Binaural Noise Reduction in the Time Domain With a Stereo Setup

Jacob Benesty, Jingdong Chen, *Senior Member, IEEE*, and Yiteng Huang, *Member, IEEE*

***Abstract*—Binaural noise reduction with a stereophonic (or simply stereo) setup has become a very important problem as stereo sound systems and devices are being more and more deployed in modern voice communications. This problem is very challenging since it requires not only the reduction of the noise at the stereo inputs, but also the preservation of the spatial information embodied in the two channels so that after noise reduction the listener can still localize the sound source from the binaural outputs. As a result, simply applying a traditional single-channel noise reduction technique to each channel individually may not work as the spatial effects may be destroyed. In this paper, we present a new formulation of the binaural noise reduction problem in stereo systems. We first form a complex signal from the stereo inputs with one channel being its real part and the other being its imaginary part. By doing so, the binaural noise reduction problem can be processed by a single-channel widely linear filter. The widely linear estimation theory is then used to derive optimal noise reduction filters that can fully take advantage of the noncircularity of the complex speech signal to achieve noise reduction while preserving the desired signal (speech) and spatial information. With this new formulation, the Wiener, minimum variance distortionless response (MVDR), maximum signal-to-noise ratio (SNR), and tradeoff filters are derived. Experiments are provided to justify the effectiveness of these filters.**

***Index Terms*—Binaural noise reduction, maximum signal-to-noise ratio (SNR) filter, minimum variance distortionless response (MVDR) filter, noncircularity, speech enhancement, stereo sound system, time domain, tradeoff filter, widely linear estimation, Wiener filter.**

I. Introduction

TELECONFERENCING capability is an integral part of modern communication networks. It facilitates group collaborations (i.e., business, military, government, and educational institutions) efficiently and at low costs. A key technical challenge for a teleconferencing (or more generally a telecollaboration) system is the ability to acquire high-fidelity speech while keeping speakers' spatial information intact so that it is possible for the remote listener to follow a panel of speakers and distinguish them by listening to the reproduction of the signals. To preserve the sound realism, it is necessary to use multiple loudspeakers and multiple microphones in both the transmission and receiving ends. The system in each end then becomes a multiple-input multiple-output (MIMO) one that is more complicated than a single-channel system in terms of signal acquisition, processing, and reproduction. A particular case of such a MIMO system is the stereophonic (or simply stereo) one, which uses two microphones and two loudspeakers. A stereo system has many unique advantages. It can record and reproduce a speech sound with spatial information and offer much more flexibility in processing the sound as compared to a single-channel system; yet it is cheaper and less complicated to design and easier to be integrated into portable devices in comparison with a general MIMO system with a large number of sensors. As a result, stereo systems are being more and more deployed in communication terminals such as smart phones, desktop phones, PCs, etc.

With a stereo setup, many signal processing problems such as echo cancellation and noise reduction are fundamentally different from those in the single-channel case. As a result, simply applying a single-channel processing technique to each of the stereo channels does not, in general, result in satisfactory performance. Therefore, these problems need to be re-addressed. In this paper, we focus on the issue of noise reduction in stereo systems. The basic problem is to process the stereo input signals such as to mitigate the noise effect, thereby producing two (binaural) outputs with less amount of noise and a higher signal-to-noise ratio (SNR), but the mitigation process is required not to add audible distortion to the desired speech signal at the two channels (this is the same as in the singe-channel case) and meanwhile the spatial information of the desired sound source should be preserved so that, after noise reduction, the remote listener will still be able to localize the sound source thanks to his/her binaural hearing mechanism. This problem is generally referred to as binaural noise reduction.

Since a stereo system with two microphones can be viewed as a particular case of the general problem of microphone arrays [1], [2], one straightforward approach to binaural noise reduction is through the adoption of beamforming techniques. Notice, however, that a beamformer only gives one monaural output. In order to have binaural outputs, we would need to use two beamformers at the same time with one beamformer generating an output for the left channel while the other producing an output for the right channel [3]–[9]. Different constraints can be applied to preserve the spatial effect. One simple way, for example, is to force the time difference of arrival (TDOA) between the two beamformers' outputs to be the same as that of the desired signals at the two stereo input channels. We should

Manuscript received February, 2011; revised January 05, 2011; accepted February 14, 2011. Date of publication February 24, 2011; date of current version August 19, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. James Johnston.

J. Benesty is with INRS-EMT, University of Quebec, Montreal QC H5A 1K6, Canada.

J. Chen is with Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China (e-mail: jingdongchen@ieee.org).

Y. Huang is with WeVoice, Inc., Bridgewater, NJ 08807 USA.

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/TASL.2011.2119313

note that beamforming generally aims at recovering the source signal. Although it is feasible to use such technique for binaural noise reduction, the amount of noise attenuation or SNR improvement of beamforming with the use of only two microphones is generally limited, particularly in teleconferencing applications where background noise can be strong, environments are reverberant, and the source can be far away from the microphones. To gain more noise reduction, a viable approach is to extend the ideas developed for the single-channel noise reduction to the binaural case. An early attempt on this was made in the hearing-aids area [10]. The basic idea in [10] is similar to the widely known spectral subtraction [11] or parametric Wiener filter [12]–[14]; but it poses a constraint on the suppression of each frequency band to preserve the spatial information of the desired sound source. If the band has the interaural time and level differences characteristic of the desired source, this band is kept unchanged (passed through without attenuation). Otherwise, the band is suppressed. This method was refined in [15] and then extended to a Wiener filter framework with the use of head-related transfer functions (HRTFs) for noise estimation [16]. While it can possibly obtain more noise reduction than beamforming, this second approach generally adds distortion to the desired speech. Furthermore, it requires the *a priori* knowledge about the interaural time and level differences of the source signal, which is not easy to acquire in conferencing applications. A more practical approach to preserving sound realism while performing noise reduction is, perhaps, through generalizing the principles of multichannel noise reduction such as the transfer function based generalized sidelobe canceller (TF-GSC) [17], the multichannel Wiener filter [18], and the spatial prediction method [19] to the binaural case. Since the multichannel noise reduction techniques are formulated to estimate the desired signals observed at the microphones, the spatial information should be naturally preserved. Indeed, it has been shown in [20] and [21] that a binaural multichannel Wiener filter is able to protect the interaural time difference (ITD) cues while achieving noise reduction [20], [21] though a tradeoff between noise reduction and preservation of the binaural cues seems inevitable.

In this paper, we take a different approach to binaural noise reduction in stereo systems. We first combine the stereo inputs together to form a complex signal: its real part corresponds to the stereo's left-channel signal and its imaginary part corresponds to the stereo's right-channel signal. By doing so, the binaural noise reduction problem is converted to one of complex linear filtering. We then apply the so-called widely linear estimation theory to derive optimal binaural noise reduction filters. These filters have the potential to fully take advantage of the noncircularity of the complex speech signal in order to achieve noise reduction while controlling the speech distortion and preserving the spatial information. With this new formulation, various noise reduction filters can be derived. We will show how to derive the Wiener, minimum variance distortionless response (MVDR), maximum SNR, and tradeoff filters. Experimental results will be provided to illustrate the effectiveness of these filters. Note that the optimal noise reduction filters derived in this paper may also be obtained through the use of a multichannel noise reduction framework with a proper formulation, but the approach taken here provides a more convenient way to define the objective functions and performance measures. Furthermore, it makes the derivation of the different filters and their evaluation much easier.

The rest of this paper is organized as follows. In Section II, we formulate the binaural noise reduction problem in stereo systems. We then briefly review the widely linear estimation theory and show how this theory can be applied to the binaural noise reduction problem in Section III. Section IV presents some performance measures that can be used to evaluate binaural noise reduction. In Section V, we discuss different noise reduction filters. Section VI presents some experiments to validate the theoretical derivations. Finally, we give our conclusions in Section VII.

II. Signal Model

In this paper, we consider the signal model in which two microphones (that we refer to as right and left) capture a source signal convolved with acoustic impulse responses in some noise field. The signals received at the right and left microphones, at the discrete time-index k, are then expressed as

$$y_{\mathrm{R}}(k) = g_{\mathrm{R}}(k) * s(k) + v_{\mathrm{R}}(k) = x_{\mathrm{R}}(k) + v_{\mathrm{R}}(k) \quad (1a)$$

$$y_{\mathrm{L}}(k) = g_{\mathrm{L}}(k) * s(k) + v_{\mathrm{L}}(k) = x_{\mathrm{L}}(k) + v_{\mathrm{L}}(k) \quad (1b)$$

where $g_{\mathrm{R}}(k)$ [resp. $g_{\mathrm{L}}(k)$] is the impulse response from the unknown speech source $s(k)$ to the microphone on the right (resp. left), $*$ stands for linear convolution, and $v_{\mathrm{R}}(k)$ [resp. $v_{\mathrm{L}}(k)$] is the additive noise at the microphone on the right (resp. left). We assume that all the signals $x_{\mathrm{R}}(k)$, $x_{\mathrm{L}}(k)$, $v_{\mathrm{R}}(k)$, and $v_{\mathrm{L}}(k)$ are zero mean and $x_{\mathrm{R}}(k)$ and $x_{\mathrm{L}}(k)$ are uncorrelated with $v_{\mathrm{R}}(k)$ and $v_{\mathrm{L}}(k)$. The two noise signals $v_{\mathrm{R}}(k)$ and $v_{\mathrm{L}}(k)$ can be either uncorrelated or correlated (e.g., from a same point source), but they are assumed to be nonspeech and more stationary than speech so that their statistics can be estimated with the help of a voice activity detector (VAD). This model can be seen as a particular case of the general problem of microphone arrays [1], [2], [22].

In this paper, we consider the problem of recovering the signals $x_{\mathrm{R}}(k)$ and $x_{\mathrm{L}}(k)$ given the observations $y_{\mathrm{R}}(k)$ and $y_{\mathrm{L}}(k)$. This means that the desired signals in our problem are the speech signals received at the right and left microphones. It is clear then that we have two objectives. The first one is to attenuate the contribution of the noise terms $v_{\mathrm{R}}(k)$ and $v_{\mathrm{L}}(k)$ as much as possible. The second objective is to preserve $x_{\mathrm{R}}(k)$ and $x_{\mathrm{L}}(k)$ with their spatial information, so that with the enhanced signals, along with our binaural hearing process, we will still be able to localize the source $s(k)$.

We have stereo signals in model (1); but we believe that it is more convenient to work in the complex domain in order that the original (stereo) problem is transformed to a single-channel noise reduction processing. Indeed, from the two real microphone signals given in (1a) and (1b), we can form the complex microphone signal as

$$y(k) = y_{\mathrm{R}}(k) + jy_{\mathrm{L}}(k) = x(k) + v(k) \quad (2)$$

where $j = \sqrt{-1}$, $x(k) = x_{\mathrm{R}}(k) + jx_{\mathrm{L}}(k)$ is the complex desired signal, and $v(k) = v_{\mathrm{R}}(k) + jv_{\mathrm{L}}(k)$ is the complex additive noise. Now, our problem may be stated as follows: given

the complex microphone signal, $y(k)$, which is a mixture of two uncorrelated complex signals $x(k)$ and $v(k)$, our goal is to preserve $x(k)$ (i.e., our desired signal) while minimizing $v(k)$.

