A Family of Maximum SNR Filters for Noise Reduction

Gongping Huang, *Student Member, IEEE*, Jacob Benesty, Tao Long, and Jingdong Chen, *Senior Member, IEEE*

***Abstract*—This paper is devoted to the study and analysis of the maximum signal-to-noise ratio (SNR) filters for noise reduction both in the time and short-time Fourier transform (STFT) domains with one single microphone and multiple microphones. In the time domain, we show that the maximum SNR filters can significantly increase the SNR but at the expense of tremendous speech distortion. As a consequence, the speech quality improvement, measured by the perceptual evaluation of speech quality (PESQ) algorithm, is marginal if any, regardless of the number of microphones used. In the STFT domain, the maximum SNR filters are formulated by considering the interframe information in every frequency band. It is found that these filters not only improve the SNR, but also improve the speech quality significantly. As the number of input channels increases so is the gain in SNR as well as the speech quality. This demonstrates that the maximum SNR filters, particularly the multichannel ones, in the STFT domain may be of great practical value.**

***Index Terms*—Maximum SNR filter, multichannel, noise reduction, short-time Fourier transform (STFT) domain, single channel, speech enhancement, time domain.**

I. Introduction

NOISE reduction, sometimes also referred to as speech enhancement, is a problem of recovering a clean speech from its microphone observations corrupted by additive noise, thereby improving the signal-to-noise ratio (SNR) to make the observation signals sound more natural and comfortable with a higher perceptual quality. This has long been a major problem in signal processing for voice communications and human-machine interfaces. A significant number of efforts have been devoted to this problem in the literature [1]–[4]. Most early studies mainly focused on using a single microphone (the problem is then referred to as the single-channel noise reduction) as most communication devices at that time were equipped with only one microphone. In this case, the problem can be attacked with either signal processing methods [4]–[6] or signal processing combined with auditory properties [7], [8]. Recently, multiple microphones or microphone arrays have been widely investigated in this context (the problem is then referred to as the multichannel noise reduction). It has been found that the flexibility in dealing with noise and the noise reduction performance can increase with the number of microphones [2], [9]–[14].

In the time domain, the noise reduction problem can be formulated as a linear filtering technique either on a sample or on a block basis [15]. In the former case, a sample of the desired clean speech is estimated by passing a vector of the noisy signal through a finite-impulse-response (FIR) filter [9], [15]. Similarly, in the block formulation, a block of the clean signal is estimated by passing a vector of the noisy signal through a filtering matrix [15]. In both situations, the most critical issue of noise reduction is to find an optimal filter or filtering matrix that can significantly mitigate the noise effect while maintaining the filtered speech signal perceptually close to its original form. Typically, the optimal filter (or filtering matrix) is designed from the mean-squared error (MSE) criterion [9], [16], [17]. Since one of the major objectives of noise reduction is to reduce noise (i.e., improve the SNR) [18], [19], thereby improving speech quality, it is natural to think of the optimal filter that maximizes the output SNR, leading to the so-called maximum SNR filter [9]. However, it has been observed that this filter is not very helpful in enhancing speech quality or intelligibility since it introduces significant speech distortion.

Another popular way of formulating the problem is to convert the original problem into the short-time Fourier transform (STFT) domain [18]–[23]. With this approach, the most critical issue of noise reduction is to design an optimal filter in every STFT frequency band. The earliest effort on this can be dated back to the well-known spectral subtraction method [1], [4], which is still popularly used in many today's systems [20], [24], [25]. However, this approach was developed in a heuristic way and it has no optimality properties associated with it. A great deal of efforts were then devoted to finding optimal noise reduction filters in a statistical estimation framework. Many such filters were deduced, including the minimum mean-squared error (MMSE) estimator [26], [27], the maximum likelihood (ML) estimator [28], the maximum *a posteriori* (MAP) estimator [29], etc. Most of these filters were then found to be closely related to the well-known Wiener filter [30], which is expected since most of these approaches make the common assumption that the speech and noise signals are Gaussian distributed. Another common assumption that these methods make is that the STFT coefficients from different frequency bands and time frames are independent of each other. With this assumption, the noise reduction filter in a given frequency band turns out to be a gain and, therefore, the problem of noise reduction becomes one of finding an optimal gain [18]. Since a gain does not change the subband input SNR, it is not possible to design a filter that can maximize the subband output SNR. However, the fullband output SNR can

Manuscript received January 15, 2014; revised May 05, 2014; accepted September 22, 2014. Date of publication September 26, 2014; date of current version October 02, 2014. This work was supported in part by the Chinese Specialized Research Fund for the Doctoral Program of High Education (20136102110010). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Roberto Togneri.

G. Huang and J. Chen are with the Center of Immersive and Intelligent Acoustics, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: gongpinghuang@gmail.com; jingdongchen@ieee.org).

J. Benesty is with the INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada (e-mail: benesty@emt.inrs.ca).

T. Long is with the Xi'an Jiaotong University, Xi'an 710049, China (e-mail: longtao2002@163.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2360643

be improved. As a matter of fact, if we put the gains from all the frequency bands into a vector, the optimal filtering vector that maximizes the fullband output SNR is a unit vector with only one non-zero component [18]. The non-zero component corresponds to the subband that has the largest subband input SNR among all the subbands. However, this filtering vector can cause significant speech distortion, making the speech unintelligible; consequently, it is never used in practice.

Recently, a new noise reduction framework was developed in the STFT domain, which considers the interframe information [16], [18], [21]. In this situation, the filter in every STFT frequency band is no longer a gain, but a filtering vector. With this new framework, it is possible to design an optimal filter that can improve both the subband and fullband SNRs. This provides an opportunity to design new forms of maximum SNR filters. This paper is, therefore, devoted to the study and analysis of the maximum SNR filters for noise reduction. Although the major focus of this paper is on the maximum SNR filters in the STFT domain, we also discuss these filters in the time domain for the purpose of completeness and comparison. In the time domain, we discuss these filters for both the single-channel and multichannel cases. We show that the maximum SNR filters can significantly increase the SNR at the expense of tremendous speech distortion and, as a consequence, the speech quality improvement is marginal if any, regardless of the number of microphones used. In the STFT domain, the maximum SNR filters are formulated by considering the interframe information in every frequency band. These STFT-domain maximum SNR filters improve not only the SNR, but also the speech quality significantly. The more the number of input channels, the better is the gain in SNR and speech quality.

The rest of this paper is organized as follows. In Section II, we discuss the single-channel maximum SNR filter for noise reduction in the time domain. Section III continues the discussion of the maximum SNR filter in the time domain but with multiple microphones. We then describe, in Section IV, how to design the maximum SNR filter in the STFT domain for the single-channel case. The multichannel maximum SNR filter in the STFT domain is addressed in Section V. In Section VI, we present some experiments to validate the theoretical analysis. Finally, some conclusions are drawn in Section VII.

II. Single-Channel Noise Reduction in the Time Domain

A. Signal Model and Problem Formulation

The noise reduction (speech enhancement) problem considered in this section is one of recovering the zero-mean desired signal (or clean signal) $x(t)$, t being the discrete-time index, from the noisy observation (microphone signal) [9], [15]:

$$y(t) = x(t) + v(t), \tag{1}$$

where $v(t)$ is the unwanted additive noise, which is assumed to be a zero-mean random process, white or colored, but uncorrelated with $x(t)$. All signals are considered to be real and broadband. The signal model given in (1) can be put into a vector form by accumulating the L most recent successive time samples, i.e.,

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{v}(t), \tag{2}$$

where

$$\mathbf{y}(t) \triangleq [y(t)\ y(t-1)\ \cdots\ y(t-L+1)]^T \tag{3}$$

is a vector of length L, the superscript ^{T} denotes transpose of a vector or a matrix, and $\mathbf{x}(t)$ and $\mathbf{v}(t)$ are defined in a similar way to $\mathbf{y}(t)$ in (3). Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the correlation matrix (of size $L \times L$) of the noisy signal can be written as

$$\mathbf{R}_{\mathbf{y}} \triangleq E\left[\mathbf{y}(t)\mathbf{y}^T(t)\right] = \mathbf{R}_{\mathbf{x}} + \mathbf{R}_{\mathbf{v}}, \tag{4}$$

where $E[\cdot]$ denotes mathematical expectation, and $\mathbf{R}_{\mathbf{x}} \triangleq E[\mathbf{x}(t)\mathbf{x}^T(t)]$ and $\mathbf{R}_{\mathbf{v}} \triangleq E[\mathbf{v}(t)\mathbf{v}^T(t)]$ are the correlation matrices of $\mathbf{x}(t)$ and $\mathbf{v}(t)$, respectively. The noise correlation matrix, $\mathbf{R}_{\mathbf{v}}$, is assumed to be full rank, i.e., its rank is equal to L. Note that the correlation matrices $\mathbf{R}_{\mathbf{y}}$ and $\mathbf{R}_{\mathbf{x}}$ are in general time-varying and $\mathbf{R}_{\mathbf{v}}$ can be either time invariant or time-varying depending on the stationarity of the noise signal. However, for the simplicity of notation, we will not consider the time dependency of these matrices for the time being; but we will come back to this point in Section VI on simulations.

Let us define the desired signal vector of length P ($1 \leq P \leq L$):

$$\widetilde{\mathbf{x}}(t) \triangleq [x(t)\ x(t-1)\ \cdots\ x(t-P+1)]^T. \tag{5}$$

The objective of single-channel noise reduction in the time domain is to estimate the desired signal vector, $\widetilde{\mathbf{x}}(t)$, given the observation signal vector, $\mathbf{y}(t)$. This should be done in such a way that the noise is reduced as much as possible with little or even no distortion to the desired signal.

B. Linear Estimation and Performance Measures

The desired signal vector, $\widetilde{\mathbf{x}}(t)$, can be estimated by applying a linear transformation to the observation signal vector, $\mathbf{y}(t)$, i.e.,

$$\widetilde{\mathbf{z}}(t) = \mathbf{H}\mathbf{y}(t) = \widetilde{\mathbf{x}}_{\mathrm{fd}}(t) + \widetilde{\mathbf{v}}_{\mathrm{rn}}(t), \tag{6}$$

where $\widetilde{\mathbf{z}}(t)$ is supposed to be the estimate of $\widetilde{\mathbf{x}}(t)$,

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \vdots \\ \mathbf{h}_P^T \end{bmatrix} \tag{7}$$

is a rectangular filtering matrix of size $P \times L$, $\mathbf{h}_p$ ($p = 1, 2, \ldots, P$) are FIR filters of length L with

$$\mathbf{h}_p = [h_{p,0}\ h_{p,1}\ \cdots\ h_{p,L-1}]^T, \tag{8}$$

$\widetilde{\mathbf{x}}_{\mathrm{fd}}(t) \triangleq \mathbf{H}\mathbf{x}(t)$ is the filtered desired signal, and $\widetilde{\mathbf{v}}_{\mathrm{rn}}(t) \triangleq \mathbf{H}\mathbf{v}(t)$ is the residual noise. The correlation matrix of $\widetilde{\mathbf{z}}(t)$ is then

$$\mathbf{R}_{\widetilde{\mathbf{z}}} \triangleq E\left[\widetilde{\mathbf{z}}(t)\widetilde{\mathbf{z}}^T(t)\right] = \mathbf{R}_{\widetilde{\mathbf{x}}_{\mathrm{fd}}} + \mathbf{R}_{\widetilde{\mathbf{v}}_{\mathrm{rn}}}, \tag{9}$$

where $\mathbf{R}_{\widetilde{\mathbf{x}}_{\mathrm{fd}}} = \mathbf{H}\mathbf{R}_{\mathbf{x}}\mathbf{H}^T$ and $\mathbf{R}_{\widetilde{\mathbf{v}}_{\mathrm{rn}}} = \mathbf{H}\mathbf{R}_{\mathbf{v}}\mathbf{H}^T$.

