ON A MULTICHANNEL MAXIMUM SNR FILTER FOR NOISE REDUCTION IN THE STFT DOMAIN

Gongping Huang[1]*, Jacob Benesty*[2]*, and Jingdong Chen*[1]

[1]CIAIC and School of Marine Science and Technology
Northwestern Polytechnical University
127 Youyi West Rd., Xi'an, Shaanxi 710072, China

[2]INRS-EMT, University of Quebec
800 de la Gauchetiere Ouest, Suite 6900
Montreal, QC H5A 1K6, Canada

ABSTRACT

This paper studies the multichannel maximum signal-to-noise ratio (SNR) filter for noise reduction in the short-time Fourier transform (STFT) domain. By considering both the interchannel and interframe information, the maximum SNR filter is formulated by maximizing the output SNR with the minimum distortion constraint. Simulations are performed to examine the performance of the deduced multichannel maximum SNR noise reduction filter in different conditions and the results demonstrated that this filter can significantly increase both the SNR and the speech quality.

Index Terms— Noise reduction, speech enhancement, multichannel, maximum SNR filter, short-time Fourier transform (STFT) domain.

1. INTRODUCTION

Since one of the major objectives of noise reduction is to reduce the amount of noise [1–5], thereby improving the signal-to-noise ratio (SNR), it is a natural motivation to investigate the filter that maximizes the output SNR. However, the traditional maximum SNR filters, which are derived by maximizing the Raleigh quotient between the variance of the filtered speech and that of the residual noise without any constraint, have been found to introduce significant speech distortion, making them useless in real applications [1]. Recently, we developed a maximum SNR approach to single-channel noise reduction in the short-time Fourier transform (STFT) domain [6]. This new maximum SNR filter differs from the traditional ones in two major aspects: 1) it considers the interframe information in every subband in the STFT domain and 2) the maximum SNR filter is derived with the minimum speech distortion constraint. Experiments showed that this filter can significantly improve both the SNR and speech quality at the same time. This paper continues our effort on the maximum SNR filter for noise reduction. It extends the basic idea in [6] from the single-channel to the multichannel cases. Same as in [6], the approach developed in this paper also works in the STFT domain; but it considers not only the interframe information but also the interchannel information provided by an array of microphones.

2. SIGNAL MODEL

We consider the signal model in which a microphone array with M sensors captures a convolved speech signal in some noise field, as illustrated in Fig. 1. The received signals are expressed as [7], [8]

$$\begin{aligned} y_m(t) &= g_m(t) * s(t) + v_m(t) \\ &= x_m(t) + v_m(t),\ m = 1, 2, \ldots, M, \end{aligned} \tag{1}$$

This work was supported in part by the NSFC "Distinguished Young Scientists Fund" under grant No. 61425005.

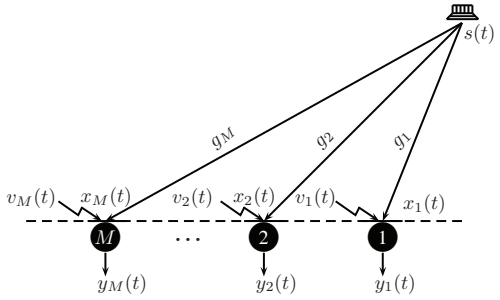

Fig. 1. Illustration of a microphone array system for sound acquisition.

where $g_m(t)$ is the acoustic impulse response from the unknown speech source, $s(t)$, to the mth microphone, $*$ stands for linear convolution, and $v_m(t)$ is the additive noise at microphone m. We assume that the convolved speech and noise signals are uncorrelated, zero mean, real, and broadband. By definition, the terms $x_m(t)$, $m = 1, 2, \ldots, M$ are coherent across the array.

In the STFT domain, the received signals are expressed as [1], [3]

$$\begin{aligned} Y_m(k,n) &= G_m(k)S(k,n) + V_m(k,n) \\ &= X_m(k,n) + V_m(k,n),\ m = 1, 2, \ldots, M, \end{aligned} \tag{2}$$

where k is the frequency index, n is the time-frame index, and $Y_m(k,n)$, $G_m(k)$, $S(k,n)$, $X_m(k,n)$, and $V_m(k,n)$ are the STFTs of $y_m(t)$, $g_m(t)$, $s(t)$, $x_m(t)$, and $v_m(t)$, respectively.