The signal model given in (2) can be put into a vector form if we accumulate L successive samples

$$\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k) \tag{3}$$

where

$$\mathbf{y}(k) \triangleq [\, y(k) \quad y(k-1) \quad \cdots \quad y(k-L+1) \,]^T \tag{4}$$

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

$$\mathbf{R}_{\mathbf{y}} \triangleq E\left[\mathbf{y}(k)\mathbf{y}^H(k)\right] = \mathbf{R}_{\mathbf{x}} + \mathbf{R}_{\mathbf{v}} \tag{5}$$

where $E[\cdot]$ denotes mathematical expectation, the superscript ^{H} denotes the conjugate-transpose operator, and $\mathbf{R}_{\mathbf{x}} \triangleq E\left[\mathbf{x}(k)\mathbf{x}^H(k)\right]$ and $\mathbf{R}_{\mathbf{v}} \triangleq E\left[\mathbf{v}(k)\mathbf{v}^H(k)\right]$ are the correlation matrices of $\mathbf{x}(k)$ and $\mathbf{v}(k)$, respectively.

III. Widely Linear Model

As can be noticed from the model given in (2), we deal with complex random variables. A very important statistical characteristic of a complex random variable (CRV) is the so-called circularity property or lack of it (noncircularity) [23], [24]. A zero-mean CRV, z, is circular if and only if the only nonnull moments and cumulants are the moments and cumulants constructed with the same power in z and z^*[25], [26], where the superscript * denotes complex conjugation. In particular, z is said to be a second-order circular CRV (CCRV) if its so-called pseudo-variance [23] is equal to zero, i.e., $E\left(z^2\right) = 0$, while its variance is nonnull, i.e., $E\left(|z|^2\right) \neq 0$. This means that the second-order behavior of a CCRV is well described by its variance. If the pseudo-variance $E\left(z^2\right)$ is not equal to 0, the CRV z is then noncircular. A good measure of the second-order circularity is the circularity quotient [23] defined as the ratio between the pseudo-variance and the variance, i.e.,

$$\gamma_z = \frac{E\left(z^2\right)}{E\left(|z|^2\right)}. \tag{6}$$

It is easy to show that $0 \leq |\gamma_z| \leq 1$. If $\gamma_z = 0$, z is a second-order CCRV; otherwise, z is noncircular, and a larger value of $|\gamma_z|$ indicates that the CRV z is more noncircular.

Now, let us examine whether the complex desired signal, $x(k)$, is second-order circular or not. We have

$$\gamma_x = \frac{E\left[x^2(k)\right]}{E[|x(k)|^2]} = \frac{E\left[x_{\mathrm{R}}^2(k)\right] - E\left[x_{\mathrm{L}}^2(k)\right] + 2jE[x_{\mathrm{R}}(k)x_{\mathrm{L}}(k)]}{\sigma_x^2} \tag{7}$$

where $\sigma_x^2 = E\left[|x(k)|^2\right]$ is the variance of $x(k)$. One can check from (7) that the CRV $x(k)$ is second-order circular (i.e., $\gamma_x = 0$) if and only if

$$E\left[x_{\mathrm{R}}^2(k)\right] = E\left[x_{\mathrm{L}}^2(k)\right] \quad \text{and} \quad E\left[x_{\mathrm{R}}(k)x_{\mathrm{L}}(k)\right] = 0. \tag{8}$$

Since the signals $x_{\mathrm{R}}(k)$ and $x_{\mathrm{L}}(k)$ come from the same source, they are in general correlated. As a result, the second condition in (8) should not be true. Therefore, we can safely state that the complex desired signal, $x(k)$, is noncircular, and so is the complex microphone signal, $y(k)$. If we assume that the noise terms at the two microphones are uncorrelated and have the same power then $\gamma_v = 0$ [i.e., $v(k)$ is a second-order CCRV].

Since we deal with noncircular CRVs as demonstrated above, the classical linear estimation technique [27], [28], which is developed for processing real signals or CCRVs, cannot be applied. Instead, an estimate of $x(k)$ should be obtained using the widely linear (WL) estimation theory as [24], [29]

$$\hat{x}(k) = \mathbf{h}^H\mathbf{y}(k) + \mathbf{h}'^H\mathbf{y}^*(k) = \tilde{\mathbf{h}}^H\tilde{\mathbf{y}}(k) \tag{9}$$

where $\mathbf{h}$ and $\mathbf{h}'$ are two complex finite-impulse-response (FIR) filters of length L and

$$\tilde{\mathbf{h}} \triangleq \begin{bmatrix} \mathbf{h} \\ \mathbf{h}' \end{bmatrix} \tag{10}$$

$$\tilde{\mathbf{y}}(k) \triangleq \begin{bmatrix} \mathbf{y}(k) \\ \mathbf{y}^*(k) \end{bmatrix} \tag{11}$$

are the augmented WL filter and observation vector, respectively, both of length $2L$. We can rewrite (9) as

$$\hat{x}(k) = \tilde{\mathbf{h}}^H\left[\tilde{\mathbf{x}}(k) + \tilde{\mathbf{v}}(k)\right] = x_{\mathrm{f}}(k) + v_{\mathrm{rn}}(k) \tag{12}$$

where $\tilde{\mathbf{x}}(k)$ and $\tilde{\mathbf{v}}(k)$ are defined in a similar way to $\tilde{\mathbf{y}}(k)$, $x_{\mathrm{f}}(k) \triangleq \tilde{\mathbf{h}}^H\tilde{\mathbf{x}}(k)$ is a filtered version of the desired signal and its conjugate of L successive time samples, and $v_{\mathrm{rn}}(k) \triangleq \tilde{\mathbf{h}}^H\tilde{\mathbf{v}}(k)$ is the residual noise. From (12), we see that $\hat{x}(k)$ depends on the vector $\tilde{\mathbf{x}}(k)$. However, our desired signal at time k is only $x(k)$ [and not the whole vector $\tilde{\mathbf{x}}(k)$]; so we should decompose the vector $\tilde{\mathbf{x}}(k)$ into two orthogonal vectors: one corresponding to the desired signal at time k and the other corresponding to the interference. Let us first decompose $\mathbf{x}(k)$ and $\mathbf{x}^*(k)$ separately.

The vector $\mathbf{x}(k)$ can be written as

$$\mathbf{x}(k) = x(k)\boldsymbol{\rho}_x + \mathbf{x}'(k) \tag{13}$$

where

$$\begin{aligned} \boldsymbol{\rho}_x &= [\, \rho_{x,0} \quad \rho_{x,1} \quad \cdots \quad \rho_{x,L-1} \,]^T \\ &= [\, 1 \quad \rho_{x,1} \quad \cdots \quad \rho_{x,L-1} \,]^T \\ &= \frac{E\left[\mathbf{x}(k)x^*(k)\right]}{\sigma_x^2} \end{aligned} \tag{14}$$

is the (normalized) correlation vector (of length L) between $\mathbf{x}(k)$ and $x(k)$,

$$\rho_{x,l} = \frac{E\left[x(k-l)x^*(k)\right]}{\sigma_x^2} \tag{15}$$

is the correlation coefficient between $x(k-l)$ and $x(k)$ with $|\rho_{x,l}| \leq 1$, and

$$\mathbf{x}'(k) = \mathbf{x}(k) - x(k)\boldsymbol{\rho}_x \tag{16}$$

is the interference signal vector. Obviously, $x(k)\boldsymbol{\rho}_x$ is correlated with $x(k)$ and

$$E\left[\mathbf{x}'(k)x^*(k)\right]=\mathbf{0} \tag{17}$$

so $\mathbf{x}'(k)$ is uncorrelated with $x(k)$.

Similarly, we have

$$\mathbf{x}^*(k)=x(k)\boldsymbol{\gamma}_x^*+\mathbf{x}''(k) \tag{18}$$

where

$$\begin{aligned}\boldsymbol{\gamma}_x&=\begin{bmatrix}\gamma_{x,0} & \gamma_{x,1} & \cdots & \gamma_{x,L-1}\end{bmatrix}^T\\&=\frac{E\left[\mathbf{x}(k)x(k)\right]}{\sigma_x^2}\end{aligned} \tag{19}$$

is the (normalized) correlation vector (of length L) between $\mathbf{x}(k)$ and $x^*(k)$

$$\gamma_{x,l}=\frac{E\left[x(k-l)x(k)\right]}{\sigma_x^2} \tag{20}$$

is the correlation coefficient[1] between $x(k-l)$ and $x^*(k)$ with $|\gamma_{x,l}|\leq 1$, and

$$\mathbf{x}''(k)=\mathbf{x}^*(k)-x(k)\boldsymbol{\gamma}_x^* \tag{21}$$

is the interference signal vector. Clearly, $x(k)\boldsymbol{\gamma}_x^*$ is correlated with $x(k)$, while $\mathbf{x}''(k)$ and $x(k)$ are uncorrelated since

$$E\left[\mathbf{x}''(k)x^*(k)\right]=\mathbf{0}. \tag{22}$$

Combining (13) and (18), we get

$$\tilde{\mathbf{x}}(k)=x(k)\mathbf{d}_x+\tilde{\mathbf{x}}'(k)=\mathbf{x}_{\mathrm{d}}(k)+\tilde{\mathbf{x}}'(k) \tag{23}$$

where

$$\mathbf{d}_x\triangleq\begin{bmatrix}\boldsymbol{\rho}_x\\ \boldsymbol{\gamma}_x^*\end{bmatrix} \tag{24}$$

$$\tilde{\mathbf{x}}'(k)\triangleq\begin{bmatrix}\mathbf{x}'(k)\\ \mathbf{x}''(k)\end{bmatrix} \tag{25}$$

$\mathbf{x}_{\mathrm{d}}(k)\triangleq x(k)\mathbf{d}_x$ is correlated with the desired signal, $x(k)$, and will contribute to its estimation, so we call it the desired signal vector. In comparison, $\tilde{\mathbf{x}}'(k)$ is uncorrelated with $x(k)$, and will interfere with the estimation, so we call it the interference signal vector.