To facilitate the analysis and interpretation of the noise reduction performance, let us give two useful performance measures: the SNRs (before and after filtering) and speech distortion index. From the signal model given in (1), we define the input SNR as

$$\mathrm{iSNR} \triangleq \frac{\sigma_x^2}{\sigma_v^2}, \tag{10}$$

where $\sigma_x^2 \triangleq E[x^2(t)]$ and $\sigma_v^2 \triangleq E[v^2(t)]$ are the variances of $x(t)$ and $v(t)$, respectively. The output SNR, after noise reduction, can be defined as

$$\mathrm{oSNR}(\mathbf{H}) \triangleq \frac{\mathrm{tr}\left(\mathbf{R}_{\widetilde{\mathbf{x}}_{\mathrm{fd}}}\right)}{\mathrm{tr}\left(\mathbf{R}_{\widetilde{\mathbf{v}}_{\mathrm{rn}}}\right)} = \frac{\sum_{p=1}^{P}\mathbf{h}_p^T\mathbf{R}_{\mathbf{x}}\mathbf{h}_p}{\sum_{p=1}^{P}\mathbf{h}_p^T\mathbf{R}_{\mathbf{v}}\mathbf{h}_p}, \tag{11}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix.

The distortion-based mean-squared error (MSE) is given by

$$J(\mathbf{H}) \triangleq E\left\{\left[\widetilde{\mathbf{x}}_{\mathrm{fd}}(t) - \widetilde{\mathbf{x}}(t)\right]^T\left[\widetilde{\mathbf{x}}_{\mathrm{fd}}(t) - \widetilde{\mathbf{x}}(t)\right]\right\}, \tag{12}$$

from which we deduce the speech distortion index [15]:

$$\upsilon\left(\mathbf{H}\right)=\frac{J\left(\mathbf{H}\right)}{P\sigma_x^2}, \tag{13}$$

which is lower bounded by 0 and expected to be upper bounded by 1 for optimal filters. A small value of $\upsilon(\mathbf{H})$ implies little distortion of the desired signal. The larger the value of this index, the more the desired signal is distorted.

C. Maximum SNR Filter

We show here how to maximize the output SNR, which is defined in (11). This procedure leads to the maximum SNR filtering matrix, which is slightly different from the one presented in [15] where no minimum distortion constraint is used.

It can be checked that [15], [18]

$$\mathrm{oSNR}\left(\mathbf{H}\right)\leq\max_p\frac{\mathbf{h}_p^T\mathbf{R}_{\mathbf{x}}\mathbf{h}_p}{\mathbf{h}_p^T\mathbf{R}_{\mathbf{v}}\mathbf{h}_p}=\chi. \tag{14}$$

Let λ_1 be the maximum eigenvalue of the matrix $\mathbf{R}_{\mathbf{v}}^{-1}\mathbf{R}_{\mathbf{x}}$ with corresponding eigenvector $\mathbf{b}_1$. The maximum SNR filtering matrix is given by

$$\mathbf{H}_{\max}=\begin{bmatrix}\beta_1\mathbf{b}_1^T\\ \beta_2\mathbf{b}_1^T\\ \vdots\\ \beta_P\mathbf{b}_1^T\end{bmatrix}, \tag{15}$$

where $\beta_p, p=1,2,\ldots,P$ are arbitrary real numbers with at least one of them different from 0. The corresponding output SNR is

$$\mathrm{oSNR}\left(\mathbf{H}_{\max}\right)=\lambda_1. \tag{16}$$

The output SNR with the maximum SNR filtering matrix is always greater than or equal to the input SNR, i.e., $\mathrm{oSNR}(\mathbf{H}_{\max})\geq\mathrm{iSNR}$. We also have $\mathrm{oSNR}(\mathbf{H})\leq\lambda_1,\forall\mathbf{H}$.

The choice of the values of $\beta_p, p=1,2,\ldots,P$, is extremely important in practice; with a poor choice of these values, the desired signal vector can be severely distorted. Therefore, the β_p's should be found in such a way that distortion is minimized. We can rewrite the distortion-based MSE as

$$\begin{aligned}J\left(\mathbf{H}\right)&=\mathrm{tr}\left(\mathbf{I}_{\mathrm{i}}\mathbf{R}_{\mathbf{x}}\mathbf{I}_{\mathrm{i}}^T\right)+\mathrm{tr}\left(\mathbf{H}\mathbf{R}_{\mathbf{x}}\mathbf{H}^T\right)-2\cdot\mathrm{tr}\left(\mathbf{H}\mathbf{R}_{\mathbf{x}}\mathbf{I}_{\mathrm{i}}^T\right)\\&=\mathrm{tr}\left(\mathbf{I}_{\mathrm{i}}\mathbf{R}_{\mathbf{x}}\mathbf{I}_{\mathrm{i}}^T\right)+\sum_{p=1}^{P}\mathbf{h}_p^T\mathbf{R}_{\mathbf{x}}\mathbf{h}_p-2\sum_{p=1}^{P}\mathbf{h}_p^T\mathbf{R}_{\mathbf{x}}\mathbf{i}_p,\end{aligned} \tag{17}$$

where

$$\mathbf{I}_{\mathrm{i}}=\begin{bmatrix}\mathbf{I}_P & \mathbf{0}_{P\times(L-P)}\end{bmatrix}, \tag{18}$$

$\mathbf{I}_P$ is the $P\times P$ identity matrix, and $\mathbf{i}_p$ is the pth column of the $L\times L$ identity matrix, $\mathbf{I}_L$, and $\mathbf{0}_{P\times(L-P)}$ is matrix of size $P\times(L-P)$ with all its elements being 0.

Substituting (15) into (19), we get

$$\begin{aligned}J(\mathbf{H}_{\max})&=\mathrm{tr}\left(\mathbf{I}_{\mathrm{i}}\mathbf{R}_{\mathbf{x}}\mathbf{I}_{\mathrm{i}}^T\right)+\mathbf{b}_1^T\mathbf{R}_{\mathbf{x}}\mathbf{b}_1\sum_{p=1}^{P}\beta_p^2\\&\quad-2\sum_{p=1}^{P}\beta_p\mathbf{b}_1^T\mathbf{R}_{\mathbf{x}}\mathbf{i}_p.\end{aligned} \tag{19}$$

Minimizing (19) with respect to the β_p's, we find

$$\beta_p=\frac{\mathbf{b}_1^T\mathbf{R}_{\mathbf{x}}\mathbf{i}_p}{\mathbf{b}_1^T\mathbf{R}_{\mathbf{x}}\mathbf{b}_1}=\frac{\mathbf{b}_1^T\mathbf{R}_{\mathbf{x}}\mathbf{i}_p}{\lambda_1}, p=1,2,\ldots,P, \tag{20}$$

where $\lambda_1=\mathbf{b}_1^T\mathbf{R}_{\mathbf{x}}\mathbf{b}_1$. Substituting these optimal values into (15), we obtain the maximum SNR filtering matrix with minimum signal distortion:

$$\mathbf{H}_{\max}=\mathbf{I}_{\mathrm{i}}\mathbf{R}_{\mathbf{x}}\frac{\mathbf{b}_1\mathbf{b}_1^T}{\lambda_1}=\mathbf{I}_{\mathrm{i}}\mathbf{R}_{\mathbf{v}}\mathbf{b}_1\mathbf{b}_1^T. \tag{21}$$

We deduce that

$$\upsilon\left(\mathbf{H}_{\max}\right)=1-\frac{1}{P\sigma_x^2\lambda_1}\sum_{p=1}^{P}\left(\mathbf{b}_1^T\mathbf{R}_{\mathbf{x}}\mathbf{i}_p\right)^2. \tag{22}$$

III. Multichannel Speech Enhancement in the Time Domain

A. Signal Model and Problem Formulation

In this section, we consider the signal model in which a microphone array with M sensors captures a convolved source signal in some noise field. The received signals are expressed as [2], [19]

$$\begin{aligned}y_m(t)&=g_m(t)*s(t)+v_m(t)\\&=x_m(t)+v_m(t), m=1,2,\ldots,M,\end{aligned} \tag{23}$$

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, $x_m(t)$, $m=1,2,\ldots,M$ are coherent across the sensors.

By processing the data by blocks of L time samples, the signal model given in (23) can be put into a vector form as

$$\mathbf{y}_m(t)=\mathbf{x}_m(t)+\mathbf{v}_m(t), m=1,2,\ldots,M, \tag{24}$$

where

$$\mathbf{y}_m(t)=\left[y_m(t)\ y_m(t-1)\ \cdots\ y_m(t-L+1)\right]^T \tag{25}$$

is a vector of length L, and $\mathbf{x}_m(t)$ and $\mathbf{v}_m(t)$ are defined similarly to $\mathbf{y}_m(t)$. It is more convenient to concatenate the M vectors $\mathbf{y}_m(t)$ together as

$$\begin{aligned}\underline{\mathbf{y}}(t)&=\begin{bmatrix}\mathbf{y}_1^T(t) & \mathbf{y}_2^T(t) & \cdots & \mathbf{y}_M^T(t)\end{bmatrix}^T\\&=\underline{\mathbf{x}}(t)+\underline{\mathbf{v}}(t),\end{aligned} \tag{26}$$

where the vectors $\underline{\mathbf{x}}(t)$ and $\underline{\mathbf{v}}(t)$ of length ML are defined in a similar way to $\underline{\mathbf{y}}(t)$. Since $x_m(t)$ and $v_m(t)$ are uncorrelated by assumption, the correlation matrix (of size $ML\times ML$) of the microphone signals is

$$\mathbf{R}_{\underline{\mathbf{y}}}=E\left[\underline{\mathbf{y}}(t)\underline{\mathbf{y}}^T(t)\right]=\mathbf{R}_{\underline{\mathbf{x}}}+\mathbf{R}_{\underline{\mathbf{v}}}, \tag{27}$$

where $\mathbf{R}_{\underline{\mathbf{x}}}$ and $\mathbf{R}_{\underline{\mathbf{v}}}$ are the correlation matrices of $\underline{\mathbf{x}}(t)$ and $\underline{\mathbf{v}}(t)$, respectively (similar to the previous section, we will not consider the time dependency of the signal statistics for the simplicity of notation).

The objective of noise reduction in this section is to estimate $\mathbf{x}_1(t)$ given the noisy signal vector, $\underline{\mathbf{y}}(t)$.

B. Linear Estimation and Performance Measures

In the time domain and with multiple microphones, the desired signal vector, $\mathbf{x}_1(t)$, can be estimated by applying a linear transformation to $\underline{\mathbf{y}}(k)$, i.e.,

$$\mathbf{z}(t) = \underline{\mathbf{H}}\underline{\mathbf{y}}(t) = \mathbf{x}_{\mathrm{fd}}(t) + \mathbf{v}_{\mathrm{rn}}(t), \tag{28}$$

where $\mathbf{z}(t)$ is the estimate of $\mathbf{x}_1(t)$, $\underline{\mathbf{H}}$ is a rectangular filtering matrix of size $L \times ML$, $\mathbf{x}_{\mathrm{fd}}(t) \triangleq \underline{\mathbf{H}}\underline{\mathbf{x}}(t)$ is the filtered desired signal, and $\mathbf{v}_{\mathrm{rn}}(t) \triangleq \underline{\mathbf{H}}\underline{\mathbf{v}}(t)$ is the residual noise. The correlation matrix of $\mathbf{z}(t)$ is then

$$\mathbf{R}_{\mathbf{z}} = \mathbf{R}_{\mathbf{x}_{\mathrm{fd}}} + \mathbf{R}_{\mathbf{v}_{\mathrm{rn}}}, \tag{29}$$

where $\mathbf{R}_{\mathbf{x}_{\mathrm{fd}}} = \underline{\mathbf{H}}\mathbf{R}_{\underline{\mathbf{x}}}\underline{\mathbf{H}}^T$ and $\mathbf{R}_{\mathbf{v}_{\mathrm{rn}}} = \underline{\mathbf{H}}\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{H}}^T$.