By considering N consecutive time-frames for each microphone, we can rewrite the observation signals into the following vector form:

$$\mathbf{y}_m(k,n) = \mathbf{x}_m(k,n) + \mathbf{v}_m(k,n),\ m = 1, 2, \ldots, M, \tag{3}$$

where

$$\begin{aligned} \mathbf{y}_m(k,n) \triangleq \big[\ & Y_m(k,n) \quad Y_m(k,n-1) \\ & \cdots \quad Y_m(k,n-N+1)\ \big]^T, \end{aligned} \tag{4}$$

the superscript ^{T} denotes transpose of a vector or a matrix, and $\mathbf{x}_m(k,n)$ and $\mathbf{v}_m(k,n)$ are defined in a similar way to $\mathbf{y}_m(k,n)$.

Stacking the M vectors in (3) into a long vector, we get

$$\begin{aligned} \underline{\mathbf{y}}(k,n) &\triangleq \big[\ \mathbf{y}_1^T(k,n) \quad \mathbf{y}_2^T(k,n) \quad \cdots \quad \mathbf{y}_M^T(k,n)\ \big]^T \\ &= \underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n), \end{aligned} \tag{5}$$

where $\underline{\mathbf{y}}(k,n)$ is a vector of length MN, and $\underline{\mathbf{x}}(k,n)$ and $\underline{\mathbf{v}}(k,n)$ are defined in a similar way to $\underline{\mathbf{y}}(k,n)$. The correlation matrix of

$\underline{\mathbf{y}}(k,n)$ is then

$$\begin{aligned}\underline{\mathbf{\Phi}}_{\underline{\mathbf{y}}}(k,n) &\triangleq E\left[\underline{\mathbf{y}}(k,n)\underline{\mathbf{y}}^H(k,n)\right] \\ &= \underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n) + \underline{\mathbf{\Phi}}_{\underline{\mathbf{v}}}(k,n),\end{aligned} \tag{6}$$

where the superscript ^{H} is the conjugate-transpose operator, and $\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)$ and $\underline{\mathbf{\Phi}}_{\underline{\mathbf{v}}}(k,n)$ are the correlation matrices of $\underline{\mathbf{x}}(k,n)$ and $\underline{\mathbf{v}}(k,n)$, respectively.

With the above signal model, the objective of noise reduction is to estimate $X_1(k,n)$ given $\underline{\mathbf{y}}(k,n)$.

3. MAXIMUM SNR NOISE REDUCTION FILTER

Given the signal model in (5), the problem of noise reduction is to achieve an estimate of $X_1(k,n)$ from the observation vector $\underline{\mathbf{y}}(k,n)$. This can be done by applying a complex finite-impulse-response (FIR) filter, $\underline{\mathbf{h}}(k,n)$, of length MN, to the noisy signal vector, $\underline{\mathbf{y}}(k,n)$, i.e

$$\begin{aligned}Z(k,n) &= \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{y}}(k,n) \\ &= X_{\text{fd}}(k,n) + V_{\text{rn}}(k,n),\end{aligned} \tag{7}$$

where $X_{\text{fd}}(k,n) \triangleq \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{x}}(k,n)$ is the filtered desired signal and $V_{\text{rn}}(k,n) \triangleq \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{v}}(k,n)$ is the residual noise. The variance of $Z(k,n)$ is then

$$\phi_Z(k,n) \triangleq E\left[|Z(k,n)|^2\right] = \phi_{X_{\text{fd}}}(k,n) + \phi_{V_{\text{rn}}}(k,n), \tag{8}$$

where $\phi_{X_{\text{fd}}}(k,n) = \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{h}}(k,n)$ and $\phi_{V_{\text{rn}}}(k,n) = \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{v}}}(k,n)\underline{\mathbf{h}}(k,n)$ are the variances of $X_{\text{fd}}(k,n)$ and $V_{\text{rn}}(k,n)$, respectively.