Substituting (23) into (12), we obtain

$$\begin{aligned}\hat{x}(k)&=\tilde{\mathbf{h}}^H\left[x(k)\mathbf{d}_x+\tilde{\mathbf{x}}'(k)+\tilde{\mathbf{v}}(k)\right]\\&=x_{\mathrm{fd}}(k)+x'_{\mathrm{ri}}(k)+v_{\mathrm{rn}}(k)\end{aligned} \tag{26}$$

where $x_{\mathrm{fd}}(k)\triangleq x(k)\tilde{\mathbf{h}}^H\mathbf{d}_x$ is the filtered desired signal and $x'_{\mathrm{ri}}(k)\triangleq\tilde{\mathbf{h}}^H\tilde{\mathbf{x}}'(k)$ is the residual interference. We observe that the estimate of the desired signal at time k is the sum of three terms that are mutually uncorrelated. Therefore, the variance of $\hat{x}(k)$ is

$$\sigma_{\hat{x}}^2=\sigma_{x_{\mathrm{fd}}}^2+\sigma_{x'_{\mathrm{ri}}}^2+\sigma_{v_{\mathrm{rn}}}^2 \tag{27}$$

where

$$\sigma_{x_{\mathrm{fd}}}^2=\sigma_x^2\left|\tilde{\mathbf{h}}^H\mathbf{d}_x\right|^2=\tilde{\mathbf{h}}^H\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\tilde{\mathbf{h}} \tag{28}$$

$$\sigma_{x'_{\mathrm{ri}}}^2=\mathbf{h}^H\mathbf{R}_{\tilde{\mathbf{x}}'}\tilde{\mathbf{h}}=\tilde{\mathbf{h}}^H\mathbf{R}_{\tilde{\mathbf{x}}}\tilde{\mathbf{h}}-\sigma_x^2\left|\tilde{\mathbf{h}}^H\mathbf{d}_x\right|^2 \tag{29}$$

$$\sigma_{v_{\mathrm{rn}}}^2=\tilde{\mathbf{h}}^H\mathbf{R}_{\tilde{\mathbf{v}}}\tilde{\mathbf{h}} \tag{30}$$

$\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}=\sigma_x^2\mathbf{d}_x\mathbf{d}_x^H$ is the correlation matrix (whose rank is equal to 1) of $\mathbf{x}_{\mathrm{d}}(k)$, and $\mathbf{R}_{\tilde{\mathbf{x}}'}=E\left[\tilde{\mathbf{x}}'(k)\tilde{\mathbf{x}}'^H(k)\right]$, $\mathbf{R}_{\tilde{\mathbf{x}}}=E\left[\tilde{\mathbf{x}}(k)\tilde{\mathbf{x}}^H(k)\right]$, $\mathbf{R}_{\tilde{\mathbf{v}}}=E\left[\tilde{\mathbf{v}}(k)\tilde{\mathbf{v}}^H(k)\right]$ are the correlation matrices of $\tilde{\mathbf{x}}'(k)$, $\tilde{\mathbf{x}}(k)$, and $\tilde{\mathbf{v}}(k)$, respectively.

It is clear from (26) that the objective of our noise reduction problem is to find optimal filters that can minimize the effect of $x'_{\mathrm{ri}}(k)+v_{\mathrm{rn}}(k)$ while preserving the desired signal $x(k)$, but before deriving such filters, we first give some very useful performance measures for the evaluation of the time-domain binaural noise reduction problem with the WL model.

IV. Performance Measures

How to assess noise reduction filters is a very important issue. In this section, we give some measures that will be used in this paper to evaluate the noise reduction performance. The first and most fundamental one is the signal-to-noise ratio (SNR). The input SNR is defined as

$$\mathrm{iSNR}\triangleq\frac{\sigma_x^2}{\sigma_v^2} \tag{31}$$

where $\sigma_v^2\triangleq E\left[|v(k)|^2\right]$ is the variance of the complex additive noise.

To quantify the level of noise remaining at the output of the complex WL filter, we define the output SNR as the ratio of the variance of the filtered desired signal over the variance of the residual interference-plus-noise,[2] i.e.,

$$\mathrm{oSNR}\left(\tilde{\mathbf{h}}\right)\triangleq\frac{\sigma_{x_{\mathrm{fd}}}^2}{\sigma_{x'_{\mathrm{ri}}}^2+\sigma_{v_{\mathrm{rn}}}^2}=\frac{\sigma_x^2\left|\tilde{\mathbf{h}}^H\mathbf{d}_x\right|^2}{\tilde{\mathbf{h}}^H\mathbf{R}_{\mathrm{in}}\tilde{\mathbf{h}}}=\frac{\tilde{\mathbf{h}}^H\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\tilde{\mathbf{h}}}{\tilde{\mathbf{h}}^H\mathbf{R}_{\mathrm{in}}\tilde{\mathbf{h}}} \tag{32}$$

where

$$\mathbf{R}_{\mathrm{in}}=\mathbf{R}_{\tilde{\mathbf{x}}'}+\mathbf{R}_{\tilde{\mathbf{v}}} \tag{33}$$

is the interference-plus-noise covariance matrix. The objective of the noise reduction filter is to make the output SNR greater than the input SNR so that the quality of the noisy signal will be enhanced. For the particular filter $\tilde{\mathbf{h}}=\mathbf{i}_1$, where $\mathbf{i}_1$ is the first column of the identity matrix $\mathbf{I}_{2L}$ of size $2L\times 2L$, we have

$$\mathrm{oSNR}\left(\mathbf{i}_1\right)=\mathrm{iSNR}. \tag{34}$$

[1]Note that $\gamma_{x,0}=\gamma_x$, which is the circularity quotient for the complex signal $x(k)$.

[2]In this paper, we consider the interference as part of the noise in the definitions of the performance measures.

With the filter $\mathbf{i}_1$, the SNR cannot be improved.

Now, let us introduce the quantity $\mathrm{oSNR}_{\max}$, which is defined as the maximum output SNR that can be achieved through filtering so that

$$\mathrm{oSNR}\left(\tilde{\mathbf{h}}\right) \leq \mathrm{oSNR}_{\max}, \ \forall \tilde{\mathbf{h}}. \tag{35}$$

It can be checked from (32) that this quantity is equal to the maximum eigenvalue of the matrix $\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}$, i.e.,

$$\mathrm{oSNR}_{\max} = \lambda_{\max}\left(\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\right). \tag{36}$$

The filter that can achieve $\mathrm{oSNR}_{\max}$ is called the maximum SNR filter and is denoted by $\tilde{\mathbf{h}}_{\max}$. It is easy to see from (36) that $\tilde{\mathbf{h}}_{\max}$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}$. Clearly, we have

$$\mathrm{oSNR}_{\max} = \mathrm{oSNR}\left(\tilde{\mathbf{h}}_{\max}\right) \geq \mathrm{oSNR}\left(\mathbf{i}_1\right) = \mathrm{iSNR}. \tag{37}$$

Since the rank of the matrix $\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}$ is equal to 1, we also have

$$\mathrm{oSNR}_{\max} = \mathrm{tr}\left(\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}\right) = \sigma_x^2\mathbf{d}_x^H\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{d}_x \tag{38}$$

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

The noise reduction factor [30], [31] quantifies the amount of noise that is rejected by the filter. This quantity is defined as the ratio of the variance of the noise over the variance of the interference-plus-noise remaining after the filtering operation, i.e.,

$$\xi_{\mathrm{nr}}\left(\tilde{\mathbf{h}}\right) \triangleq \frac{\sigma_v^2}{\sigma_{x'_{\mathrm{ri}}}^2 + \sigma_{v_{\mathrm{rn}}}^2} = \frac{\sigma_v^2}{\tilde{\mathbf{h}}^H\mathbf{R}_{\mathrm{in}}\tilde{\mathbf{h}}}. \tag{39}$$

The higher is the value of $\xi_{\mathrm{nr}}\left(\tilde{\mathbf{h}}\right)$, the more the noise is reduced. This factor is expected to be lower bounded by 1 for optimal filters.

In practice, the WL filter $\tilde{\mathbf{h}}$ may distort the desired signal. In order to evaluate the level of this distortion, we define the speech reduction factor [27] as the variance of the desired signal over the variance of the filtered desired signal, i.e.,

$$\xi_{\mathrm{sr}}\left(\tilde{\mathbf{h}}\right) \triangleq \frac{\sigma_x^2}{\sigma_{x_{\mathrm{fd}}}^2} = \frac{1}{\left|\tilde{\mathbf{h}}^H\mathbf{d}_x\right|^2}. \tag{40}$$

An important observation is that the design of a filter that does not distort the complex desired signal requires the constraint

$$\tilde{\mathbf{h}}^H\mathbf{d}_x = 1. \tag{41}$$

Thus, the speech reduction factor is equal to 1 if there is no distortion and expected to be greater than 1 when distortion occurs.

By making the appropriate substitutions, one can derive the relationship among the four previous measures:

$$\frac{\mathrm{oSNR}\left(\tilde{\mathbf{h}}\right)}{\mathrm{iSNR}} = \frac{\xi_{\mathrm{nr}}\left(\tilde{\mathbf{h}}\right)}{\xi_{\mathrm{sr}}\left(\tilde{\mathbf{h}}\right)}. \tag{42}$$

When no distortion occurs in the desired signal, the gain in SNR coincides with the noise reduction factor. Expression (42) indicates the equivalence between gain/loss in SNR and distortion. In other words, a gain in SNR can be achieved only if the desired signal and/or noise are/is distorted.

Another useful performance measure is the speech distortion index [30], [31] defined as

$$\upsilon_{\mathrm{sd}}\left(\tilde{\mathbf{h}}\right) \triangleq \frac{E\left[\left|x_{\mathrm{fd}}(k) - x(k)\right|^2\right]}{\sigma_x^2} = \left|\tilde{\mathbf{h}}^H\mathbf{d}_x - 1\right|^2. \tag{43}$$

The speech distortion index is always greater than or equal to 0 and should be upper bounded by 1 for optimal noise reduction filters. The higher is the value of $\upsilon_{\mathrm{sd}}\left(\tilde{\mathbf{h}}\right)$, the more the desired signal is distorted.

V. Optimal Filters

In this part, we derive the optimal filters for binaural noise reduction in stereo systems. For that, we need to derive first the mean-square error (MSE) criterion and its relation with the MSEs of speech distortion and residual interference-plus-noise.

We define the error signal between the estimated and desired signals as

$$e(k) \triangleq \hat{x}(k) - x(k) = \tilde{\mathbf{h}}^H\tilde{\mathbf{y}}(k) - x(k) \tag{44}$$

which can be written as the sum of two uncorrelated error signals

$$e(k) = e_{\mathrm{d}}(k) + e_{\mathrm{r}}(k) \tag{45}$$

where

$$e_{\mathrm{d}}(k) \triangleq x_{\mathrm{fd}}(k) - x(k) \tag{46}$$

is the signal distortion due to the WL filter and

$$e_{\mathrm{r}}(k) \triangleq x'_{\mathrm{ri}}(k) + v_{\mathrm{rn}}(k) \tag{47}$$

represents the residual interference-plus-noise.