By choosing microphone 1 as the reference, the input SNR is given by

$$\mathrm{iSNR} \triangleq \frac{\mathrm{tr}\left(\mathbf{R}_{\mathbf{x}_1}\right)}{\mathrm{tr}\left(\mathbf{R}_{\mathbf{v}_1}\right)}, \tag{30}$$

where $\mathbf{R}_{\mathbf{x}_1}$ and $\mathbf{R}_{\mathbf{v}_1}$ are the correlation matrices of $\mathbf{x}_1(t)$ and $\mathbf{v}_1(t)$, respectively. The output SNR is given by

$$\mathrm{oSNR}\left(\underline{\mathbf{H}}\right) \triangleq \frac{\mathrm{tr}\left(\mathbf{R}_{\mathbf{x}_{\mathrm{fd}}}\right)}{\mathrm{tr}\left(\mathbf{R}_{\mathbf{v}_{\mathrm{rn}}}\right)} = \frac{\mathrm{tr}\left(\underline{\mathbf{H}}\mathbf{R}_{\underline{\mathbf{x}}}\underline{\mathbf{H}}^T\right)}{\mathrm{tr}\left(\underline{\mathbf{H}}\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{H}}^T\right)}. \tag{31}$$

The distortion-based MSE is defined as

$$J\left(\underline{\mathbf{H}}\right) \triangleq E\left\{\left[\mathbf{x}_{\mathrm{fd}}(t) - \mathbf{x}_1(t)\right]^T\left[\mathbf{x}_{\mathrm{fd}}(t) - \mathbf{x}_1(t)\right]\right\}. \tag{32}$$

Hence, the speech distortion index is

$$\upsilon\left(\underline{\mathbf{H}}\right) = \frac{J\left(\underline{\mathbf{H}}\right)}{\mathrm{tr}\left(\mathbf{R}_{\mathbf{x}_1}\right)}. \tag{33}$$

C. Maximum SNR Filter

It is clear from Section II that

$$\underline{\mathbf{H}}_{\max} = \begin{bmatrix} \beta_1\underline{\mathbf{b}}_1^T \\ \beta_2\underline{\mathbf{b}}_1^T \\ \vdots \\ \beta_L\underline{\mathbf{b}}_1^T \end{bmatrix}, \tag{34}$$

where $\beta_l, l = 1, 2, \ldots, L$, are arbitrary real numbers with at least one of them different from 0 and $\underline{\mathbf{b}}_1$ is the eigenvector corresponding to the maximum eigenvalue, $\underline{\lambda}_1$, of the matrix $\mathbf{R}_{\underline{\mathbf{v}}}^{-1}\mathbf{R}_{\underline{\mathbf{x}}}$.

Following the same line of derivation in Section II, one can deduce the optimal β_l's that minimize the distortion-based MSE. As a result, the maximum SNR filtering matrix is

$$\underline{\mathbf{H}}_{\max} = \underline{\mathbf{I}}_{\mathrm{i}}\mathbf{R}_{\underline{\mathbf{x}}}\frac{\underline{\mathbf{b}}_1\underline{\mathbf{b}}_1^T}{\underline{\lambda}_1} = \underline{\mathbf{I}}_{\mathrm{i}}\mathbf{R}_{\underline{\mathbf{v}}}\underline{\mathbf{b}}_1\underline{\mathbf{b}}_1^T, \tag{35}$$

where

$$\underline{\mathbf{I}}_{\mathrm{i}} = \begin{bmatrix} \mathbf{I}_L & \mathbf{0}_{L\times(ML-L)} \end{bmatrix} \tag{36}$$

and $\mathbf{I}_L$ is the $L \times L$ identity matrix. The speech distortion index is then

$$\upsilon\left(\underline{\mathbf{H}}_{\max}\right) = 1 - \frac{1}{\mathrm{tr}\left(\mathbf{R}_{\mathbf{x}_1}\right)\underline{\lambda}_1}\sum_{l=1}^{L}\left(\underline{\mathbf{b}}_1^T\mathbf{R}_{\underline{\mathbf{x}}}\underline{\mathbf{i}}_l\right)^2, \tag{37}$$

where $\underline{\mathbf{i}}_l$ is the lth column of the $ML \times ML$ identity matrix, $\mathbf{I}_{ML}$.

IV. Single-Channel Noise Reduction in the STFT Domain

A. Signal Model and Problem Formulation

In the STFT domain, the signal model in (1) can be rewritten as

$$Y(k,n) = X(k,n) + V(k,n), \tag{38}$$

where the zero-mean complex random variables $Y(k,n)$, $X(k,n)$, and $V(k,n)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at frequency bin $k \in \{0, 1, \ldots, K-1\}$ and time frame n. Since $X(k,n)$, and $V(k,n)$ are uncorrelated by assumption, the variance of $Y(k,n)$ is

$$\phi_Y(k,n) \triangleq E\left[\left|Y(k,n)\right|^2\right] = \phi_X(k,n) + \phi_V(k,n), \tag{39}$$

where $\phi_X(k,n)$ and $\phi_V(k,n)$ are, respectively, the variances of $X(k,n)$ and $V(k,n)$ defined similarly to $\phi_Y(k,n)$.

By considering the N most recent successive time frames of the observations, we can put (38) into the following form:

$$\begin{aligned}\mathbf{y}(k,n) &\triangleq \left[Y(k,n)Y(k,n-1)\cdots Y(k,n-N+1)\right]^T \\ &= \mathbf{x}(k,n) + \mathbf{v}(k,n),\end{aligned} \tag{40}$$

where $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ are the clean speech and noise signal vectors defined in a similar way to $\mathbf{y}(k,n)$. The correlation matrix of $\mathbf{y}(k,n)$ is then

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) &\triangleq E\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right] \\ &= \mathbf{\Phi}_{\mathbf{x}}(k,n) + \mathbf{\Phi}_{\mathbf{v}}(k,n),\end{aligned} \tag{41}$$

where the superscript ^{H} is the conjugate-transpose operator, and $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ are the correlation matrices of $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$, respectively. The objective of this section is then to estimate $X(k,n)$ from $\mathbf{y}(k,n)$ with the maximum SNR filter.

B. Linear Estimation and Performance Measures

In the STFT domain, the desired signal, $X(k,n)$, can be estimated by applying a complex FIR filter, $\mathbf{h}(k,n)$ of length N, to the noisy signal vector, $\mathbf{y}(k,n)$, i.e.,

$$\begin{aligned}Z(k,n) &= \mathbf{h}^H(k,n)\mathbf{y}(k,n) \\ &= X_{\mathrm{fd}}(k,n) + V_{\mathrm{rn}}(k,n),\end{aligned} \tag{42}$$

where $Z(k,n)$ is supposed to be the estimate of $X(k,n)$, $X_{\mathrm{fd}}(k,n) \triangleq \mathbf{h}^H(k,n)\mathbf{x}(k,n)$ is the filtered desired signal, and $V_{\mathrm{rn}}(k,n) \triangleq \mathbf{h}^H(k,n)\mathbf{v}(k,n)$ is the residual noise. The variance of $Z(k,n)$ is

$$\phi_Z(k,n) = \phi_{X_{\mathrm{fd}}}(k,n) + \phi_{V_{\mathrm{rn}}}(k,n), \tag{43}$$

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

The subband input SNR at frequency bin k is defined as

$$\mathrm{iSNR}(k,n) \triangleq \frac{\phi_X(k,n)}{\phi_V(k,n)}, \tag{44}$$

while the subband output SNR at frequency bin k is given by

$$\begin{aligned}\mathrm{oSNR}\left[\mathbf{h}(k,n)\right] &\triangleq \frac{\phi_{X_{\mathrm{fd}}}(k,n)}{\phi_{V_{\mathrm{rn}}}(k,n)} \\ &= \frac{\mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{h}(k,n)}{\mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{h}(k,n)}.\end{aligned} \tag{45}$$

The distortion-based MSE at frequency bin k is defined as

$$J\left[\mathbf{h}(k,n)\right] \triangleq E\left\{\left|X(k,n)-\mathbf{h}^H(k,n)\mathbf{x}(k,n)\right|^2\right\}, \quad (46)$$

from which we define the subband speech distortion index at frequency bin k:

$$\upsilon\left[\mathbf{h}(k,n)\right] = \frac{J\left[\mathbf{h}(k,n)\right]}{\phi_X(k,n)}. \quad (47)$$

C. Maximum SNR Filter

Let $\lambda_1(k,n)$ be the maximum eigenvalue of the matrix $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)$. We denote by $\mathbf{b}_1(k,n)$ the eigenvector associated with $\lambda_1(k,n)$. It is obvious that the filter that maximizes the subband output SNR is

$$\mathbf{h}_{\max}(k,n) = \beta(k,n)\mathbf{b}_1(k,n), \quad (48)$$

where $\beta(k,n) \neq 0$ is an arbitrary complex number. We also have

$$\mathrm{oSNR}\left[\mathbf{h}_{\max}(k,n)\right] = \lambda_1(k,n) \geq \mathrm{iSNR}(k,n). \quad (49)$$

The factor $\beta(k,n)$ must be found in such a way that distortion is minimized. The distortion-based MSE can be rewritten as

$$\begin{aligned} J\left[\mathbf{h}(k,n)\right] = \phi_X(k,n) &+ \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{h}(k,n) \\ &- \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_{N,1} \\ &- \mathbf{i}_{N,1}^T\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{h}(k,n), \end{aligned} \quad (50)$$

where $\mathbf{i}_{N,1}$ is the first column of the $N \times N$ identity matrix $\mathbf{I}_N$. Now, substituting (48) into (50), we get

$$\begin{aligned} J\left[\mathbf{h}_{\max}(k,n)\right] = \phi_X(k,n) &+ \left|\beta(k,n)\right|^2\mathbf{b}_1^H(k,n) \\ &\times \mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{b}_1(k,n) \\ &- \beta^*(k,n)\mathbf{b}_1^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_{N,1} \\ &- \beta(k,n)\mathbf{i}_{N,1}^T\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{b}_1(k,n), \end{aligned} \quad (51)$$

where the superscript * is the complex-conjugate operator. Minimizing $J[\mathbf{h}_{\max}(k,n)]$ with respect to $\beta^*(k,n)$, we obtain

$$\begin{aligned} \beta(k,n) &= \frac{\mathbf{b}_1^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_{N,1}}{\mathbf{b}_1^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{b}_1(k,n)} \\ &= \frac{\mathbf{b}_1^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_{N,1}}{\lambda_1(k,n)}. \end{aligned} \quad (52)$$

Hence, the optimal maximum SNR filter with minimum distortion is

$$\mathbf{h}_{\max}(k,n) = \frac{\mathbf{b}_1^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_{N,1}}{\lambda_1(k,n)}\mathbf{b}_1(k,n). \quad (53)$$

We also find that

$$\upsilon\left[\mathbf{h}_{\max}(k,n)\right] = 1 - \frac{\left|\mathbf{b}_1^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i}_{N,1}\right|^2}{\phi_X(k,n)\lambda_1(k,n)}. \quad (54)$$

V. Multichannel Speech Enhancement in the STFT Domain

A. Signal Model and Problem Formulation

In the STFT domain, the model shown in (23) can be written as

$$\begin{aligned} \underline{\mathbf{y}}(k,n) &= \begin{bmatrix}\mathbf{y}_1^T(k,n) & \mathbf{y}_2^T(k,n) & \cdots & \mathbf{y}_M^T(k,n)\end{bmatrix}^T \\ &= \underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n), \end{aligned} \quad (55)$$

where

$$\begin{aligned} \mathbf{y}_m(k,n) = [&Y_m(k,n) \quad Y_m(k,n-1) \\ &\cdots \quad Y_m(k,n-N+1)]^T, \end{aligned} \quad (56)$$

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

$$\begin{aligned} \mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n) &\triangleq E\left[\underline{\mathbf{y}}(k,n)\underline{\mathbf{y}}^H(k,n)\right] \\ &= \mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n) + \mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n), \end{aligned} \quad (57)$$

where $\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)$ and $\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$ are the correlation matrices of $\underline{\mathbf{x}}(k,n)$ and $\underline{\mathbf{v}}(k,n)$, respectively. The objective of noise reduction in this section is to estimate $X_1(k,n)$ from $\underline{\mathbf{y}}(k,n)$.