The subband input and output SNRs are defined, respectively, as

$$\text{iSNR}(k,n) \triangleq \frac{\phi_{X_1}(k,n)}{\phi_{V_1}(k,n)} \tag{9}$$

and

$$\text{oSNR}\left[\underline{\mathbf{h}}(k,n)\right] \triangleq \frac{\phi_{X_{\text{fd}}}(k,n)}{\phi_{V_{\text{rn}}}(k,n)} = \frac{\underline{\mathbf{h}}^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{h}}(k,n)}{\underline{\mathbf{h}}^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{v}}}(k,n)\underline{\mathbf{h}}(k,n)}, \tag{10}$$

where $\phi_{X_1}(k,n)$ and $\phi_{V_1}(k,n)$ are the variances of $X_1(k,n)$ and $V_1(k,n)$, respectively.

Now, we want to find a filter that maximizes the subband output SNR, i.e., $\text{oSNR}\left[\underline{\mathbf{h}}(k,n)\right]$. It is seen from (10) that $\text{oSNR}\left[\underline{\mathbf{h}}(k,n)\right]$ is in the form of the generalized Rayleigh quotient. Let $\underline{\lambda}_1(k,n)$ be the maximum eigenvalue of the matrix $\underline{\mathbf{\Phi}}_{\underline{\mathbf{v}}}^{-1}(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)$. We denote by $\underline{\mathbf{b}}_1(k,n)$ the eigenvector associated with $\underline{\lambda}_1(k,n)$. It can be shown that the filter that maximizes $\text{oSNR}\left[\underline{\mathbf{h}}(k,n)\right]$ is

$$\underline{\mathbf{h}}_{\max}(k,n) = \underline{\beta}(k,n)\underline{\mathbf{b}}_1(k,n), \tag{11}$$

where $\underline{\beta}(k,n) \neq 0$ is an arbitrary complex number. It can be checked that

$$\text{oSNR}\left[\underline{\mathbf{h}}_{\max}(k,n)\right] = \underline{\lambda}_1(k,n) \geq \text{oSNR}\left[\underline{\mathbf{i}}_1\right] = \text{iSNR}(k,n), \tag{12}$$

where $\underline{\mathbf{i}}_1$ is the first column of the $MN \times MN$ identity matrix $\mathbf{I}_{MN}$.

The choice of the value of $\underline{\beta}(k,n)$ is extremely important in practice. With a poor choice of this parameter, the desired signal can be severely distorted. To find the proper value of $\underline{\beta}(k,n)$, let us define the distortion-based mean-square error (MSE) at frequency index k:

$$J\left[\underline{\mathbf{h}}(k,n)\right] \triangleq E\left\{\left|X_1(k,n) - \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{x}}(k,n)\right|^2\right\}, \tag{13}$$

which can be rewritten as

$$\begin{aligned}J\left[\underline{\mathbf{h}}(k,n)\right] &= \phi_{X_1}(k,n) + \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{h}}(k,n) \\ &\quad - \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}_1 - \underline{\mathbf{i}}_1^T\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{h}}(k,n).\end{aligned} \tag{14}$$

Substituting (11) into (14), we get the distortion-based MSE associated with the maximum SNR filter, i.e.,

$$\begin{aligned}J\left[\underline{\mathbf{h}}_{\max}(k,n)\right] &= \phi_{X_1}(k,n) + \left|\underline{\beta}(k,n)\right|^2\underline{\mathbf{b}}_1^H(k,n)\times \\ &\quad \underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{b}}_1(k,n) - \underline{\beta}^*(k,n)\underline{\mathbf{b}}_1^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}_1 \\ &\quad - \underline{\beta}(k,n)\underline{\mathbf{i}}_1^T\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{b}}_1(k,n),\end{aligned} \tag{15}$$

where the superscript * is the complex-conjugate operator.

Now, we want to find the $\underline{\beta}(k,n)$ parameter that minimizes $J\left[\underline{\mathbf{h}}_{\max}(k,n)\right]$. Differentiating $J\left[\underline{\mathbf{h}}_{\max}(k,n)\right]$ with respect to $\underline{\beta}^*(k,n)$ and equating the result to 0, We obtain

$$\begin{aligned}\underline{\beta}(k,n) &= \frac{\underline{\mathbf{b}}_1^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}_1}{\underline{\mathbf{b}}_1^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{b}}_1(k,n)} \\ &= \frac{\underline{\mathbf{b}}_1^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}_1}{\underline{\lambda}_1(k,n)}.\end{aligned} \tag{16}$$