The MSE is then

$$J\left(\tilde{\mathbf{h}}\right) \triangleq E\left[\left|e(k)\right|^2\right] = J_{\mathrm{d}}\left(\tilde{\mathbf{h}}\right) + J_{\mathrm{r}}\left(\tilde{\mathbf{h}}\right) \tag{48}$$

where

$$J_{\mathrm{d}}\left(\tilde{\mathbf{h}}\right) \triangleq E\left[\left|e_{\mathrm{d}}(k)\right|^2\right] = \sigma_x^2\left|\tilde{\mathbf{h}}^H\mathbf{d}_x - 1\right|^2 \tag{49}$$

and

$$J_{\mathrm{r}}\left(\tilde{\mathbf{h}}\right) \triangleq E\left[\left|e_{\mathrm{r}}(k)\right|^2\right] = \sigma_{x'_{\mathrm{ri}}}^2 + \sigma_{v_{\mathrm{rn}}}^2. \tag{50}$$

For the particular filter $\tilde{\mathbf{h}} = \mathbf{i}_1$, the MSE is

$$J\left(\mathbf{i}_1\right) = \sigma_v^2 \tag{51}$$

so there is neither noise reduction nor speech distortion. We can now define the normalized MSE (NMSE) as

$$J_{\rm n}\left(\tilde{\mathbf{h}}\right) \triangleq \frac{J\left(\tilde{\mathbf{h}}\right)}{J\left(\mathbf{i}_1\right)} = {\rm iSNR}\cdot \upsilon_{\rm sd}\left(\tilde{\mathbf{h}}\right) + \frac{1}{\xi_{\rm nr}\left(\tilde{\mathbf{h}}\right)}$$
$$= {\rm iSNR}\left[\upsilon_{\rm sd}\left(\tilde{\mathbf{h}}\right) + \frac{1}{{\rm oSNR}\left(\tilde{\mathbf{h}}\right)\cdot \xi_{\rm sr}\left(\tilde{\mathbf{h}}\right)}\right] \quad (52)$$

where

$$\upsilon_{\rm sd}\left(\tilde{\mathbf{h}}\right) \triangleq \frac{J_{\rm d}\left(\tilde{\mathbf{h}}\right)}{\sigma_x^2} \quad (53)$$
$$\xi_{\rm nr}\left(\tilde{\mathbf{h}}\right) \triangleq \frac{\sigma_v^2}{J_{\rm r}\left(\tilde{\mathbf{h}}\right)}. \quad (54)$$

This shows the relationship between the MSEs and the performance measures defined in the previous section.

It is clear that the objective of noise reduction with the WL model is to find optimal WL filters that would either minimize $J\left(\tilde{\mathbf{h}}\right)$ or minimize $J_{\rm r}\left(\tilde{\mathbf{h}}\right)$ or $J_{\rm d}\left(\tilde{\mathbf{h}}\right)$ subject to some constraint.

A. Wiener

The Wiener filter is easily derived by taking the gradient of the MSE, $J\left(\tilde{\mathbf{h}}\right)$, with respect to $\tilde{\mathbf{h}}$ and equating the result to zero

$$\tilde{\mathbf{h}}_{\rm W} = \mathbf{R}_{\tilde{\mathbf{y}}}^{-1}\mathbf{R}_{\tilde{\mathbf{x}}}\mathbf{i}_1. \quad (55)$$

Using the fact that $\mathbf{R}_{\tilde{\mathbf{y}}} = \mathbf{R}_{\tilde{\mathbf{x}}} + \mathbf{R}_{\tilde{\mathbf{v}}}$, we can rewrite (55) as

$$\tilde{\mathbf{h}}_{\rm W} = \left(\mathbf{I}_{2L} - \mathbf{R}_{\tilde{\mathbf{y}}}^{-1}\mathbf{R}_{\tilde{\mathbf{v}}}\right)\mathbf{i}_1. \quad (56)$$

Since

$$\mathbf{R}_{\tilde{\mathbf{x}}}\mathbf{i}_1 = \sigma_x^2\mathbf{d}_x \quad (57)$$

we also have

$$\tilde{\mathbf{h}}_{\rm W} = \sigma_x^2\mathbf{R}_{\tilde{\mathbf{y}}}^{-1}\mathbf{d}_x. \quad (58)$$

It can be verified from Section III that

$$\mathbf{R}_{\tilde{\mathbf{y}}} = \sigma_x^2\mathbf{d}_x\mathbf{d}_x^H + \mathbf{R}_{\rm in}. \quad (59)$$

Determining the inverse of $\mathbf{R}_{\tilde{\mathbf{y}}}$ from (59) with the Woodbury's identity

$$\mathbf{R}_{\tilde{\mathbf{y}}}^{-1} = \mathbf{R}_{\rm in}^{-1} - \frac{\mathbf{R}_{\rm in}^{-1}\mathbf{d}_x\mathbf{d}_x^H\mathbf{R}_{\rm in}^{-1}}{\sigma_x^{-2} + \mathbf{d}_x^H\mathbf{R}_{\rm in}^{-1}\mathbf{d}_x} \quad (60)$$

and substituting the result into (58) leads to another interesting form of the Wiener filter

$$\tilde{\mathbf{h}}_{\rm W} = \frac{\mathbf{R}_{\rm in}^{-1}\mathbf{d}_x}{\sigma_x^{-2} + \mathbf{d}_x^H\mathbf{R}_{\rm in}^{-1}\mathbf{d}_x} \quad (61)$$

that we can rewrite as

$$\tilde{\mathbf{h}}_{\rm W} = \frac{\mathbf{R}_{\rm in}^{-1}\mathbf{R}_{\tilde{\mathbf{y}}} - \mathbf{I}_{2L}}{1 - 2L + {\rm tr}\left(\mathbf{R}_{\rm in}^{-1}\mathbf{R}_{\tilde{\mathbf{y}}}\right)}\mathbf{i}_1 = \frac{\mathbf{R}_{\rm in}^{-1}\mathbf{R}_{\mathbf{x}_{\rm d}}}{1 + {\rm oSNR}_{\rm max}}\mathbf{i}_1. \quad (62)$$

Using (61), we deduce that the output SNR of the Wiener filter is

$${\rm oSNR}\left(\tilde{\mathbf{h}}_{\rm W}\right) = {\rm oSNR}_{\rm max} = {\rm tr}\left(\mathbf{R}_{\rm in}^{-1}\mathbf{R}_{\tilde{\mathbf{y}}}\right) - 2L \quad (63)$$

and the corresponding speech distortion index is a clear function of the maximum output SNR

$$\upsilon_{\rm sd}\left(\tilde{\mathbf{h}}_{\rm W}\right) = \frac{1}{\left(1 + {\rm oSNR}_{\rm max}\right)^2}. \quad (64)$$

So, the higher is the value of ${\rm oSNR}_{\rm max}$, the less the desired signal is distorted.

Clearly,

$${\rm oSNR}\left(\tilde{\mathbf{h}}_{\rm W}\right) \geq {\rm iSNR} \quad (65)$$

since the Wiener filter maximizes the output SNR. It is of great interest to observe that the two filters $\tilde{\mathbf{h}}_{\rm W}$ and $\tilde{\mathbf{h}}_{\rm max}$ both maximize the output SNR and they are equivalent up to a scaling factor.

With the Wiener filter, the noise reduction factor is

$$\xi_{\rm nr}\left(\tilde{\mathbf{h}}_{\rm W}\right) = \frac{\left(1 + {\rm oSNR}_{\rm max}\right)^2}{{\rm iSNR}\cdot{\rm oSNR}_{\rm max}} \geq \left(1 + \frac{1}{{\rm oSNR}_{\rm max}}\right)^2. \quad (66)$$

Substituting (64) and (66) into (52), we find the minimum NMSE (MNMSE):

$$J_{\rm n}\left(\tilde{\mathbf{h}}_{\rm W}\right) = \frac{\rm iSNR}{1 + {\rm oSNR}\left(\tilde{\mathbf{h}}_{\rm W}\right)}. \quad (67)$$

B. Minimum Variance Distortionless Response

The celebrated minimum variance distortionless response (MVDR) filter proposed by Capon [32], [33] can be derived in this context by minimizing the MSE of the residual interference-plus-noise, $J_{\rm r}\left(\tilde{\mathbf{h}}\right)$, with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\min_{\tilde{\mathbf{h}}} \tilde{\mathbf{h}}^H\mathbf{R}_{\rm in}\tilde{\mathbf{h}} \quad \text{subject to} \quad \tilde{\mathbf{h}}^H\mathbf{d}_x = 1 \quad (68)$$

for which the solution is

$$\tilde{\mathbf{h}}_{\rm MVDR} = \frac{\mathbf{R}_{\rm in}^{-1}\mathbf{d}_x}{\mathbf{d}_x^H\mathbf{R}_{\rm in}^{-1}\mathbf{d}_x} \quad (69)$$

that we can rewrite as

$$\tilde{\mathbf{h}}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\tilde{\mathbf{y}}} - \mathbf{I}_{2L}}{\mathrm{tr}\left(\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\tilde{\mathbf{y}}}\right) - 2L}\mathbf{i}_1 = \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{R}_{\mathbf{x}_{\mathrm{d}}}}{\mathrm{oSNR}_{\max}}\mathbf{i}_1. \quad (70)$$

Obviously, we can rewrite the MVDR as

$$\tilde{\mathbf{h}}_{\mathrm{MVDR}} = \frac{\mathbf{R}_{\tilde{\mathbf{y}}}^{-1}\mathbf{d}_x}{\mathbf{d}_x^H\mathbf{R}_{\tilde{\mathbf{y}}}^{-1}\mathbf{d}_x}. \quad (71)$$

The Wiener and MVDR filters are related as follows:

$$\tilde{\mathbf{h}}_{\mathrm{W}} = \alpha\tilde{\mathbf{h}}_{\mathrm{MVDR}} \quad (72)$$

where

$$\alpha = \tilde{\mathbf{h}}_{\mathrm{W}}^H\mathbf{d}_x = \frac{\mathrm{oSNR}_{\max}}{1+\mathrm{oSNR}_{\max}}. \quad (73)$$

Here again the two filters $\tilde{\mathbf{h}}_{\mathrm{W}}$ and $\tilde{\mathbf{h}}_{\mathrm{MVDR}}$ are equivalent up to a scaling factor. From a theoretical point of view, this scaling is not significant as it does not affect the output SNR, but from a practical point of view it can be important since it is time-varying and can cause discontinuity in the residual noise level. Therefore, it is essential to have this scaling factor right from one frame to another in order to avoid large distortions. Therefore, it is recommended to use the MVDR filter rather than the Wiener filter in speech enhancement applications.