B. Linear Estimation and Performance Measures

The desired signal, $X_1(k,n)$, is estimated as follows:

$$\begin{aligned} Z(k,n) &= \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{y}}(k,n) \\ &= X_{\mathrm{fd}}(k,n) + V_{\mathrm{rn}}(k,n), \end{aligned} \quad (58)$$

where $\underline{\mathbf{h}}(k,n)$ is a complex filter of length MN, $X_{\mathrm{fd}}(k,n) \triangleq \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{x}}(k,n)$ is the filtered desired signal and $V_{\mathrm{rn}}(k,n) \triangleq \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{v}}(k,n)$ is the residual noise. We see that the variance of $Z(k,n)$ is

$$\phi_Z(k,n) = \phi_{X_{\mathrm{fd}}}(k,n) + \phi_{V_{\mathrm{rn}}}(k,n), \quad (59)$$

where $\phi_{X_{\mathrm{fd}}}(k,n) = \underline{\mathbf{h}}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{h}}(k,n)$ and $\phi_{V_{\mathrm{rn}}}(k,n) = \underline{\mathbf{h}}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)\underline{\mathbf{h}}(k,n)$.

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

$$\mathrm{iSNR}(k,n) \triangleq \frac{\phi_{X_1}(k,n)}{\phi_{V_1}(k,n)} \quad (60)$$

and

$$\mathrm{oSNR}\left[\underline{\mathbf{h}}(k,n)\right] \triangleq \frac{\underline{\mathbf{h}}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{h}}(k,n)}{\underline{\mathbf{h}}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)\underline{\mathbf{h}}(k,n)}, \quad (61)$$

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.

The subband speech distortion index is

$$\upsilon\left[\underline{\mathbf{h}}(k,n)\right] = \frac{E\left\{\left|X_1(k,n)-\underline{\mathbf{h}}^H(k,n)\underline{\mathbf{x}}(k,n)\right|^2\right\}}{\phi_{X_1}(k,n)}. \quad (62)$$

C. Maximum SNR Filter

Following the same line of derivation given in the previous sections, it can be shown that the maximum SNR filter with minimum distortion is

$$\underline{\mathbf{h}}_{\max}(k,n) = \frac{\underline{\mathbf{b}}_1^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}_{MN,1}}{\underline{\lambda}_1(k,n)}\underline{\mathbf{b}}_1(k,n), \quad (63)$$

where $\underline{\lambda}_1(k,n)$ is the maximum eigenvalue of $\underline{\mathbf{\Phi}}_{\underline{\mathbf{y}}}^{-1}(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)$, $\underline{\mathbf{b}}_1(k,n)$ is the corresponding eigenvector, and $\underline{\mathbf{i}}_{MN,1}$ is the first column of the $MN \times MN$ identity matrix, $\mathbf{I}_{MN}$. We also find that

$$\upsilon\left[\underline{\mathbf{h}}_{\max}(k,n)\right] = 1 - \frac{\left|\underline{\mathbf{b}}_1^H(k,n)\underline{\mathbf{\Phi}}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}_{MN,1}\right|^2}{\phi_{X_1}(k,n)\underline{\lambda}_1(k,n)}. \quad (64)$$

VI. Experiments and Simulations

In the previous sections, we have formulated both the single-channel and multichannel maximum SNR filters for noise reduction in the time and STFT domains. In this section, we study their performance through experiments.

A. Experimental Setup

The clean speech signal used in the single-channel case is recorded in a quiet office room. It is sampled at 8 kHz. The overall length of the signal is approximately 90-s long. The noisy speech is obtained by adding noise to the clean speech (the noise signal is properly scaled to control the input SNR level). We consider three types of noise: a white Gaussian random process, a babble noise signal recorded in a New York Stock Exchange (NYSE) room, and a car noise signal. All the noise signals are sampled at 8 kHz.

The multichannel experiments are conducted with the impulse responses measured in the varechoic chamber at Bell Labs [31]. For a detailed description of the varechoic chamber and how the reverberation time, T_{60}, is controlled, see [31], [32].

The layout of the multichannel experimental setup is illustrated in Fig. 1, where a linear array of 10 omnidirectional microphones is mounted 1.4 m ($z = 1.400$) above the floor and parallel to the north wall at a distance of 0.5 m. The ten microphones are located, respectively, at (x, 5.600, 1.400), where $x = 3.337 : 0.1 : 4.237$. To simulate a sound source, we placed a loudspeaker at (3.337, 4.162, 1.600), playing back a clean speech signal as used in the single-channel case. To make the experiments repeatable, the acoustic channel impulse responses from the source to the ten microphones are first measured (at 48 kHz and then downsampled to 8 kHz) [32]. These measured impulse responses are then regarded as the true ones. During experiments, the microphone outputs are generated by convolving the source signal with the corresponding measured impulse responses and noise is then added to the convolved signals to control the SNR level.

B. Single-Channel Maximum SNR Filter in the Time Domain

To implement the maximum SNR filter derived in Section II-C, we need to know the correlation matrices $\mathbf{R}_{\mathbf{y}}$ and $\mathbf{R}_{\mathbf{v}}$. In this experiment, we compute these matrices directly from the respective signals using a recursive method [19], i.e.,

$$\widehat{\mathbf{R}}_{\mathbf{y}}(t) = \alpha_y\widehat{\mathbf{R}}_{\mathbf{y}}(t-1) + (1-\alpha_y)\mathbf{y}(t)\mathbf{y}^T(t), \quad (65)$$

$$\widehat{\mathbf{R}}_{\mathbf{v}}(t) = \alpha_v\widehat{\mathbf{R}}_{\mathbf{v}}(t-1) + (1-\alpha_v)\mathbf{v}(t)\mathbf{v}^T(t), \quad (66)$$

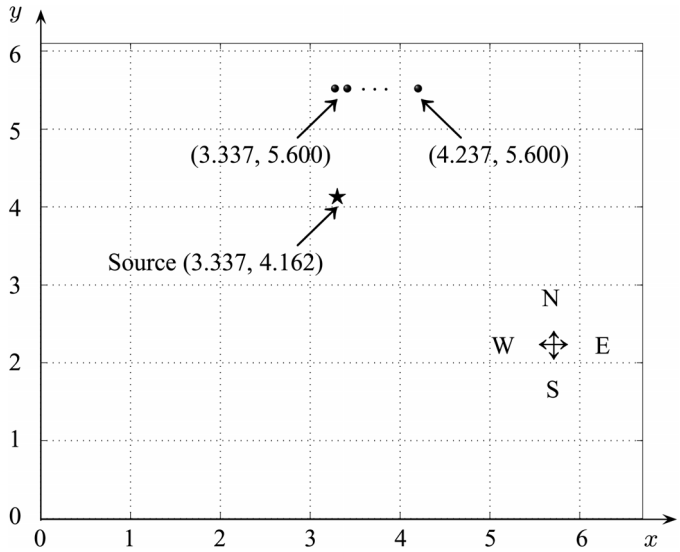


Fig. 1. Layout of the experimental setup in the varechoic chamber (coordinate values measured in meters). The sound source (a loudspeaker) is located at (3.337, 4.162, 1.600). The ten microphones of the linear array are located, respectively, at (x, 5.600, 1.400), where $x = 3.337 : 0.1 : 4.237$.

where $\alpha_y \in (0,1)$ and $\alpha_v \in (0,1)$ are two forgetting factors that control the influence of the previous data samples on the current estimate (the initial estimate is obtained from the first 4000 signal samples with a short-time average). After we obtain the estimated matrices $\widehat{\mathbf{R}}_{\mathbf{y}}(t)$ and $\widehat{\mathbf{R}}_{\mathbf{v}}(t)$, the clean speech signal correlation matrix is then computed as $\widehat{\mathbf{R}}_{\mathbf{x}}(t) = \widehat{\mathbf{R}}_{\mathbf{y}}(t) - \widehat{\mathbf{R}}_{\mathbf{v}}(t)$ [note that in order to ensure that $\widehat{\mathbf{R}}_{\mathbf{x}}(t)$ is positive semidefinite, we apply the eigenvalue decomposition to $\widehat{\mathbf{R}}_x(t)$ and force all the very small eigenvalues to zero]. These estimated correlation matrices are substituted into (21) to implement the maximum SNR filter.

To evaluate the performance of the maximum SNR filter, we adopt three metrics: the output SNR, the speech distortion index [9], and the perceptual evaluation of speech quality (PESQ) [33] (note that many methods can be used to evaluate noise reduction, such as the measures in [35], [34], but we focus on the aforementioned three objective metrics in most experiments of this paper for concise and clear presentation). The former two measures are computed according to (11) and (13), respectively, by replacing the expectation with a long time average, i.e., we first estimate the overall filtered desired signal and residual noise from the 90-s long noisy signal and these estimated signals are used to compute the output SNR and speech distortion index using a long time average. The PESQ score is computed by comparing the 90-s long enhanced signal with the original clean speech.

Fig. 2 plots the experimental results as a function of the forgetting factor (here we assume that $\alpha_y = \alpha_v = \alpha$ for simplicity) for four different filter lengths, i.e., $L = 10$,20,30, and 40. The background noise is white Gaussian, the input SNR is 10 dB, and the block size, P, is equal to 1. It is seen that the output SNR first increases with the forgetting factor and then decreases in all the four different filter-length situations. One can see that the maximum SNR filter can significantly increase the SNR. In comparison, the speech distortion index, υ_{sd}, in the four studied cases increases with the forgetting factor monotonously, i.e., the larger the value of the forgetting factor, the more the speech distortion. Similarly to the output SNR, the PESQ score also first increases with the forgetting factor, but then decreases. It is seen that when the forgetting factor is small, the maximum SNR filter

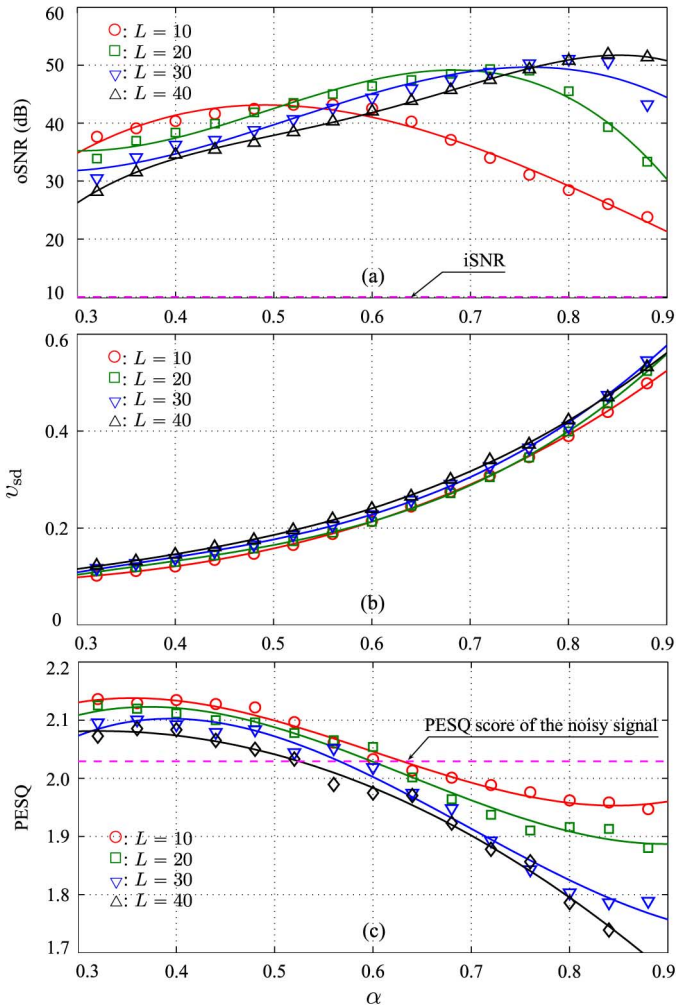


Fig. 2. Performance of the single-channel maximum SNR filter in the time domain as a function of the forgetting factor, for four different filter lengths in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: iSNR $= 10$ dB, $P = 1$, and the PESQ score of the noisy signal is 2.029.