Substituting (16) into (11), we find that the optimal maximum SNR filter with minimum distortion is

$$\underline{\mathbf{h}}_{\max}(k,n) = \frac{\underline{\mathbf{b}}_1^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}_1}{\underline{\lambda}_1(k,n)}\underline{\mathbf{b}}_1(k,n). \tag{17}$$

4. SIMULATIONS

In this section, we study the performance of the optimal maximum SNR filter with minimum distortion through simulations. The layout of the simulation setup is illustrated in Fig. 2, where a linear array of 4 omnidirectional microphones is placed in a square room of size 4 m × 4 m × 4 m. The four microphones are located, respectively, at (x, 2.0, 1.6), where $x = 2.1 : 0.1 : 2.3$. A loudspeaker is placed at (2.0, 2.0, 1.4). The acoustic channel impulse responses from the

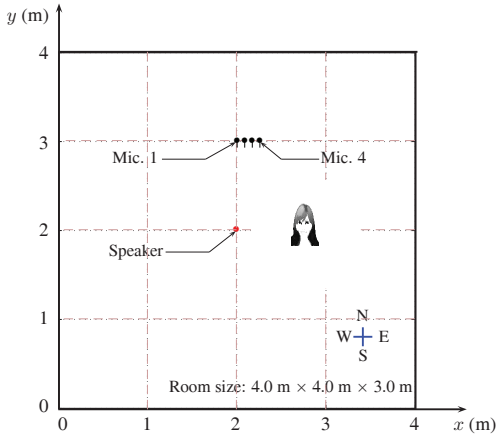

Fig. 2. Layout of the simulation setup (coordinates in meters).

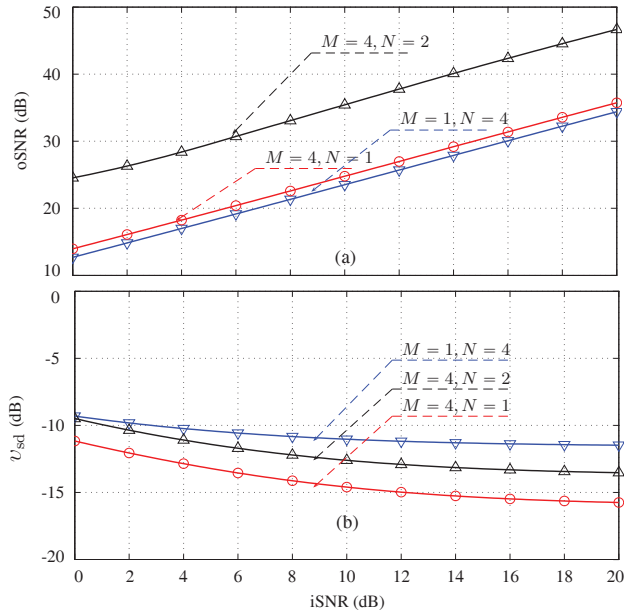

Fig. 3. Performance of the multichannel maximum SNR filter in the STFT domain as a function of the input SNR in white Gaussian noise: (a) output SNR and (b) speech distortion index.

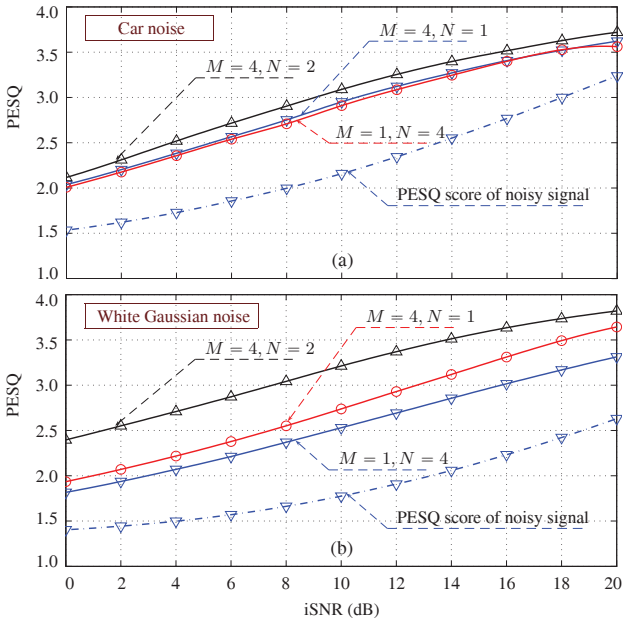

Fig. 4. PESQ score of the multichannel maximum SNR filter in the STFT domain as a function of the input SNR in different noises: (a) car noise and (b) white Gaussian noise.

source to the four microphones are generated with the image model method [9]. Then, the microphone outputs are generated by convolving the source signal with the corresponding impulse responses and noise is then added to the convolved signals to control the SNR level.