It is clear that we always have

$$\mathrm{oSNR}\left(\tilde{\mathbf{h}}_{\mathrm{MVDR}}\right) = \mathrm{oSNR}\left(\tilde{\mathbf{h}}_{\mathrm{W}}\right) \quad (74)$$

$$\upsilon_{\mathrm{sd}}\left(\tilde{\mathbf{h}}_{\mathrm{MVDR}}\right) = 0 \quad (75)$$

$$\xi_{\mathrm{sr}}\left(\tilde{\mathbf{h}}_{\mathrm{MVDR}}\right) = 1 \quad (76)$$

$$\xi_{\mathrm{nr}}\left(\tilde{\mathbf{h}}_{\mathrm{MVDR}}\right) = \frac{\mathrm{oSNR}_{\max}}{\mathrm{iSNR}} \leq \xi_{\mathrm{nr}}\left(\tilde{\mathbf{h}}_{\mathrm{W}}\right) \quad (77)$$

and

$$1 \geq J_{\mathrm{n}}\left(\tilde{\mathbf{h}}_{\mathrm{MVDR}}\right) = \frac{\mathrm{iSNR}}{\mathrm{oSNR}_{\max}} \geq J_{\mathrm{n}}\left(\tilde{\mathbf{h}}_{\mathrm{W}}\right). \quad (78)$$

C. Tradeoff

In the tradeoff approach, we try to maintain a compromise between noise reduction and speech distortion. Here we minimize the speech distortion index with the constraint that the noise reduction factor is equal to a positive value that is greater than 1. Mathematically, this is equivalent to

$$\min_{\tilde{\mathbf{h}}} J_{\mathrm{d}}\left(\tilde{\mathbf{h}}\right) \quad \text{subject to} \quad J_{\mathrm{r}}\left(\tilde{\mathbf{h}}\right) = \beta\sigma_v^2 \quad (79)$$

where $0 < \beta < 1$ to ensure that we get some noise reduction. By using a Lagrange multiplier, $\mu \geq 0$, to adjoin the constraint to the cost function, we easily deduce the tradeoff filter

$$\begin{aligned}\tilde{\mathbf{h}}_{\mathrm{T},\mu} &= \sigma_x^2\left(\sigma_x^2\mathbf{d}_x\mathbf{d}_x^H + \mu\mathbf{R}_{\mathrm{in}}\right)^{-1}\mathbf{d}_x \\ &= \frac{\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{d}_x}{\mu\sigma_x^{-2} + \mathbf{d}_x^H\mathbf{R}_{\mathrm{in}}^{-1}\mathbf{d}_x}\end{aligned} \quad (80)$$

where the Lagrange multiplier μ satisfies $J_{\mathrm{r}}\left(\tilde{\mathbf{h}}_{\mathrm{T},\mu}\right) = \beta\sigma_v^2$. Setting $\mu = 1$ leads to the Wiener filter while taking $\mu = 0$, gives the MVDR. By playing on the value of μ, we can make a tradeoff between the amount of noise reduction and the amount of speech distortion. However, the output SNR of the tradeoff filter is independent of μ and is identical to the output SNR of the Wiener filter, i.e.,

$$\mathrm{oSNR}\left(\tilde{\mathbf{h}}_{\mathrm{T},\mu}\right) = \mathrm{oSNR}\left(\tilde{\mathbf{h}}_{\mathrm{W}}\right), \ \forall\mu \geq 0. \quad (81)$$

Again, we observe here as well that the tradeoff and Wiener filters are equivalent up to a scaling factor.

VI. Experimental Results

We have carried out a number of experiments to study the previously developed noise reduction filters in practical acoustic environments under different operation conditions. In this section, we will present some of the results, which highlight the merits and limitations inherent in each noise reduction filter, and justify what we learned through theoretical analysis in the previous sections.

A. Experimental Setup

The experiments were conducted with the impulse responses measured in the varechoic chamber at Bell Labs [34], [35]. The chamber is a rectangular room, which measures 6700 mm long by 6100 mm wide by 2900 mm high and is equipped with 368 electronically controlled panels. Each panel consists of two perforated sheets whose holes, if aligned, expose sound absorbing material (fiberglass) behind, but if shifted to misalign, form a highly reflective surface. Each panel can be individually controlled so that the holes on a particular panel are either fully open (absorbing) or fully closed (reflective). As a result, a total of 2^{368} different room characteristics can be generated by varying the binary states of the 368 panels in different combinations.

A diagram of the floor layout of the experimental setup is illustrated in Fig. 1. For convenience, positions in the floor plan are designated by (x, y) coordinates with reference to the northwest corner and corresponding to millimeters along the (north, west) walls. A stereo system with two microphones and two loudspeakers is configured. The two microphones are located respectively at (3437, 500) and (3537, 500) and their outputs are processed and then sent to the two loudspeakers (which are not shown in the figure) for listening. Another loudspeaker, which plays back a speech signal prerecorded from a female talker, is used to simulate a moving speech source and it moves back and forth from positions P1 to P7 as shown in Fig. 1. The seven positions are uniformly spaced along the line $y = 1938$ with the first position P1 at (337, 1938) and the last position P7 at (6337, 1938). Note that the elevation for microphones is 1400 while it is 1600 for source positions.

To make the experiments repeatable, the acoustic channel impulse responses were measured from the seven source positions to the two microphones. The measurement was carried out with a sampling rate of 8 kHz when 89% of the chamber panels were

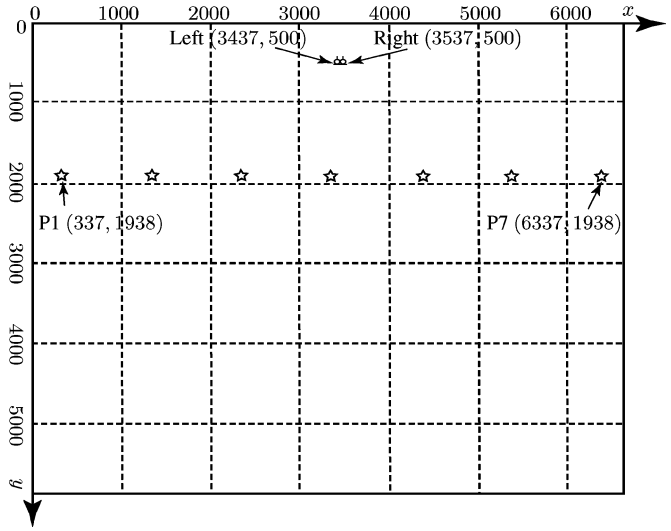

Fig. 1. Floor layout of the experimental setup in the varechoic chamber (coordinate values measured in millimeters). The two microphones are located at (3437, 500) and (3537, 500) respectively (with $z = 1400$). A loudspeaker is placed at one of the seven positions from P1 to P7 to simulate a moving speech source. The seven positions are uniformly spaced along the line connecting P1 and P7 (with $z = 1600$).

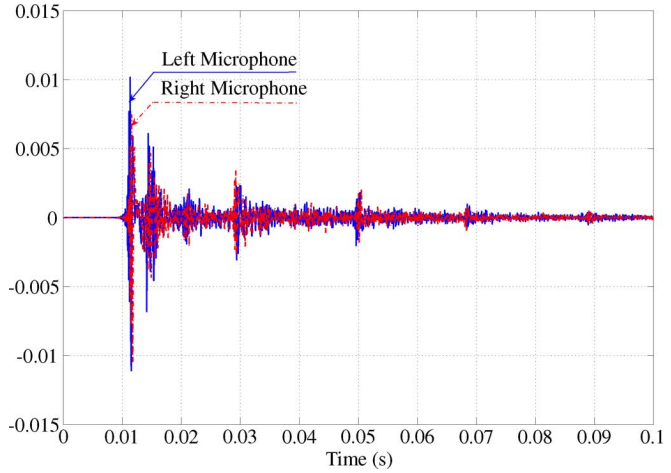

Fig. 2. The impulse responses (first 0.1 s) from the position P1 to the two microphones that are measured when 89% of the varechoic chamber panels are open and the corresponding reverberation time T_{60} is approximately 0.24 s. The sampling rate is 8 kHz.

open (corresponding to a reverberation time T_{60} of approximately 0.24 s). As an example, Fig. 2 plots the two impulse responses measured from P1 to the two microphones. The difference between the two impulse responses determines the spatial effect of the sound source at position P1.

The microphone outputs were generated by convolving the source signal with the corresponding measured impulse responses and noise was then added to the convolved results to control the SNR level. The source signal used in our experiments was recorded from a female talker in a quiet office room. It was sampled at 8 kHz. The overall length of the signal is approximately 4 minutes. To simulate a moving source, we changed the source position (i.e., used a new set of impulse responses) every 4.3 s first from P1 to P7 and then back and forth. Each time the movement was restricted to the position immediately next to the current one. Fig. 3 shows the waveforms of the two microphones' outputs (only the first 10 s) in the absence

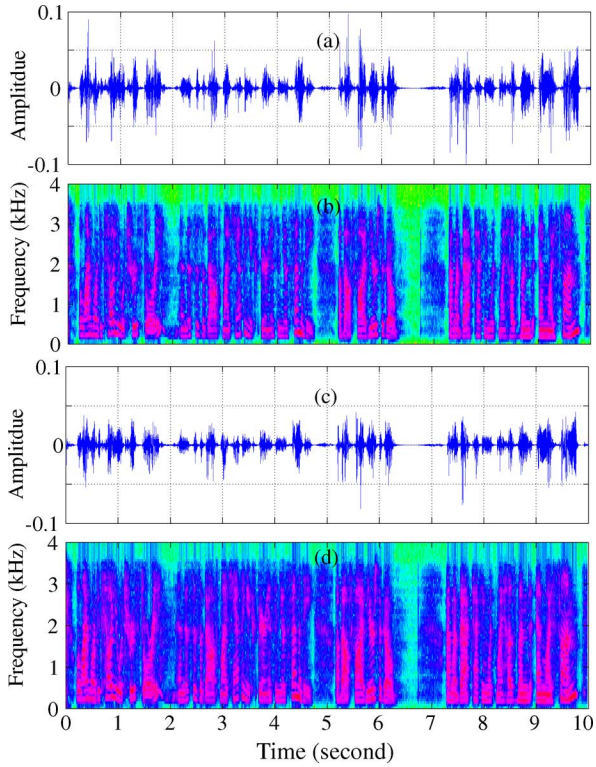

Fig. 3. First 10-s signals of the stereo system and their spectrograms. (a) Left-channel speech. (b) Spectrogram of the left-channel speech. (c) Right-channel speech. (d) Spectrogram of the right-channel speech.

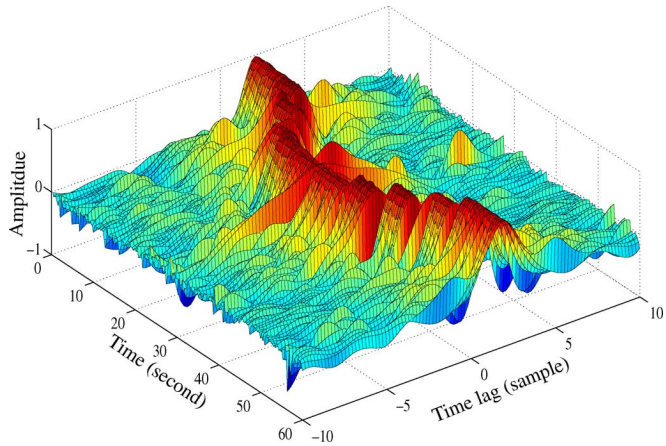

Fig. 4. Short-time cross-correlation function between the left- and right-channel speech signals. It is computed using a short-time average with a window length of 64 ms (no overlap).

of noise and the corresponding spectrograms. To visualize the spatial sound effect, we computed the cross-correlation function between the two channels every 64 ms using a short-time average with a frame size of 64 ms. The results are plotted in Fig. 4, where the peak of the cross-correlation function at each time corresponds to the current source position.