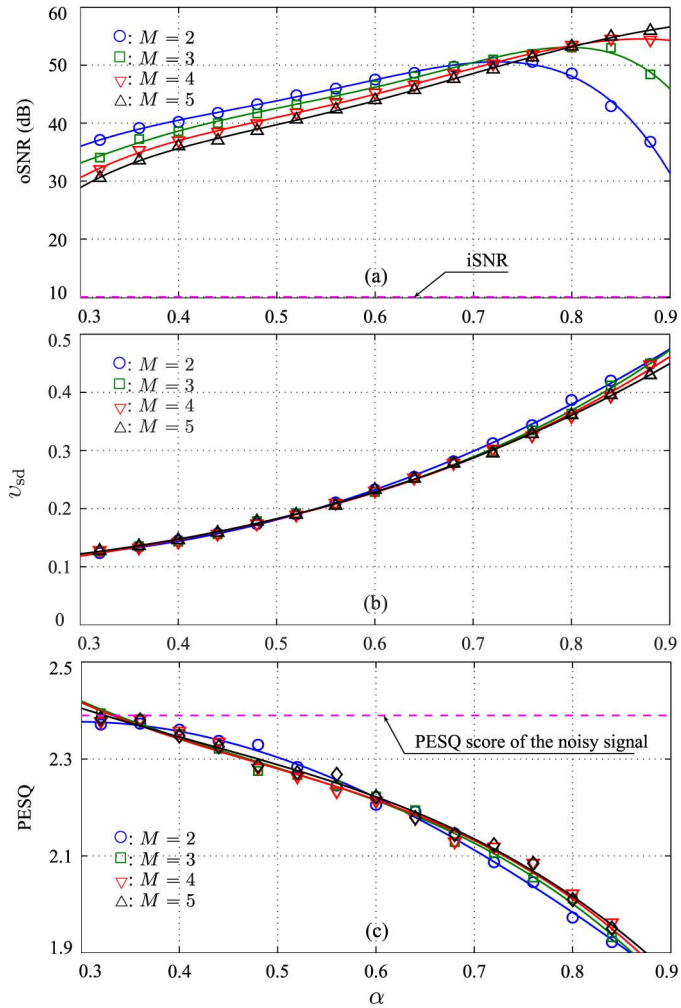


Fig. 3. Performance of the multichannel maximum SNR filter in the time domain as a function of the forgetting factor, α, for four different numbers of microphones in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: iSNR $= 10$ dB, $T_{60} = 240$ ms, $L = 10$, and the PESQ score of the noisy signal is 2.399.

can increase the PESQ, but there is not much gain in PESQ, indicating that the maximum SNR filter does not improve much the speech quality. The underlying reason is that the maximum SNR filter introduces tremendous speech distortion as seen in Fig. 2(b) even though the SNR is significantly improved. These results corroborate with what was observed in the literature of noise reduction [9].

Several other experiments were carried out to assess the performance of the maximum SNR filter given in (21) in different noise and SNR conditions. Similar to the previous experiment, the results showed that this filter can dramatically improve the SNR, but it also introduces a significant amount of speech distortion. As a consequence, the quality improvement is small as indicated by the PESQ score. The results are not reported here for lack of space.

C. Multichannel Maximum SNR Filter in the Time Domain

In this subsection, we study the performance of the multichannel maximum SNR filter given in (35). Similar to the previous experiments, we use a recursive approach to estimate the correlation matrices $\mathbf{R}_{\underline{\mathbf{y}}}$, $\mathbf{R}_{\underline{\mathbf{v}}}$, and $\mathbf{R}_{\underline{\mathbf{x}}}$. Also, we evaluate the noise reduction performance using the output SNR, speech distortion index, and PESQ score as the performance metrics. Note that as shown in Section III, we choose the first microphone as the reference one in the multichannel case. So, all the performance measures are computed using the signals at the first microphone.

Fig. 3 plots the results as a function of the forgetting factor for different numbers of microphones in white Gaussian noise where $L = 10$ and $T_{60} = 240$ ms. It is seen that the maximum SNR filter can dramatically increase the SNR, but at a price of very large speech distortion, regardless of the number of channels. When the forgetting factor is small, the maximum SNR filter can slightly improve the PESQ score; but the gain in PESQ is marginal if any and does not change much with the number of channels.

Several other experiments were conducted to examine the performance of the multichannel maximum SNR filter as a function of the filter length, L, and in different noise and SNR conditions. Similar to the single-channel case, the maximum SNR filter improves significantly the SNR, but the corresponding speech distortion is tremendous at the same time. As a consequence, the quality improvement is marginal if any.

D. Single-Channel Maximum SNR Filter in the STFT Domain

This subsection is concerned with the performance study of the single-channel maximum SNR filter in the STFT domain. To implement this filter of length N, the signals are partitioned into overlapping frames with a frame size of $K = 128$ and an overlapping factor of 75%. 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. The noisy speech spectra is then passed through the maximum SNR filter. Finally, the inverse FFT (with the overlap add technique) is used to obtain the time-domain speech estimate.

To compute the maximum SNR filter, we need to know the correlation matrices $\mathbf{\Phi}_{\mathbf{y}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$. Similar to the previous experiments, these two matrices are estimated from the respective signals using a recursive method [19] (but now the initial estimates are obtained from the first 100 frames with a short-time average), i.e.,

$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) = \alpha_{y,k}\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n-1) + (1-\alpha_{y,k})\mathbf{y}(k,n)\mathbf{y}^H(k,n), \quad (67)$$

$$\widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n) = \alpha_{v,k}\widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n-1) + (1-\alpha_{v,k})\mathbf{v}(k,n)\mathbf{v}^H(k,n), \quad (68)$$

where $\alpha_{y,k} \in (0,1)$ and $\alpha_{v,k} \in (0,1)$ are two forgetting factors. For simplicity, we assume that $\alpha_{y,k} = \alpha_{v,k} = \alpha$. After obtaining the estimates of the correlation matrices $\mathbf{\Phi}_{\mathbf{y}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$, the clean speech correlation matrix is computed as $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,n) = \widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) - \widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n)$.

Again, we assess the performance of the maximum SNR filter using the output SNR, speech distortion index, and PESQ in the time domain, i.e., we first estimate $Z(k,n)$, $X_{\mathrm{fd}}(k,n)$, and $V_{\mathrm{rn}}(k,n)$ in the STFT domain with the maximum SNR filter, and they are then transformed into the time domain to obtain the enhanced and filtered desired signals as well as the residual noise. All performance measures are then computed using a long time average.

In the first experiment, we investigate the impact of the forgetting factor, α, on the performance. The clean speech is the same as the one used in Section VI-B. The background noise is white Gaussian and the input SNR is 10 dB. The results are plotted in Fig. 4. It is seen that the output SNR slightly decreases with α for small values of N while if N is large the output SNR increases with α till it reaches its maximum and then decreases. A similar trend is observed for the PESQ score. The maximum output SNR and the highest PESQ score are achieved at different values of α for different filter lengths. Table I summarizes the value of α that produces the highest PESQ score for different filter lengths. Generally, the larger the filter length, N, the larger is the forgetting factor that achieves the best PESQ score. The underlying reason can be explained as follows. As the filter length increases, the dimension of the correlation matrices becomes larger and, as a result, we would need to use more historic data to achieve a robust matrix estimate. In contrast to the output SNR and PESQ score, the speech distortion index bears a monotonic relationship with the forgetting factor. It is noticed that the value of the speech distortion index of the maximum SNR filter in the STFT domain is much smaller than that of its counterpart in the time domain and, as a result, the STFT-domain maximum SNR filter can more noticeably increase the speech quality as indicated by the PESQ score.

It is noticed from Fig. 4 that the filter length, N, plays a very important role on the noise reduction performance. Fig. 5 plots

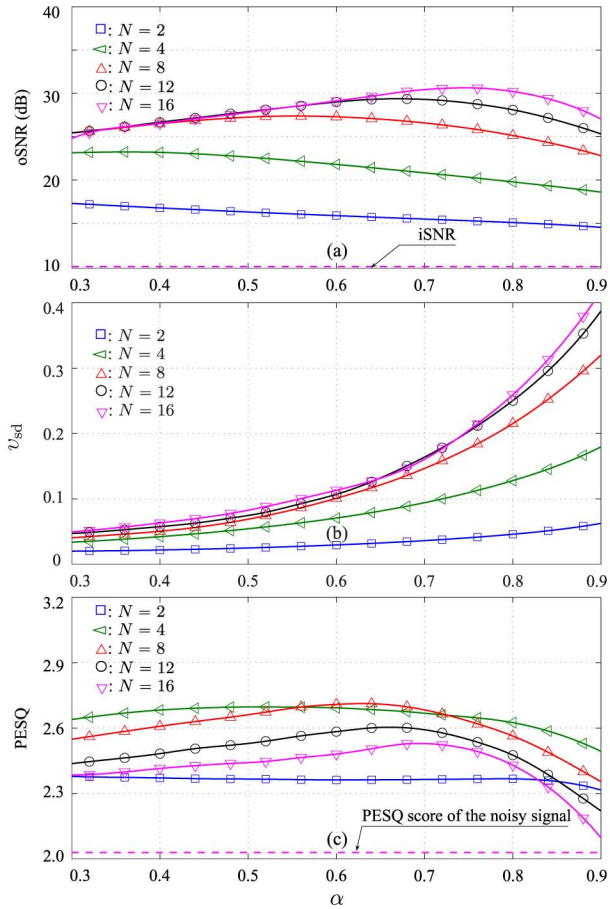


Fig. 4. Performance of the single-channel maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the forgetting factor, α, for five different filter lengths in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: iSNR $= 10$ dB and the PESQ score of the noisy speech is 2.029.

the output SNR, speech distortion index, and PESQ score, all as a function of N, where the experimental conditions are the same as in the previous one. Note that the values of the forgetting factor are chosen according to Table I. It is seen from Fig. 5 that both the output SNR and speech distortion index increase with N. In other words, one can increase the output SNR by using a larger filter length, but the speech distortion index increases at the same time. In contrast, the PESQ score first increases with the filter length and then decreases, as shown in Fig. 5(c). This clearly shows that the quality of the enhanced speech is a tradeoff between noise reduction and speech distortion. When the speech distortion is small, increasing the amount of noise reduction can help improve speech quality. However, when the speech distortion increases to a certain threshold, it will start to be the main factor that degrades speech quality. In our experiment, it is observed that good speech quality is obtained with N in the range between 4 and 8.