To implement the maximum SNR filter given in (17), we first divide the array signals into overlapping frames (of size $K = 128$) with 75% overlapping. A Kaiser window is then applied to each frame and the windowed frame signal is subsequently transformed into the STFT domain using a 128-point FFT. To compute the maximum SNR filter, we need to know the correlation matrices $\mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n)$ and $\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$. The former can be directly computed from the noisy signal. However, to compute the $\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$ matrix, we need a noise estimator such as the ones developed in [10] and [11]. But to place our focal point on illustrating the performance with the maximum SNR filter, we directly compute the noise correlation matrix from the noise signal in this paper. Specifically, the initial estimates $\mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n)$ and $\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$ are computed use the first 100 frames with a sample average. Then at every time-frame, the two matrices are updated using a recursive method as in [12].

After obtaining the estimate of the correlation matrices $\mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n)$ and $\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$, the clean speech correlation matrix is computed as $\widehat{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n) = \widehat{\mathbf{\Phi}}_{\underline{\mathbf{y}}}(k,n) - \widehat{\mathbf{\Phi}}_{\underline{\mathbf{v}}}(k,n)$. In order to ensure that the matrix $\widehat{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)$ is positive semi-definite, we apply the eigenvalue decomposition and set the non-positive eigenvalues to zero.

We use the output SNR, the speech distortion index [13], and the perceptual evaluation of speech quality (PESQ) [14] as the performance measures. To compute these measures, we first estimate $Z(k,n)$, $X_{\mathrm{fd}}(k,n)$, and $V_{\mathrm{rn}}(k,n)$ in the STFT domain, and then transform them into the time domain. As a result, we can compute the output SNR after noise reduction according to the following definition [13]:

$$\mathrm{oSNR} = \frac{E\left[x_{\mathrm{fd}}^2(t)\right]}{E\left[v_{\mathrm{rn}}^2(t)\right]}, \tag{18}$$

where $x_{\mathrm{fd}}(t)$ is the time domain filtered desired signal [inverse FFT of $X_{\mathrm{fd}}(k,n)$], and $v_{\mathrm{rn}}(t)$ is the time domain residual noise [inverse FFT of $V_{\mathrm{rn}}(k,n)$]. Similarly, the speech distortion index is computed according to its definition [13]:

$$v_{\mathrm{sd}} = \frac{E\left\{\left[x_{\mathrm{fd}}(t) - x(t)\right]^2\right\}}{E\left[x^2(t)\right]}. \tag{19}$$

In the first experiment, we examine the performance of multichannel maximum SNR filter in different SNR conditions. We set the reflection coefficients to 0.8 (the reverberation time T_{60} is approximately 240 ms), and change the input SNR from 0 dB to 20 dB. We study the maximum SNR filter in three different conditions: $M = 1, N = 4$ (single-channel maximum SNR filter), $M = 4, N = 1$ (multichannel maximum SNR gain without using interframe information), and $M = 4, N = 2$ (multichannel maximum SNR filter considering both the interchannel and interframe information). The results are plotted in Fig. 3. As one can see from Fig. 3, the single-channel maximum SNR filter ($M = 1, N = 4$) and the multichannel maximum SNR gain ($M = 4, N = 1$) have similar output SNRs in a given input SNR condition. In comparison, the multichannel maximum SNR filter ($M = 4, N = 2$) achieves much higher output SNRs than the other two.

To further evaluate the multichannel maximum SNR filter, we examine, in this simulation, the PESQ performance, which has been found to have higher correlations, than many other widely known objective measures, with the subjective ratings of overall quality of enhanced speech signals. Same as in the previous simulation, we set the room reflection coefficients to 0.8 ($T_{60} \simeq 240$ ms). We consider two types of noise: white Gaussian noise and car noise (recorded in a sedan car running at 50 miles/hour on a highway). The input SNR varies from 0 dB to 20 dB. Signals from a female talker and a male talker are used as the source signal. For every given noise condition, the PESQ mean opinion score (MOS) for each talker is computed first. These raw PESQ MOS scores are then mapped to the PESQ

Table 1. Noise reduction performance of the multichannel maximum SNR filter ($M = 4$, $N = 2$, input SNR is 10 dB).

