where the $M$-dimensional principal subspace vector $\hat{q}_1$ is the first column vector of $Q$. From (9) and (10), note that $\hat{q}_1$ is the estimate of the acoustic transfer function vector $H$ up to a scaling factor [4]. In summary, the principal subspace based MWF can be expressed as

$$
W = \lambda_{N,1}^{-1} R_N^{-1} \hat{q}_1 \left( \frac{\lambda_{Y,1} - \lambda_{N,1}}{\lambda_{Y,1}} \right)^{-1} \hat{q}_1^H e_1
(11)
$$

and the reverberation time is 0 ms. However, the error between the steering vector and the multi-channel transfer function gets larger as the reverberation time increases even if there is no microphone mismatch. In this work, therefore, we assume a moderately reverberant environment (reverberation time $\leq 300$ ms).

Considering the steering vector as a reference for the acoustic transfer function vector, we modify the principal subspace toward the reference according to the deviation between the principal subspace vector $\hat{q}_1$ and the steering vector of the target signal $v_s$. First, the angle between $\hat{q}_1$ and $v_s$ is calculated to measure the closeness of two vectors. The angle between two vectors $v_1$ and $v_2$ is a measure for closeness and can be defined as

$$
\angle(v_1, v_2) = \cos^{-1} \left( \frac{v_1^H v_2}{||v_1|| \cdot ||v_2||} \right)
(13)
$$
where $||\cdot||$ denotes the vector norm and the range of the angle is $[0, \pi/2]$. Before calculating the angle between $\hat{q}_1$ and $v_s$, each element of $\hat{q}_1$ is divided by its absolute value as

$$
\bar{q}_1 = \left[ \frac{q_1}{|q_1|} \frac{q_2}{|q_2|} \ldots \frac{q_M}{|q_M|} \right]^T
(14)
$$
with $\bar{q}_1 = [\bar{q}_1, \bar{q}_2, \ldots, \bar{q}_M]^T$. By calculating the angle between $\hat{q}_1$ and $v_s$ instead of the angle between $\hat{q}_1$ and $v_s$, we alleviate the error caused by the microphone gain mismatch. We adopt a simple way to modify the principal subspace using linear interpolation between $\bar{q}_1$ and $v_s$ as

$$
\bar{q}'_1 = (1 - \alpha) \bar{q}_1 + \alpha \frac{v_s}{||v_s||}
(15)
$$

$$
\alpha = \angle(v_s, \bar{q}_1) / \pi / 2.
(16)
$$

After the interpolation, each element of $\bar{q}'_1$ is multiplied by each absolute value of the element of $\bar{q}_1$ as

$$
\bar{q}_1 = \bar{q}'_1 \cdot ||\bar{q}_1||
(17)
$$
where $\cdot$ denotes the elementwise product. After the modification of the principal subspace vector, the MWF is calculated by replacing the original principal subspace vector $\hat{q}_1$ with $\bar{q}_1$ in (11) and (12).

In (16), the two vectors, $\bar{q}_1$ and $v_s$, 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 (16) tend to be larger at high frequencies compared to angles at low frequencies (see Fig. 1). Motivated by this, we propose a frequency-band dependent interpolation to consider the frequency dependent angle between the two vectors as

$$
\alpha' = \begin{cases}
\sqrt{\alpha}, & f < 1 \text{kHz} \\
\alpha, & 1 \text{kHz} \leq f < 4 \text{kHz} \\
\alpha^2, & f \geq 4 \text{kHz}
\end{cases}
(18)
$$

As proposed in (18), at low frequency-band, the interpolation coefficient is boosted by applying square-root function. On the contrary, $\alpha^2$ is used as an interpolation coefficient to reduce the effect of the angle between $\hat{q}_1$ and $v_s$ at high frequency-band ($f \geq 4$kHz).
4. SIMULATION RESULTS

4.1. 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 16 kHz sampling rate for 16 seconds. For the target signal, we used connected digits taken from TIDIGITS database [8]. 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.

The multi-channel 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 [9] which were measured 2 m away from the center of the microphone array in real environments. The reverberation time was around 300 ms. The microphone array is a linear type and has 7 microphones with 5.66 cm uniform intervals. Multi-channel 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, for the interference, two different multi-channel competing speech were prepared coming simultaneously at the angle of 40° and 80° to the direction of the target speech (see Fig. 2). The multi-channel interference corrupted the target signal at the SIR levels of -5, 0, 5, and 10 dB.