We now evaluate the maximum SNR filter (with $N = 4$) in two types of noise and different SNR conditions. For the purpose of comparison, we also compare the performance to that of the MMSE [26] and Wiener filters [9], [18]. Note that for the Wiener and MMSE filters, no interframe information is used, i.e., $N = 1$. The results are plotted in Fig. 6. It is seen that the output SNR is a linear function of the input SNR, while the speech distortion index decrease with the input SNR. It is also

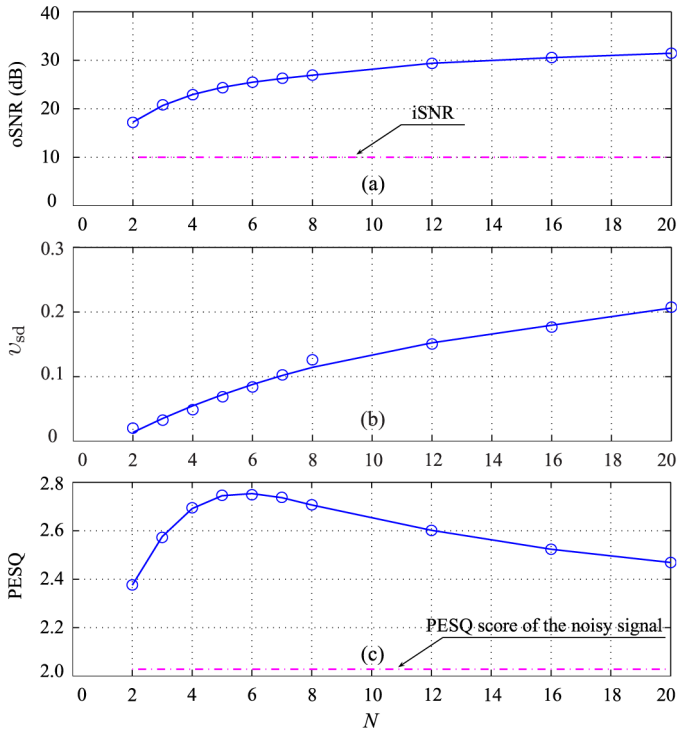


Fig. 5. Performance of the single-channel maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the filter length, N, in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: iSNR $= 10$ dB and the PESQ score of the noisy signal is 2.029.

TABLE I
VALUE OF THE FORGETTING FACTOR CORRESPONDING TO THE HIGHEST PESQ SCORE FOR DIFFERENT FILTER LENGTHS

N	2	3	4	5	6
α	0.32	0.36	0.46	0.54	0.58
N	7	8	12	16	20
α	0.62	0.66	0.68	0.72	0.76

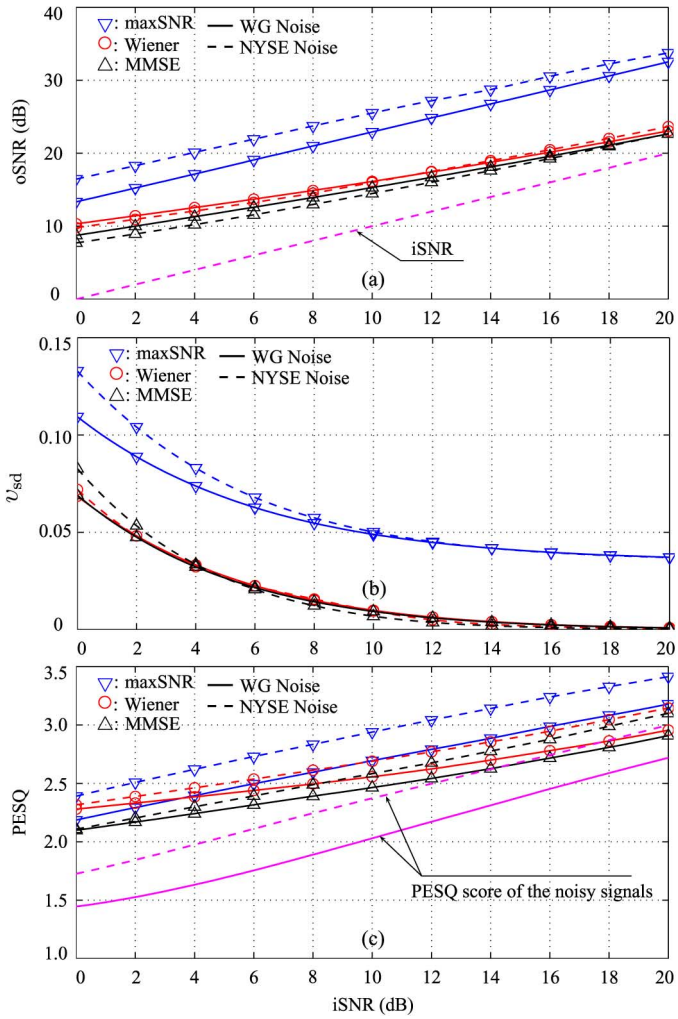


Fig. 6. Performance of the single-channel maximum SNR (with $N = 4$), Wiener, and MMSE filters (window size $K = 128$ with 75% overlap) as a function of the input SNR in different noise conditions: (a) output SNR, (b) speech distortion index, and (c) PESQ score.

observed that the maximum SNR filter has a better performance in the white Gaussian noise. This may be due to the fact that the white Gaussian noise is stationary and is, therefore, easier to deal with.

One can see from Fig. 6 that the maximum SNR filter has achieved a higher PESQ score than both the MMSE and Wiener filters in most cases, especially in the NYSE noise environments when the SNR is low, showing the advantage of the maximum SNR filter.

E. Multichannel Maximum SNR Filter in the STFT Domain

This subsection studies the performance of the multichannel maximum SNR filter given in (63) through experiments. Similar to the single-channel case in the STFT domain, we use a recursive method to estimate the correlation matrices $\mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n)$ and $\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$. Again, we evaluate the noise reduction performance using the output SNR, speech distortion index, and PESQ as the performance metrics, which are computed in the time domain with a long-time average.

As revealed in the previous experiments, the forgetting factor plays an important role on the noise reduction performance. So, in the first set of experiments, we study the impact of the forgetting factor on the performance of the multichannel maximum SNR filter in the STFT domain. The conditions are the following. The background noise is white Gaussian, the reverberation time, T_{60}, is approximately 240 ms, the filter length is set to $N = 2$, and the input SNR is 10 dB. We study three different cases, i.e., $M = 1$, 2, and 4. The results are plotted in Fig. 7. It is seen that when there are multiple microphones ($M \geq 2$), the output SNR and PESQ score first increase with α and then decrease. The maximum output SNR and the highest PESQ score are obtained for different values of α and for different numbers of microphones. Table II presents the value of α that produces the highest PESQ score for different number of microphones. It is seen that the more the microphones, the larger is the forgetting factor to achieve the best PESQ score. It is noticed that increasing the number of microphones can improve the SNR without increasing much additional speech distortion. As a result, the PESQ score is significantly improved as the number of microphones, M, increases. When $M = 1$, the highest PESQ score is approximately 2.8 (for $\alpha = 0.32$). When M is increased to 4, the highest PESQ score is approximately 3.3 (for $\alpha = 0.64$). The difference in PESQ score is 0.5, which is significant. In comparison, the speech distortion index remains almost the same as the number of microphones increases.

To see more clearly the impact of the number of microphones on the noise reduction performance, we show in Fig. 8 the output

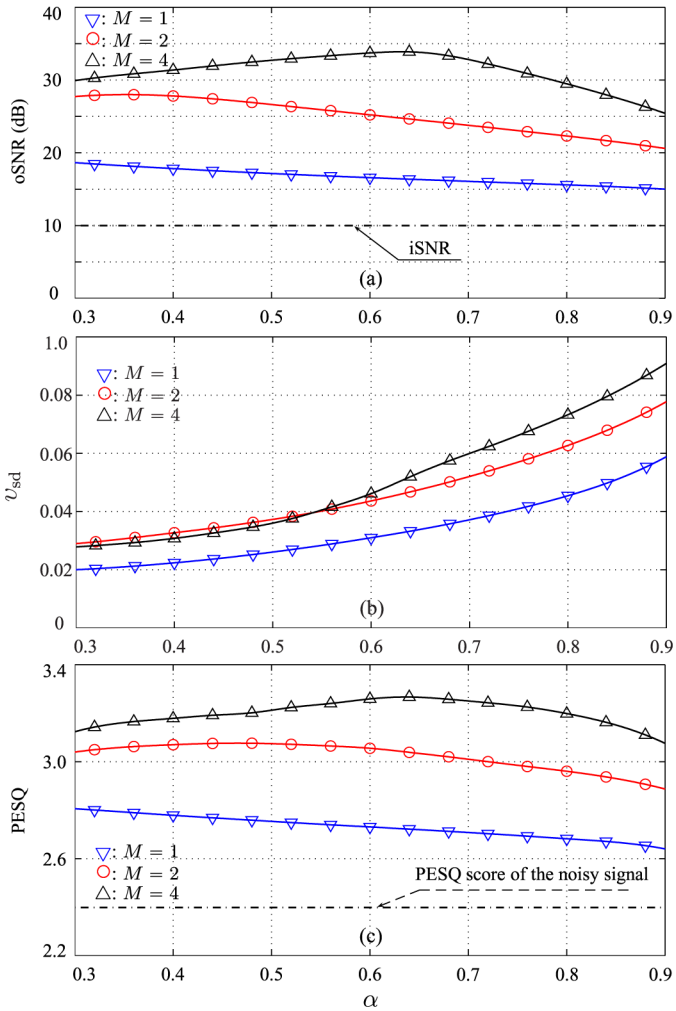


Fig. 7. Performance of the multichannel maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the forgetting factor, α, for three different numbers of microphones in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: iSNR $= 10$ dB, $T_{60} = 240$ ms, $N = 2$, and the PESQ score of the noisy signal is 2.399.

TABLE II
VALUE OF THE FORGETTING FACTOR CORRESPONDING TO THE HIGHEST PESQ SCORE FOR DIFFERENT NUMBERS OF MICROPHONES (THE FILTER LENGTH IS N = 2)

M	1	2	3	4	5
α	0.32	0.48	0.55	0.64	0.67
M	6	7	8	9	10
α	0.74	0.77	0.81	0.82	0.86

SNR, speech distortion index, and PESQ score as a function of the number of microphones, M, in the condition where $N = 2$. It is clearly seen from Fig. 8 that all the three performance metrics increase with M. However, the output SNR increases more dramatically with the number of microphones than the speech distortion index. As a result, we see the PESQ score increases (first quickly and then slowly) with M. With 10 microphones, this maximum SNR filter can increase the PESQ score from approximately 2.4 to more than 3.4, which indicates a significant improvement of speech quality.