	White Gaussian noise				Car noise			
T_{60} (ms)	oSNR (dB)	$\upsilon_{\rm sd}$	$\rm PESQ_{\it y}$	$\rm PESQ_{\it z}$	oSNR (dB)	$\upsilon_{\rm sd}$	$\rm PESQ_{\it y}$	$\rm PESQ_{\it z}$
168	35.5	0.059	2.19	3.35	22.1	0.081	2.46	3.13
256	35.3	0.068	2.30	3.41	22.4	0.096	2.55	3.22
459	35.3	0.074	2.41	3.45	24.6	0.109	2.65	3.31
628	35.5	0.076	2.46	3.45	25.1	0.114	2.70	3.35

listening quality objective (LQO) according to [14]

$$\rm PESQ = 0.999 + \frac{4}{1 + e^{-1.4945 \times PESQ_{MOS}} + 4.6607}. \quad (20)$$

The results are plotted in Fig. 4. As seen, the multichannel maximum SNR filter can improve the PESQ score significantly in all studied input SNR conditions with the two types of noise.

In the last simulation, we examine the performance of the multichannel maximum SNR filter in different reverberation conditions. The parameters are chosen same as in the previous simulation. Also, we consider both the white Gaussian noise and the car noise with an input SNR of 10 dB. We study four reverberation conditions with the room reflection coefficients are, respectively, 0.7, 0.8, 0.9, and 0.95 (the corresponding reverberation time T_{60} are approximately 168 ms, 256 ms, 459 ms, and 628 ms). The results are shown in Table 1, where the $\rm PESQ_{\it y}$ is the PESQ of the noisy signal and $\rm PESQ_{\it z}$ is the PESQ of the enhanced signal. It is seen that, given a noise condition, both the output SNR and speech distortion index of the maximum SNR filter change with reverberation. In contrast, the speech distortion index changes more dramatically with the reverberation time. The larger the reverberation time, the larger is the speech distortion index. As a result, there is less PESQ improvement as reverberation increases. This can be easily explained. As the reverberation becomes stronger, it becomes more difficult to predict the signal observed at one microphone from the signals received at the array. Consequently, the speech distortion index increases with the reverberation time while the PESQ gain decreases accordingly.

5. CONCLUSIONS

In this paper, we studied the problem of multichannel noise reduction in the STFT domain. Since one of the major objectives of noise reduction is to reduce the amount of noise, thereby enhancing the desired signal, we derived an optimal filter that maximizes the output SNR. In contrast with the maximum SNR filters studied in the literature and also in our recent work, the approach in this paper considered both the interframe and interchannel information in the STFT domain and the multichannel maximum SNR filter was derived with the minimum distortion constraint. Simulations showed that the developed multichannel maximum SNR filter can significantly improve both the output SNR and the speech quality (as indicated by the PESQ improvements), which indicate the great potential of this filter.

6. REFERENCES

[1] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Berlin, Germany: Springer-Verlag, 2009.

[2] P. Vary, "Noise suppression by spectral magnitude estimation–mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, July 1985.

[3] J. Benesty, J. Chen, and E. A. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer-Verlag, 2012.

[4] K. K. Paliwal, B. Schwerin, and K. K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Commun.*, vol. 54, pp. 282–305, Jan. 2012.

[5] A. Schasse and R. Martin, "Online inter-frame correlation estimation methods for speech enhancement in frequency subbands," in *Proc. IEEE ICASSP*, pp. 7482–7486, 2013.

[6] G. Huang, J. Benesty, T. Long, and J. Chen, "A family of maximum SNR filters for noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 2034–2047, Dec. 2014.

[7] M. Brandstein and E. D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[8] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *J. Acoust. Soc. Am.*, vol. 65, p. 943–950, Apr. 1979.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[11] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 11, pp. 466–475, Sep. 2003.

[12] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1256–1269, 2012.

[13] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1218–1234, Jul. 2006.

[14] Mapping Function for Transforming Raw Result Scores to MOS-LQO, ITU-T Rec. P. 862.1, 2003.