The multi-channel 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 [9] which were measured 2 m away from the center of the microphone array in real environments. The reverberation time was around 300 ms. The microphone array is a linear type and has 7 microphones with 5.66 cm uniform intervals. Multi-channel 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, for the interference, two different multi-channel competing speech were prepared coming simultaneously at the angle of 40° and 80° to the direction of the target speech (see Fig. 2). The multi-channel interference corrupted the target signal at the SIR levels of -5, 0, 5, and 10 dB.

Fig. 1. 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 (16) and dependent (18) interpolation coefficients, respectively.

Fig. 2. Simulation environment: target source and interfering signals arriving at a linear microphone array.

4.2. Speech recognition test

The proposed algorithms were evaluated using the Sphinx-4 automatic speech recognizer [10]. 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. 500 audio files were taken from 50 speakers (25 men, 25 women) for the test database, which were not included in the training of 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 contain 7 digits. To assess the performance of the speech recognizer, a common metric, word error rate (WER) was computed as

\[
W = \frac{S + D + I}{N_{\text{ref}}} \quad (19)
\]

where S, D, I, and N_{\text{ref}} are the numbers of substitutions, deletions, insertions, and words in the reference, respectively.

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

4.3. Tested algorithms

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

- **MWF_PS**: the original principal subspace based MWF [(11), (12)].
- **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 (15) with \( \alpha = 1 \).
- **MVDR**: the minimum variance distortionless response beamformer.
- **MWF_PS_MOD**: the proposed algorithm with interpolation coefficients in (16).
- **MWF_PS_MOD_2**: the proposed algorithm with interpolation coefficients in (18).

The interference suppression procedure for the test data is as follows. The multi-channel noisy signal was first segmented into
32 ms (512 samples for 16 kHz sampling) frames with 50 % overlap between adjacent frames. Each frame was Hann windowed and applied with 512 point FFT (Fast Fourier Transform). 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.

4.4. Results and discussions

Fig. 1 illustrates the angles between the acoustic transfer function vector $\mathbf{H}$ and the principal subspace vector as a function of input SIR before and after the modification in the presence of competing speech for five frequency bands. The input SIR is not a global SIR but indicates the local SIR of time-frequency unit in the short-time frequency analysis. In this figure, $\mathbf{H}$ was approximated by the principal eigenvector of the target speech correlation matrix $\mathbf{R}_x$ which was estimated using the oracle multi-channel target signals without interferer corrupting. Before modification, the angle between two vectors increases at lower input SIR, which implies that the principal subspace vector $\mathbf{q}_1$ deviates from $\mathbf{H}$. Note that the modified principal subspace vector $\mathbf{q}_1^*$ becomes closer to $\mathbf{H}$ at lower SIRs. At higher SIRs, little benefit is observed and $\mathbf{q}_1^*$ is even closer to $\mathbf{H}$ in the high frequencies (4000-8000 Hz). However, this disadvantage does not much affect the performance of the proposed methods since most energy of speech signal resides below 4000 Hz, at least for voiced segments (e.g., vowels). Note that angles in Fig. 1 before modification tend to be larger at high frequency bands, which motivates the frequency-band dependent interpolation coefficient (18) for the principal subspace modification (indicated as “modification II”). At lower frequency bands (0-500 Hz and 500-1000 Hz), the “modification II” provides more angle reduction at low SIRs compared to the “modification I” which uses frequency-band independent interpolation coefficient (16). At the highest frequency band (4000-8000 Hz), the “modification II” yields smaller angles at high SIRs while slightly larger angles at low SIRs compared to the “modification I”.

Figs. 3 and 4 show the performance achieved by each method. The recognition test was done for the clean signal and multi-channel clean signal (convolved with the multi-channel impulse response as described in Section 4.1). The WERs were 0.52 % and 0.98 %, respectively, which shows the multi-channel target component received at the microphones does not significantly degrade the performance of speech recognizer. The proposed methods (MWF_PS_MOD and MWF_PS_MOD_2) provide better performances in terms of WER of speech recognition than other tested algorithms. While the MVDR shows large SIR gain, it fails to reduce WER except -5 dB SIR. The main reason of the large WER of the MVDR in spite of the high SIR gain is due to signal distortion caused by lack of accounting for the microphone gain mismatch, which was considered in the other methods.

5. CONCLUSION

In this paper, we proposed a principal subspace modification for the principal subspace based MWF. The principal subspace vector was modified by the interpolation (frequency-band independent/dependent) between the principal subspace vector and the steering vector of the target 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 methods as a front processing for a speech recognition system in a distant-talking with interfering noise.

6. REFERENCES