The noisy speech is obtained by adding noise to the speech where the noise signal is properly scaled to control the input SNR level. We consider two types of noise: a computer generated white Gaussian random process and a babble noise signal recorded in a New York Stock Exchange (NYSE) room. The NYSE noise was also digitized with a sampling rate of 8 kHz. Compared to the Gaussian random noise which is stationary and

white, the NYSE noise is nonstationary and colored. It consists of sounds from various sources such as electrical fans, telephone rings, and even some background speech. Note that we do not have actual stereo recordings of the NYSE noise that matches our system configuration. Instead, we take two independent segments of this noise and add them to the two microphone signals.

B. Estimation of Correlation Matrices and Vectors

The implementation of the noise reduction filters derived in Section V requires the estimation of the correlation matrices $\mathbf{R}_{\tilde{\mathbf{y}}}$, $\mathbf{R}_{\tilde{\mathbf{x}}}$, and $\mathbf{R}_{\tilde{\mathbf{v}}}$, the correlation vector $\mathbf{d}_x$, and the variance σ_x^2. Computation of $\mathbf{R}_{\tilde{\mathbf{y}}}$ is relatively easy because the noisy signal vector $\tilde{\mathbf{y}}(k)$ is accessible, but we need a noise estimator or a VAD in practice to compute all the other parameters. While it is a very important issue (see [36] and references therein), how to effectively estimate the noise or its statistics in a stereo system is not the main thrust of this paper. So, we will set aside this issue and directly compute the noise statistics from the noise signal in the following experiments. In this way, we can focus on illustrating the basic ideas of how to achieve binaural noise reduction. Specifically, at each time instant k, an estimate of the matrices $\mathbf{R}_{\tilde{\mathbf{y}}}$ and $\mathbf{R}_{\tilde{\mathbf{x}}}$ are computed using the most recent 640 samples (40-ms long) of the noisy and clean signals, respectively, with a short-time average. The matrix $\mathbf{R}_{\tilde{\mathbf{v}}}$ is also computed using a short-time average; but noise is more stationary than speech, so the matrix $\mathbf{R}_{\tilde{\mathbf{v}}}$ is computed using 1280 samples (80-ms long). Then all the other parameters are computed in the following way: $\hat{\sigma}_x^2$ is taken as the first element of $\hat{\mathbf{R}}_{\tilde{\mathbf{x}}}$ and $\hat{\mathbf{d}}_x$ is equal to the first column of $\hat{\mathbf{R}}_{\tilde{\mathbf{x}}}$ normalized by $\hat{\sigma}_x^2$.

C. Computation of the Inverse Correlation Matrices

To implement the noise reduction filters given in Section V, we need to compute the inverse of either $\mathbf{R}_{\tilde{\mathbf{y}}}$ or $\mathbf{R}_{\text{in}}$. The size of both $\mathbf{R}_{\tilde{\mathbf{y}}}$ and $\mathbf{R}_{\text{in}}$ is $2L \times 2L$; the complexity of finding their inverse is, therefore, in the order of $(2L)^3$, which is very high. A slightly more efficient way to invert these matrices can be achieved through partitioning the correlation matrices into a block form. Take, for example, $\mathbf{R}_{\tilde{\mathbf{y}}}$, which can be written into the following partitioned form:

$$\mathbf{R}_{\tilde{\mathbf{y}}} = \begin{bmatrix} \mathbf{R}_{\mathbf{y}} & \mathbf{R}_{\mathbf{y}\mathbf{y}^*} \\ \mathbf{R}_{\mathbf{y}\mathbf{y}^*}^H & \mathbf{R}_{\mathbf{y}}^T \end{bmatrix} \tag{82}$$

where $\mathbf{R}_{\mathbf{y}\mathbf{y}^*} \triangleq E\{\mathbf{y}(k)[\mathbf{y}^*(k)]^H\} = E[\mathbf{y}(k)\mathbf{y}^T(k)]$. Now assuming that $\mathbf{R}_{\mathbf{y}}$ is nonsingular, which is true in practice, and denoting the Schur complement of $\mathbf{R}_{\mathbf{y}}$ in $\mathbf{R}_{\tilde{\mathbf{y}}}$ as

$$\mathbf{S} = \mathbf{R}_{\mathbf{y}} - \mathbf{R}_{\mathbf{y}\mathbf{y}^*}\mathbf{R}_{\mathbf{y}}^{-T}\mathbf{R}_{\mathbf{y}\mathbf{y}^*}^H \tag{83}$$

we can write the inverse matrix of $\mathbf{R}_{\tilde{\mathbf{y}}}$ into the following form:

$$\mathbf{R}_{\tilde{\mathbf{y}}}^{-1} = \begin{bmatrix} \mathbf{S}^{-1} & -\mathbf{R}_{\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{y}\mathbf{y}^*}\left(\mathbf{S}^{-1}\right)^* \\ -\mathbf{R}_{\mathbf{y}}^{-T}\mathbf{R}_{\mathbf{y}\mathbf{y}^*}^H\mathbf{S}^{-1} & \left(\mathbf{S}^{-1}\right)^* \end{bmatrix}. \tag{84}$$

Using the right-hand side of (84) to compute $\mathbf{R}_{\tilde{\mathbf{y}}}^{-1}$, we need to compute the inverse of $\mathbf{R}_{\mathbf{y}}$ and $\mathbf{S}$, both of size $L \times L$, which can be implemented more efficiently than directly computing the inverse of the $2L \times 2L$ matrix.

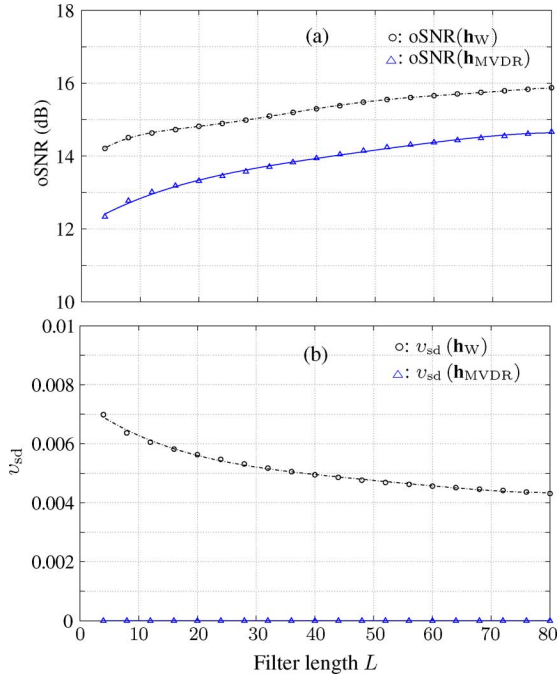

Fig. 5. Performance of the Wiener and MVDR filters as a function of the filter length, L, in the white Gaussian noise case with $\text{iSNR} = 10$ dB.

D. Comparison Between the Wiener and MVDR Filters

We derived both the Wiener and MVDR filters in Section V. A legitimate question one would ask is: which filter would perform better and more consistently in practice? In this experiment, we compare the two filters for binaural noise reduction in the stereo case. With the different correlation components computed using the method described in the previous subsection, we implemented the Wiener filter using (58) and the MVDR filter according to (71). We used the output SNR and speech distortion index defined in Section IV as the performance measures to evaluate the two filters. The two performance measures were computed in a global manner, i.e., with the constructed noise reduction filter and the computed correlation vector at each time instant, we first estimate the three signals $x_{\text{fd}}(k)$, $x'_{\text{ri}}(k)$, and $v_{\text{rn}}(k)$. A long-time average was then used to replace the expectation operation in (32) and (43) to compute the output SNR and speech distortion index. The results as a function of the filter length L for the white Gaussian noise case (with an input SNR of 10 dB) are depicted in Fig. 5. It is seen that the output SNR of both the Wiener and MVDR filters increases with the filter length L; so the longer the filter, the more the noise reduction, but once the value of L is larger than 60, further increasing L does not lead to much SNR improvement. Therefore, 60 is a sufficient value of the filter length for both the Wiener and MVDR filters with a sampling rate of 8 kHz. A larger length will not significantly improve the speech quality but would dramatically increase the complexity of the algorithm. It is also seen that the Wiener filter yields a higher output SNR than the MVDR filter, but the latter does not introduce speech distortion; as seen from Fig. 5(b), the speech distortion index for the MVDR is approximately 0. Comparatively, the value of the speech distortion index for the Wiener filter is much larger even though its value decreases as L increases.

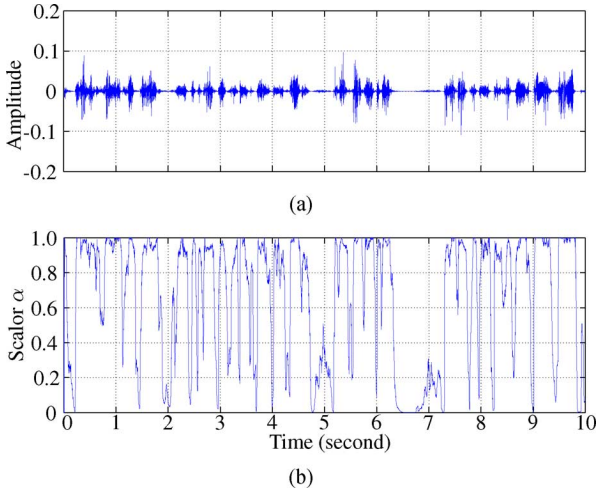

Fig. 6. (a) The speech waveform and (b) the scaling factor α between the Wiener and MVDR filters given in (73) with $L = 40$ and $\mathrm{iSNR} = 10$ dB.

In Section V, we showed that both the MVDR and Wiener filters maximize the output SNR. The only difference between these two filters is a scaling factor, given in (73), which depends on the statistics of the noisy and desired speech signals. If this scaling factor is time-invariant, the two filters should have the same performance. However, speech signals are always nonstationary and the background noise can possibly be nonstationary too. So, we need to estimate all the signal correlation matrices and vectors on a short-time basis in practice. As a result, the correlation matrices and vectors that are used to construct the noise reduction filters are time-varying, which subsequently make the scaling factor change over time, thereby causing performance difference between the Wiener and MVDR filters. To study this difference, we examine the scaling factor. We set the filter length L to 40 and the rest of the experimental conditions are the same as in the previous experiment. The scaling factor is computed according to (73) and the results are plotted in Fig. 6. It is seen that the value of the scaling factor is large (close to 1) during the presence of speech; but it is very small (close to 0) in silence periods. This illustrates that the Wiener and MVDR filters behave almost the same during the presence of speech while the Wiener filter is more aggressive in suppressing silence periods. As a result, the Wiener filter has more overall noise reduction, thereby yielding a higher output SNR, if we evaluate the performance on a global basis, but this comes with a cost: discontinuity of the background noise level in the output signal, i.e., the noise level is higher in the presence of speech than that in the absence of speech. This discontinuity is manifested in the form of speech distortion and that explains why the Wiener filter has a much higher value of the speech distortion index.