Another important factor that affects the noise reduction performance is the filter length, N. In this set of experiments, we

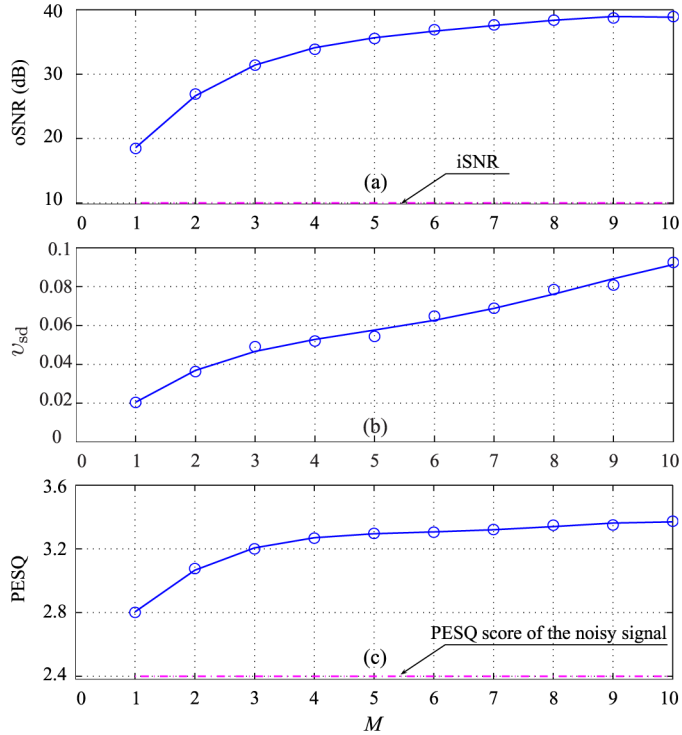


Fig. 8. Performance of the multichannel maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the number of microphones, M, in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: iSNR $= 10$ dB, $T_{60} = 240$ ms, $N = 2$, and the PESQ score of the noisy signal is 2.399.

TABLE III
VALUE OF THE FORGETTING FACTOR CORRESPONDING TO THE HIGHEST PESQ SCORE FOR DIFFERENT FILTER LENGTHS (THE NUMBER OF MICROPHONES IS M = 4)

N	1	2	3	4	5	6
α	0.40	0.64	0.78	0.80	0.84	0.84
N	8	10	12	16	20	
α	0.84	0.84	0.84	0.84	0.84	

choose $M = 4$ and investigate how the noise reduction performance changes with N. We first carried out an experiment to find the optimal values of the forgetting factor for different filter lengths, i.e., for each specified value of the filter length, we vary the forgetting factor in the range between 0 and 1 and check the corresponding noise reduction performance. The factor that produces the highest PESQ score is considered as the optimal value of the forgetting factor for that filter length. The results are summarized in Table III.

Based on the values of the forgetting factor in Table III, experiments were carried out to study the noise reduction performance as a function of the filter length, N. The results are plotted in Fig. 9. It is seen that the output SNR first increases with N and then decreases. In comparison, the speech distortion index monotonously increases with N. So the longer the filter length, the more the speech distortion. When N is small, e.g., $N < 4$, it is seen that the output SNR increases dramatically while the speech distortion index is still small. In this case, the output SNR is more important than the speech distortion index that affects the noise reduction performance. As a result, one can see that the PESQ score increases significantly with N. Therefore, the interframe information is helpful in improving the noise reduction performance. However, when $N \geq 4$, if we keep increasing

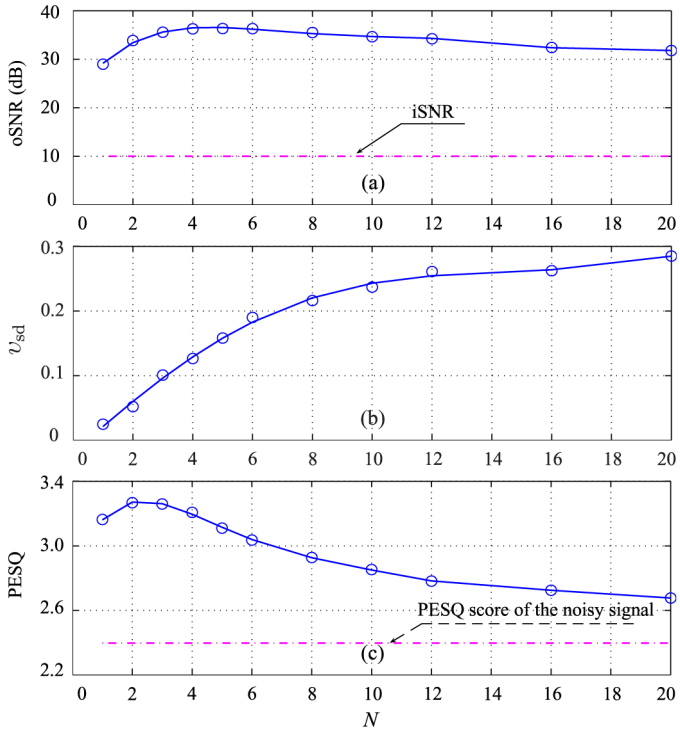


Fig. 9. Performance of the multichannel maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the filter length, N, in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: iSNR $= 10$ dB, $T_{60} = 240$ ms, $M = 4$, and the PESQ score of clean and noisy signal is 2.399.

the filter length, it is seen that the speech distortion index continues to increase while the output SNR starts to drop with N. Consequently, the speech quality starts to degrade with N, as indicated by the PESQ score. This is due to the fact that correlation exists only among neighboring frames while there is not much correlation between far-distance frames.

Experiments were also conducted to evaluate the performance of the multichannel SNR filter in the STFT domain with different input SNRs. Again, the background noise are white Gaussian and car noise, the filter length is $N = 2$, and the forgetting factor is 0.32 and 0.64 for $M = 1$ and 4, respectively. The results are plotted in Fig. 10. In all the studied input SNR conditions, the maximum SNR filter can improve the output SNR and PESQ score significantly.

In this experiment, we examine the performance of the multichannel maximum SNR filter in different reverberation conditions. For the purpose of comparison, the multichannel Wiener filter is also evaluated. The parameters are chosen as $M = 4$, $N = 2$, and $\alpha = 0.64$. The input SNR changes from 0 dB to 20 dB. The results in two reverberation conditions ($T_{60} = 240$ ms and 580 ms) are plotted in Fig. 11. We see that the output SNR is almost the same in different reverberation conditions. In contrast, the speech distortion index increases with reverberation time, which indicates that higher reverberation will lead to more distortion. As a result, the improvement in PESQ score becomes less as reverberation increases as seen Fig. 11(c). This can be easily explained. As the reverberation time becomes longer, it becomes more difficult to predict the signal observed at one microphone from that received at other microphones. Consequently, the speech distortion index increases with the reverberation time while the PESQ gain decreases accordingly.

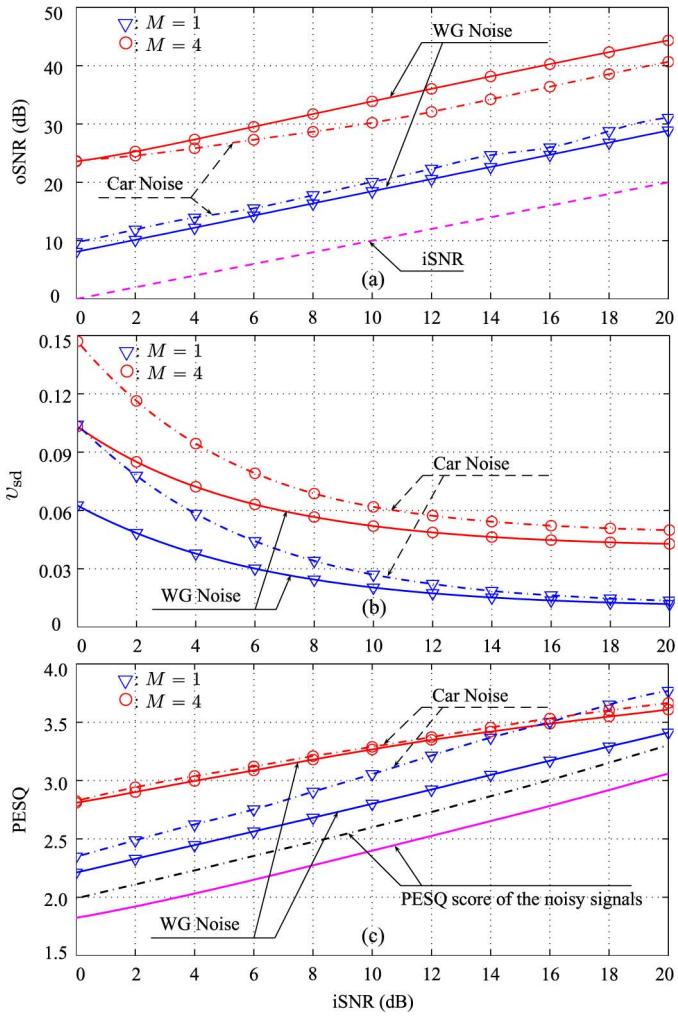


Fig. 10. Performance of the multichannel maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the input SNR with two different numbers of microphones: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: $T_{60} = 240$ ms and $N = 2$.

In comparison with the multichannel Wiener filter, the multichannel maximum SNR filter achieves significantly higher output SNRs; but its speech distortion index is also larger. If the reverberation time is not too long and the input SNR is low, the maximum SNR filter always achieves a higher PESQ improvement. But when the reverberation time is long (e.g., $T_{60} = 580$ ms) or the input SNR is high, the Wiener filter can yield a better PESQ score. This is reasonable since the maximum SNR filter is derived to maximize the output SNR without considering reverberation.

F. Evaluation of the Maximum SNR Filter with POLQA

To further validate the experimental results, we evaluate the maximum SNR filter in this experiment with the Perceptual Objective Listening Quality Assessment (POLQA), which is a new ITU standard (ITU-T Rec. P.863) and a successor of the well-known PESQ (ITU-T Rec. P.862) [35]. The evaluation is performed with the PEXQ software, which is developed by OPTICOM. We consider two situations: the single-channel case with $N = 4$ and the multichannel case with $M = 4$ and $N = 2$. Similar to the previous experiments, in the single-channel case, two types of noise (white Gaussian noise and NYSE noise) are

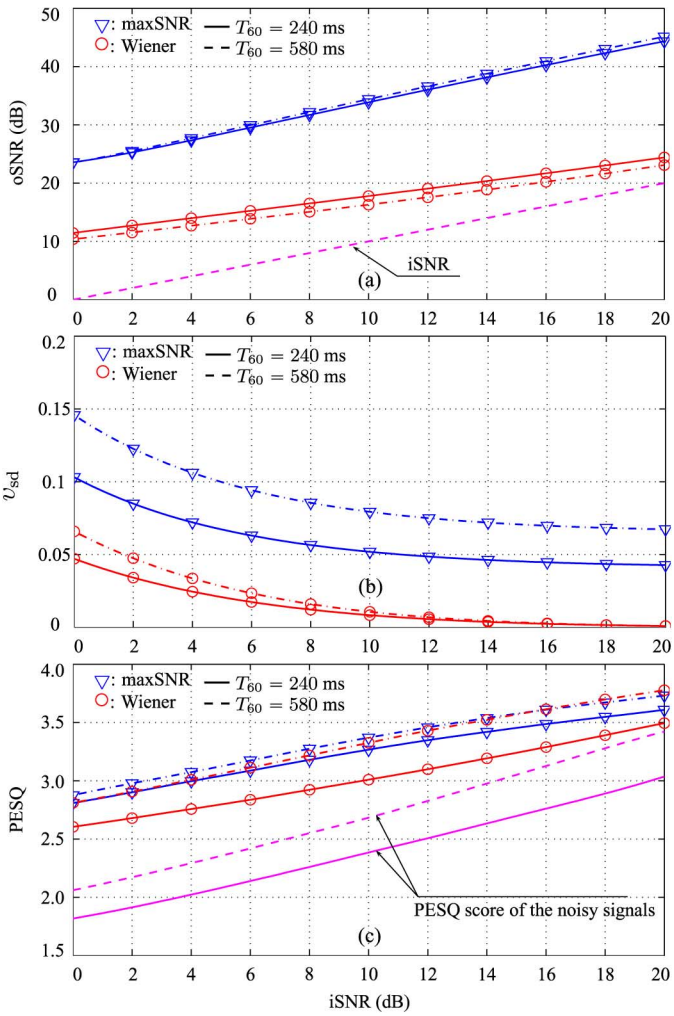


Fig. 11. Performance of the multichannel maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the input SNR with two different reverberation conditions in white Gaussian noise: (a) output SNR, (b) speech distortion index, and (c) PESQ score. Simulation conditions: $M = 4$, $N = 2$.