We also compared the Wiener and MVDR filters in different SNR conditions. The results for $L = 40$ are depicted in Fig. 7. It is seen that in the studied SNR range between 0 and 30 dB, both filters can improve the SNR, but the SNR improvement decreases as the input SNR increases. This, of course, makes sense since as the input SNR increases, there is less noise to be reduced. Comparatively, the Wiener filter yields a higher output SNR, particularly in low input SNR conditions, which is, again, due to the fact that the Wiener filter suppresses more

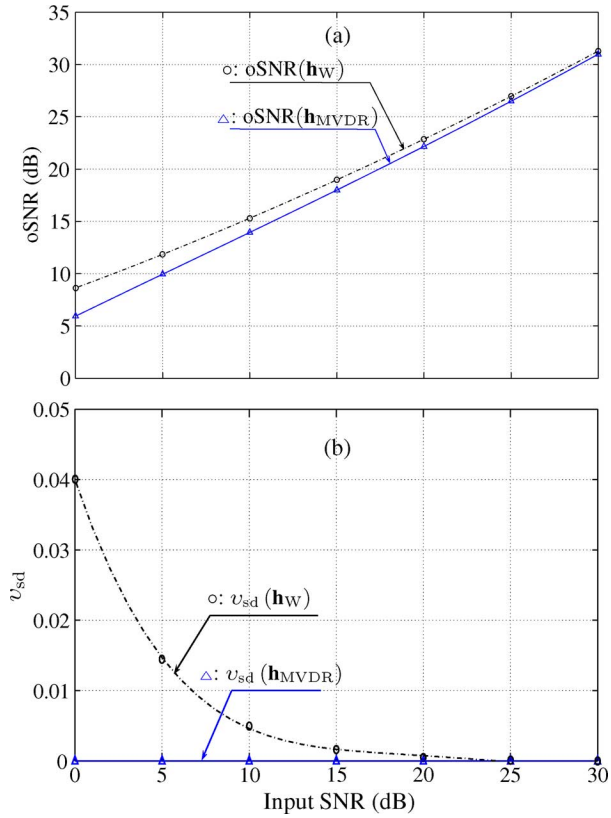

Fig. 7. Performance of the Wiener and MVDR filters as a function of the input SNR in the white Gaussian noise case with $L = 40$.

noise than the MVDR filter during the absence of speech. The MVDR filter does not introduce speech distortion, as can be seen in Fig. 7(b) where the value of the speech distortion index for the MVDR filter is approximately 0, regardless of the input SNR level. In comparison, the Wiener filter introduces speech distortion, which is inversely proportional to the input SNR. One can notice from Fig. 7 that the performance difference between the Wiener and MVDR filters is marginal when the input SNR is high (e.g., ≥ 20 dB). So, if the application environments are not very noisy, we can choose to use either of the two filters.

E. Performance of the Tradeoff Filter

A tradeoff filter was developed in Section V, where a parameter μ was introduced to control the compromise between the amount of noise reduction and the degree of speech distortion. We have shown that the tradeoff filter turns to the MVDR and Wiener for μ equal to 0 and 1, respectively. In this experiment, we validate the tradeoff filter through experiments. Based on the previous results, we set the filter length L to 40. The performance of the tradeoff filter as a function of μ in both the white Gaussian and NYSE noise conditions (with an input SNR of 10 dB) is plotted in Fig. 8. As expected, the output SNR increases as μ increases and so is the speech distortion index. It is also noticed in Fig. 8 that the Wiener ($\mu = 1$), MVDR ($\mu = 0$), and tradeoff filters have achieved a better noise reduction performance (a higher output SNR and a lower speech distortion index) in the NYSE noise than in the white Gaussian noise. This result is somehow unexpected as we know from the single-channel noise reduction that babble noise is in general

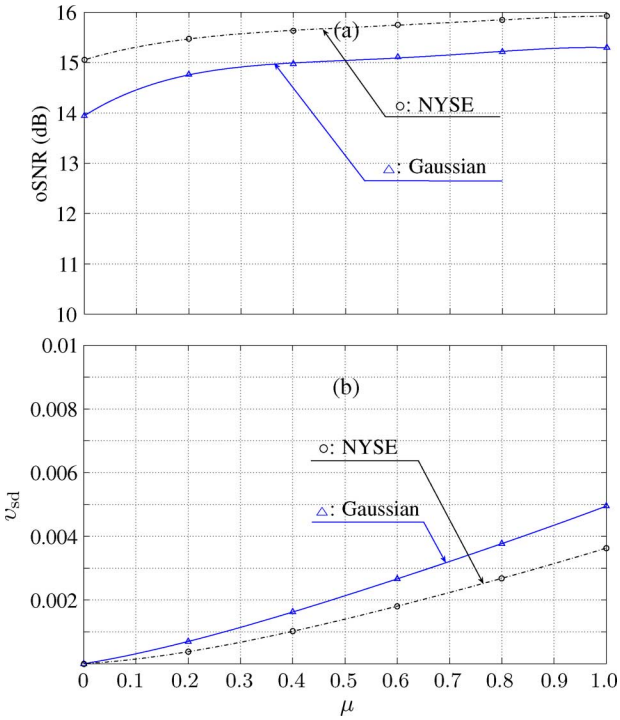

Fig. 8. Performance of the tradeoff filter as a function of the parameter μ in the white Gaussian and NYSE noise cases with $L = 40$ and $\mathrm{iSNR} = 10$ dB.

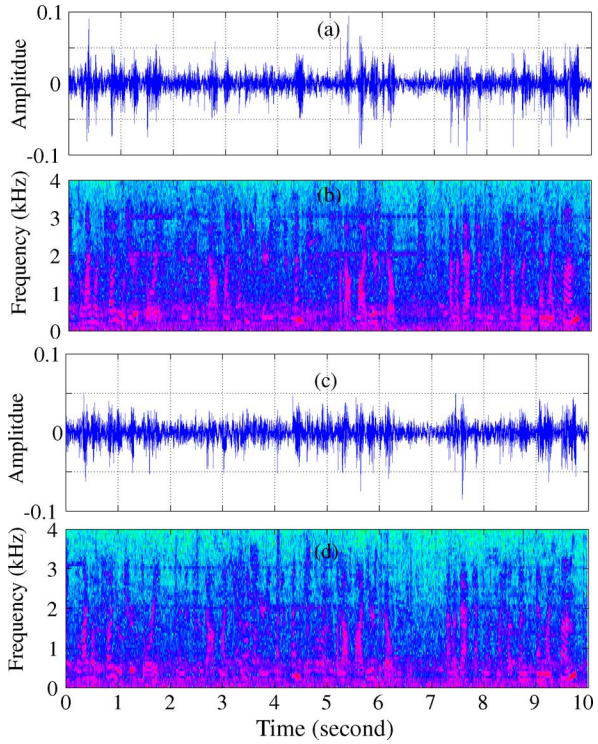

Fig. 9. First 10-s noisy signals of the stereo system and their spectrograms in the NYSE noise with $\mathrm{iSNR} = 5$ dB. (a) Left-channel noisy speech. (b) Spectrogram of the left-channel noisy speech. (c) Right-channel noisy speech. (d) Spectrogram of the right-channel noisy speech.

more difficult to deal with than the stationary white Gaussian noise. The reason for this can likely be explained as follows. 1) In white Gaussian noise, the noise signals received at the two microphones are completely uncorrelated while we noticed some small correlation between the noise signals at the two

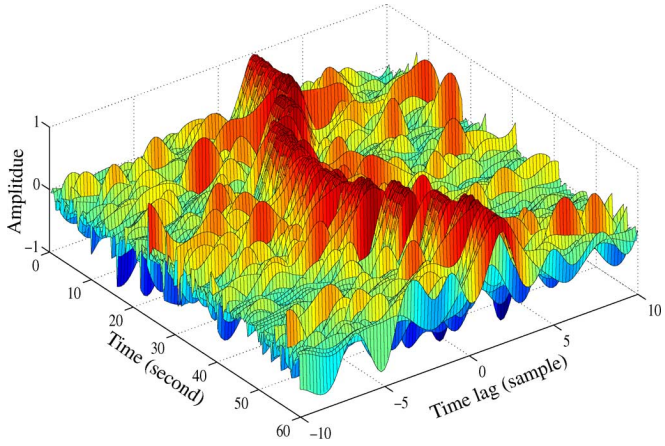

Fig. 10. Short-time cross-correlation function between the left- and right-channel noisy signals. It is computed using a short-time average with a window length of 64 ms (no overlap).

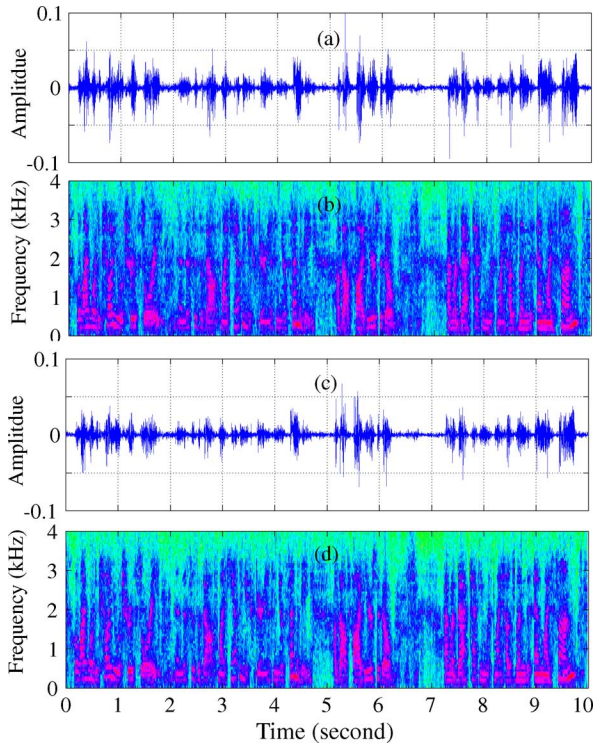

Fig. 11. First 10-s enhanced signals of the stereo system with the MVDR filter and their spectrograms in the NYSE noise with $\mathrm{iSNR} = 5$ dB. (a) Left-channel enhanced speech. (b) Spectrogram of the left-channel enhanced speech. (c) Right-channel enhanced speech. (d) Spectrogram of the right-channel enhanced speech.