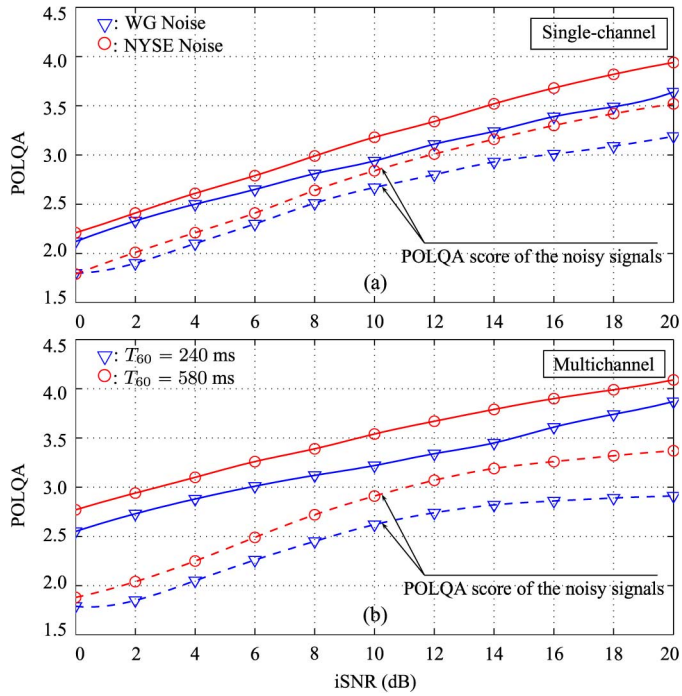


Fig. 12. POLQA Performance of the maximum SNR filter in the STFT domain (window size $K = 128$ with 75% overlap) as a function of the input SNR: (a) single-channel case ($N = 4$) with two different noise conditions, (b) multichannel case ($M = 4$, $N = 2$) with two different reverberation conditions in white Gaussian noise.

used while in the multichannel case, two different reverberation conditions ($T_{60} = 240$ ms and $T_{60} = 580$ ms) are tested. The results are plotted in Fig. 12. It is seen that the maximum SNR filter improves the POLQA score significantly in both the studied single-channel and multichannel cases. In comparison with the single-channel case, the multichannel one has a higher POLQA score, which, again, indicates the advantage of using multiple microphones. We also observe that the POLQA gain with the multichannel maximum SNR filter is slightly higher than that of the PESQ gain; but the difference is not significant.

Before finishing the section, we want to make some remarks on the complexity of the maximum SNR filters in the STFT domain. In the single-channel case, the complexity of the maximum SNR filter at every frequency band consists of three parts: computing the two correlation matrices $\mathbf{\Phi_y}$ and $\mathbf{\Phi_v}$, finding the maximum eigenvalue λ_1 and the eigenvector $\mathbf{b}_1$, and computing the filter $\mathbf{h}_{\max}$. The first part requires $4N^2 + 2N$ multiplications; the complexity of the second one is in the order of N^2 [36]; and the last part requires $N + 2$ multiplications. Therefore, the complexity of the single-channel maximum SNR filter in the STFT domain is in the order of N^2 at every subband or in the order of KN^2 at every frame. For the multichannel maximum SNR filter, the complexity is in the order of MKN^2

VII. Conclusions

Noise reduction is a challenging problem in acoustic signal processing and voice communications. Since one of the major objectives of noise reduction is to reduce the amount of noise, thereby improving the SNR, it is a natural motivation to study the maximum SNR filter. In this paper, we derived and studied a class of the maximum SNR filters including both the single-channel and multichannel ones in the time and STFT domains. A large number of experiments were carried out to examine the performance of the maximum SNR filters in terms of the amount of speech distortion, the gain in SNR, and PESQ and POLQA scores. While it was found that the maximum SNR filters in the time domain, regardless of the number of input channels, introduce significant speech distortion, which limits their effectiveness in improving speech quality, the filters in the STFT domain can significantly improve the SNR and PESQ and POLQA scores. It is also interesting to see that, in the STFT domain, the SNR and PESQ gains increase with the number of input channels. This indicates that the maximum SNR filter in the STFT domain has some great potential in practical environments.

Acknowledgment

We would like to thank the associate editor and four anonymous reviewers for their constructive comments, which helped improve the clarity and quality of this paper. We are also grateful to TRANSCOM International Ltd and OPTICOM for helping evaluate our algorithm with POLQA (ITU-T Rec. P.863).

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

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

[3] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

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

[6] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi, and P. C. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English," *J. Acoust. Soc. Amer.*, vol. 125, pp. 3291–3301, May 2011.

[7] A. Coy and J. Barker, "An automatic speech recognition system based on the scene analysis account of auditory perception," *Speech Commun.*, vol. 49, pp. 384–401, 2007.

[8] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. New York, NY, USA: Springer, 2005, pp. 371–402.

[9] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.

[10] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE ICASSP*, 2011, pp. 45–48.

[11] J. Fritsch and M. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE ICASSP*, 2013, pp. 888–891.

[12] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[13] W. Herbordt, H. Buchner, S. Nakamura, and W. Kellermann, "Multichannel bin-wise robust frequency-domain adaptive filtering and its application to adaptive beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1340–1351, May 2007.

[14] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and L1-norm minimization," *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 81–81, Jan. 2007.

[15] J. Benesty and J. Chen, *Optimal Time-domain Noise Reduction Filters–A Theoretical Study*. New York, NY, USA: Springer Briefs in Electrical and Computer Engineering, 2011.

[16] 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.

[17] J. I. Marin-Hurtado, D. N. Parikh, and D. V. Anderson, "Perceptually inspired noise-reduction method for binaural hearing aids," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1372–1382, May 2012.

[18] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. New York, NY, USA: Springer Briefs in Electrical and Computer Engineering, 2012.

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

[20] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

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

[22] R. M. Nickel, R. F. Astudillo, D. Kolossa, and R. Martin, "Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 983–997, May 2013.

[23] S. Markovich-Golan, S. Gannot, and I. Cohen, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 21, no. 7, pp. 1513–1523, Jul. 2013.

[24] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[25] W. Charoenruengkit and N. Erdöl, "The effect of spectral estimation on speech enhancement performance," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1170–1179, Jul. 2011.

[26] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[27] M. McCallum and B. Guillemin, "Stochastic-deterministic MMSE STFT speech enhancement with general a priori information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1445–1457, Jul. 2013.

[28] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[29] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. IEEE ICASSP*, 2001, pp. 496–499.

[30] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook on Speech Processing and Speech Communication*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 843–871.

[31] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald, "The new Varechoic chamber at AT&T Bell Labs," in *Proc. Wallance Clement Sabine Centennial Symp.*, 1994.

[32] A. Härmä, "Acoustic measurement data from the varechoic chamber," *Tech. Memo. , Agere Syst.*, Nov. 2001.

[33] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[35] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement (Part I and II)," *J. Audio Eng. Soc.*, vol. 61, pp. 366–384, Jun. 2013.

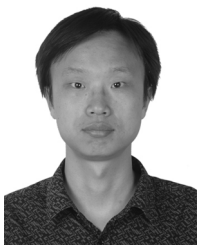
[36] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, Jul. 2007.

Gongping Huang (S'13) received the bachelor degree in electronic information engineering from the Northwestern Polytechnical University in 2012. He is currently a Ph.D. student in communication engineering at the Center of Immersive and Intelligent Acoustics, Northwestern Polytechnical University. His research interests include noise reduction, speech enhancement, and microphone array and audio signal processing.

Jacob Benesty was born in 1963. He received a master degree in microwaves from Pierre & Marie Curie University, France, in 1987, and a Ph.D. degree in control and signal processing from Orsay University, France, in April 1991. During his Ph.D. (from Nov. 1989 to Apr. 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecomunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ, USA. In May 2003, he joined the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, as a Professor. He is also a Visiting Professor at the Technion, in Haifa, Israel, and an Adjunct Professor at Aalborg University, in Denmark and at Northwestern Polytechnical University, in Xi'an, Shaanxi, China.

His research interests are in signal processing, acoustic signal processing, and multimedia communications. He is the inventor of many important technologies. In particular, he was the lead researcher at Bell Labs who conceived and designed the world-first real-time hands-free full-duplex stereophonic teleconferencing system. Also, he conceived and designed the world-first PC-based multi-party hands-free full-duplex stereo conferencing system over IP networks.

He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control and the general co-chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. He is the recipient, with Morgan and Sondhi, of the IEEE Signal Processing Society 2001 Best Paper Award. He is the recipient, with Chen, Huang, and Doclo, of the IEEE Signal Processing Society 2008 Best Paper Award. He is also the co-author of a paper for which Huang received the IEEE Signal Processing Society 2002 Young Author Best Paper Award. In 2010, he received the Gheorghe Cartianu Award from the Romanian Academy. In 2011, he received the Best Paper Award from the IEEE WASPAA for a paper that he co-authored with Chen.



Tao Long is currently a biomedical engineering Ph.D. student at Xian Jiaotong University in China. He received the bachelor degree in applying physics from Northwest University in 2006. He was a visitor at Taiwan National Chiao Tung University in 2010 and at Lubeck University in Germany in 2012. His research interests are in noise reduction and image processing.

Jingdong Chen (M'99–SM'09) received the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences in 1998.

From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis, speech analysis, as well as objective measurements for evaluating speech synthesis. He then joined Griffith University, Brisbane, Australia, where he engaged in research on robust speech recognition and signal processing. From 2000 to 2001, he worked at ATR Spoken Language Translation Research Laboratories on robust speech recognition and speech enhancement. From 2001 to 2009, he was a Member of Technical Staff at Bell Laboratories, Murray Hill, New Jersey, working on acoustic signal processing for telecommunications. He subsequently joined WeVoice Inc. in New Jersey, serving as the Chief Scientist. He is currently a professor at the Northwestern Polytechnical University in Xi'an, China. His research interests include acoustic signal processing, adaptive signal processing, speech enhancement, adaptive noise/echo control, microphone array signal processing, signal separation, and speech communication.

Dr. Chen was an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING from 2007 to 2014. He is currently a member of the IEEE Audio and Electroacoustics Technical Committee, and a member of the editorial advisory board of the *Open Signal Processing Journal*. He was the Technical Program Co-Chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) and IEEE ChinaSIP 2014, the Technical Program Chair of IEEE TENCON 2013, and helped organize many other conferences. He co-authored the books *Study and Design of Differential Microphone Arrays* (Springer-Verlag, 2013), *Speech Enhancement in the STFT Domain* (Springer-Verlag, 2011), *Optimal Time-Domain Noise Reduction Filters: A Theoretical Study* (Springer-Verlag, 2011), *Speech Enhancement in the Karhunen-Loève Expansion Domain* (Morgan&Claypool, 2011), *Noise Reduction in Speech Processing* (Springer-Verlag, 2009), *Microphone Array Signal Processing* (Springer-Verlag, 2008), and *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006). He is also a co-editor/co-author of the book *Speech Enhancement* (Berlin, Germany: Springer-Verlag, 2005).

Dr. Chen received the 2008 Best Paper Award from the IEEE Signal Processing Society (with Benesty, Huang, and Doclo), the best paper award from the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2011 (with Benesty), the Bell Labs Role Model Teamwork Award twice, respectively, in 2009 and 2007, the NASA Tech Brief Award twice, respectively, in 2010 and 2009, the Japan Trust International Research Grant from the Japan Key Technology Center in 1998, and the Young Author Best Paper Award from the 5th National Conference on Man-Machine Speech Communications in 1998.