microphones in the NYSE noise conditions. This correlation, though very small, may help improve binaural noise reduction. 2) The most difficult problem in dealing with babble noise is the estimation and tracking of the noise statistics. In our study, we directly computed the noise statistics from the noise signal, thereby avoiding the estimation issue. In reality, some error in noise statistics estimation is unavoidable and the error would grow as the noise becomes more nonstationary, which will subsequently lead to degradation in noise reduction performance. Once noise estimation error is taken into account, whether the developed filters can still achieve better performance in NYSE than in Gaussian noise needs further verification, but we will

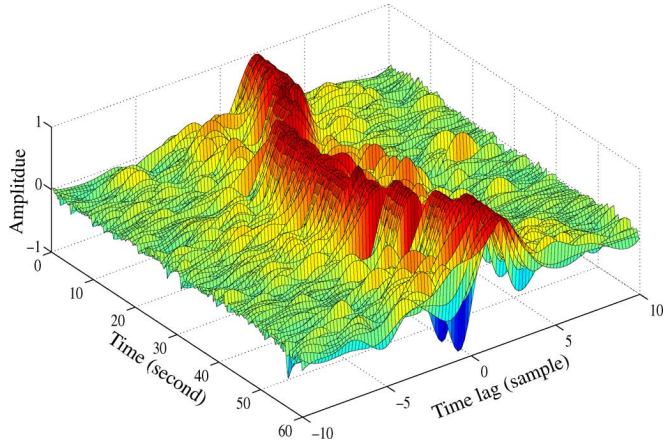

Fig. 12. Short-time cross-correlation function between the left- and right-channel enhanced signals with the MVDR filter. It is computed using a short-time average with a window length of 64 ms (no overlap).

leave this problem for the future study as the noise estimation problem is well beyond the scope of this paper.

F. Illustration of Noise Reduction Performance

We selected one set of experiments to illustrate the performance of the MVDR filter derived in Section V. The conditions for this set is $L = 40$, the input SNR is 5 dB, and the background noise is the NYSE noise. The waveforms and spectrograms of the noisy stereo signals are shown in Fig. 9 and a 3-D plot of the cross-correlation function between the two noisy outputs is shown in Fig. 10. Comparing Figs. 3 and 9, one can see that the presence of noise has significantly changed the spectra of the clean speech. Furthermore, we see that the noise has modified the sound spatial effect by comparing Figs. 4 and 10.

The enhanced stereo signals by the MVDR filter is plotted in Fig. 11 and the corresponding 3-D plot of the cross-correlation function between the enhanced stereo signals is plotted in Fig. 12. It is clearly seen that the MVDR filter has not only enhanced the speech spectrogram but also recovered the spatial effect.

VII. Conclusion

This paper focused on the binaural noise reduction problem in stereo systems that have two inputs and two outputs. By merging the two real input signals into one complex signal, we formulated the problem into a WL filtering framework. Under this new framework, we discussed some important performance measures and then derived the maximum SNR, Wiener, MVDR, and tradeoff filters. We discussed the relationship among these filters. In particular, we showed that all filters are equivalent up to a scaling factor. However, this scaling factor is in general time-varying because of speech nonstatonarity and can cause significant discontinuity in the residual noise level that is unpleasant to listen to. In order to avoid such discontinuity, it is essential to have the scaling factor right from one frame to another. For this purpose, it is recommended to use the MVDR filter in practice. We also showed that the derived filters do not only enhance the noisy speech, but also recover the spatial effects of the clean speech source.

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.
[2] M. Brandstein and D. B. Ward, Eds., Microphone Arrays: Signal Processing Techniques and Applications Berlin, Germany, Springer-Verlag, 2001.
[3] J. E. Greenberg and P. M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *J. Acoustic. Soc. Amer.*, vol. 93, pp. 1662–1676, Mar. 1992.
[4] M. Kompis and N. Dillier, "Noise reduction for hearing aids: Combining directional microphones with an adpative beamformer," *J. Acoustic. Soc. Amer.*, vol. 96, pp. 1910–1913, Sep. 1994.
[5] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-array hearing aids with binaural output–Part I: Fixed-processing systems," *IEEE Trans. Speech Audio Process.*, vol. 5, pp. 529–542, Nov. 1997.
[6] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Microphone-array hearing aids with binaural output–Part II: Two-microphone adaptive system," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 543–551, Nov. 1997.
[7] J. V. Berghe and J. Wouters, "An adaptive noise canceller for hearing aids using two nearby microphones," *J. Acoust. Soc. Amer.*, vol. 103, pp. 3621–3626, Jun. 1998.
[8] Y. Suzuki, S. Tsukui, F. Asano, R. Nishimura, and T. Sone, "New design method of a binaural microphone array using multiple constraints," *IEICE Trans. Fundamentals*, vol. E82-A, pp. 588–596, 1999.
[9] T. Lotter, B. Sauert, and P. Vary, "A stereo input-output superdirective beamformer for dual channel noise reduction," in *Proc. Eurospeech*, 2005, pp. 2285–2288.
[10] B. Kollmeier, J. Peissig, and V. Hohmann, "Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain," *Scand. Audiol. Suppl.*, vol. 38, pp. 28–38, 1993.
[11] 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.
[12] P. Vary, "Noise suppression by spectral magnitude estimation–mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, Jul. 1985.
[13] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.
[14] 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.
[15] H. Nakashima, Y. Chisaki, T. Usagawa, and M. Ebata, "Frequency domain binaural model based on interaural phase and level differences," *Acoustic. Sci. Tech.*, vol. 24, pp. 172–178, 2003.
[16] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter based on equalization-cancellation model," in *Proc. IEEE WASPAA*, 2009, pp. 133–136.
[17] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, pp. 1614–1626, Aug. 2001.
[18] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, pp. 2230–2244, Sep. 2002.
[19] J. Chen, J. Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 481–493, Mar. 2008.
[20] T. J. Klasen, M. Moonen, T. Van den Bogaert, and J. Wouters, "Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction," in *Proc. IEEE ICASSP*, 2005, vol. 3, pp. 29–32.
[21] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1579–1585, Mar. 2007.
[22] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008, ch. 47, pp. 945–978.
[23] E. Ollila, "On the circularity of a complex random variable," *IEEE Signal Process. Lett.*, vol. 15, pp. 841–844, 2008.
[24] D. P. Mandic and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. New York: Wiley, 2009.

[25] P. O. Amblard, M. Gaeta, and J. L. Lacoume, "Statistics for complex variables and signals–Part I: Variables," *Signal Process.*, vol. 53, pp. 1–13, 1996.
[26] P. O. Amblard, M. Gaeta, and J. L. Lacoume, "Statistics for complex variables and signals–Part II: Signals," *Signal Process.*, vol. 53, pp. 15–25, 1996.
[27] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.
[28] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Chichester, U.K.: Wiley, 2006.
[29] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 2030–2033, Aug. 1995.
[30] J. Benesty, J. Chen, Y. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Germany: Springer-Verlag, 2005, ch. 2, pp. 9–41.
[31] 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, no. 4, pp. 1218–1234, Jul. 2006.
[32] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[33] R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophysics*, vol. 36, pp. 661–675, Aug. 1971.
[34] A. Härmä, "Acoustic measurement data from the varechoic chamber," Agere Systems, Tech. Memo., Nov. 2001.
[35] 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 Centenn. Symp.*, 1994.
[36] A. H. Kamkar-Parsi and M. Bouchard, "Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 4, pp. 1141–1154, Apr. 2011.

Jacob Benesty was born in 1963. He received the M.S. degree in microwaves from Pierre and Marie Curie University, Paris, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, Orsay, France, in April 1991.

During the Ph.D. degree (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecomunications (CNET), Paris. 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. In May 2003, he joined INRS-EMT, University of Quebec, Montreal, QC, Canada, as a Professor. 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 and T. Gaensler conceived and designed the world-first PC-based multi-party hands-free full-duplex stereo conferencing system over IP networks. He is the editor of the book series: *Springer Topics in Signal Processing* (Springer, 2008). He has coauthored and coedited/coauthored many books in the area of acoustic signal processing. He is also the lead editor-in-chief of the reference *Springer Handbook of Speech Processing* (Springer-Verlag, 2007).

Prof. Benesty 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 was a member of the IEEE Signal Processing Society Technical Committee on Audio and Electroacoustics and a member of the editorial board of the *EURASIP Journal on Applied Signal Processing*. 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 coauthor of a paper for which Y. 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.

Jingdong Chen (SM'09) received the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences, Beijing, 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 the 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, NJ, working on acoustic signal processing for telecommunications. He subsequently joined WeVoice, Inc., Bridgewater, NJ, serving as the Chief Scientist. He is currently a Professor at Northwestern Polytechnical University, 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. He coauthored the books *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 coeditor/coauthor of the book *Speech Enhancement* (Springer-Verlag, 2005) and a section coeditor of the reference *Springer Handbook of Speech Processing* (Springer-Verlag, 2007).

Dr. Chen is currently an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 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 helped organize the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), and was the technical Co-Chair of the 2009 WASPAA. He received the 2008 Best Paper Award from the IEEE Signal Processing Society, 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 1998-1999 Japan Trust International Research Grant from the Japan Key Technology Center, the Young Author Best Paper Award from the 5th National Conference on Man–Machine Speech Communications in 1998, and the CAS (Chinese Academy of Sciences) President's Award in 1998.

Yiteng (Arden) Huang (S'97–M'01) received the B.S. degree from the Tsinghua University, Beijing, China, in 1994 and the M.S. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1998 and 2001, respectively, all in electrical and computer engineering.

From March 2001 to January 2008, he was a Member of Technical Staff at Bell Laboratories, Murray Hill, NJ. In January 2008, he joined the WeVoice, Inc., Bridgewater, NJ, and served as its CTO. His current research interests are in acoustic signal processing and multimedia communications. He is a coeditor/coauthor of the books *Noise Reduction in Speech Processing* (Springer-Verlag, 2009), *Microphone Array Signal Processing* (Springer-Verlag, 2008), *Springer Handbook of Speech Processing* (Springer-Verlag, 2007), *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006), *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Kluwer, 2004), and *Adaptive Signal Processing: Applications to Real-World Problems* (Springer-Verlag, 2003).

Dr. Huang served as an Associate Editor for the *EURASIP Journal on Applied Signal Processing* from 2004 and 2008 and for the IEEE SIGNAL PROCESSING LETTERS from 2002 to 2005. He served as a technical co-chair of the 2005 Joint Workshop on Hands-Free Speech Communication and Microphone Array and the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. He received the 2008 Best Paper Award and the 2002 Young Author Best Paper Award from the IEEE Signal Processing Society, the 2000–2001 Outstanding Graduate Teaching Assistant Award from the School Electrical and Computer Engineering, Georgia Tech, the 2000 Outstanding Research Award from the Center of Signal and Image Processing, Georgia Tech, and the 1997–1998 Colonel Oscar P. Cleaver Outstanding Graduate Student Award from the School of Electrical and Computer Engineering, Georgia Tech.