A MULTICHANNEL WIDELY LINEAR APPROACH TO BINAURAL NOISE REDUCTION USING AN ARRAY OF MICROPHONES

Jacob Benesty[1] *and Jingdong Chen*[2]

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

[2]: Northwestern Polytechnical University
127 Youyi West Road
Xi'an, Shaanxi 710072, China

ABSTRACT

This paper deals with the problem of binaural noise reduction using an array of microphones. This is a very important problem in applications such as teleconferencing and hearing aids where there is a need to mitigate the noise effect from the noisy signals picked up by multiple microphones and produce two "clean" outputs. The mitigation of the noise should be made in such a way that no audible distortion is added to the two outputs (this is the same as in the single-channel case) and meanwhile the spatial information of the desired sound source should be preserved so that, after noise reduction, the listener will still be able to localize the sound source thanks to his/her binaural hearing mechanism. In this paper, we present a novel approach to this problem where we first form a number of complex input signals from the multiple and real microphone observations. We also merge the two expected real outputs into a complex output signal. The widely linear estimation theory is then used to derive optimal noise reduction filters that can achieve noise reduction while preserving the desired signal (speech) and its spatial information. With this new formulation, the Wiener and minimum variance distortionless response (MVDR) filters are derived. Experiments are provided to justify the effectiveness of these filters.

Index Terms— Binaural noise reduction, microphone arrays, widely linear (WL) estimation, Wiener Filter, minimum variance distortionless response (MVDR) filter.

1. SIGNAL MODEL AND PROBLEM FORMULATION

Binaural noise reduction has a wide range of applications in areas such as teleconferencing, telecollaboration, social networks, and hearing aids. It has emerged as a very important research and engineering problem over the last two decades. Traditionally, this problem is tackled by extending the single-channel spectral modification based noise reduction technique to the binaural case by either posing some constraints on the suppression of each frequency band or using head-related transfer functions (HRTFs) to preserve the spatial information [1], [2]. Recently, we developed a new technique, which works for a stereo input and binaural output system [3]. The basic idea is to form a complex input signal from the stereo inputs and a complex output signal from the two expected real output channels. By doing so, the binaural noise reduction problem can be processed by a single-channel widely linear filter. In this paper, we attempt to extend the basic principle shown in [3] to a more general case where an array of microphones is used. Without loss of generality, we consider the signal model in which $2N$ microphones[1] capture a source signal convolved with acoustic impulse responses in some noise field. The signal received at the ith microphone is then expressed as

$$\begin{aligned} y_{\mathrm{r},i}(t) &= g_i(t) * s(t) + v_{\mathrm{r},i}(t) \\ &= x_{\mathrm{r},i}(t) + v_{\mathrm{r},i}(t),\ i = 1, 2, \ldots, 2N, \end{aligned} \tag{1}$$

where $g_i(t)$ is the acoustic impulse response from the unknown speech source, $s(t)$, location to the ith microphone, $*$ stands for linear convolution, and $x_{\mathrm{r},i}(t)$ and $v_{\mathrm{r},i}(t)$ are, respectively, the convolved speech and additive noise received at microphone i. We assume that the impulse responses are time invariant. We also assume that the signals $x_{\mathrm{r},i}(t) = g_i(t) * s(t)$ and $v_{\mathrm{r},i}(t)$ are uncorrelated, zero mean, real, and broadband.

In binaural noise reduction, it is desired to simultaneously recover the convolved speech signals at two microphones. In this paper, we consider recovering the signals $x_{\mathrm{r},1}(t)$ and $x_{\mathrm{r},N+1}(t)$ given the observations $y_{\mathrm{r},i}(t),\ i = 1, 2, \ldots, 2N$. This means that the desired signals in our problem are the speech signals received at the first and $(N+1)$th microphones[2]. It is clear then that we have two objectives. The first one is to attenuate the contribution of the noise terms $v_{\mathrm{r},1}(t)$ and $v_{\mathrm{r},N+1}(t)$ as much as possible. The second objective is to preserve $x_{\mathrm{r},1}(t)$ and $x_{\mathrm{r},N+1}(t)$ 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 $s(t)$. This is the well-known problem of binaural noise reduction.

We have $2N$ real input and two real output signals. It is convenient, however, to work in the complex domain in order that the original binaural problem is transformed to a single-output noise reduction problem with a microphone array. Indeed, from the $2N$ real microphone signals given in (1), we can form N complex microphone signals as

$$\begin{aligned} y_n(t) &= y_{\mathrm{r},n}(t) + j y_{\mathrm{r},N+n}(t) \\ &= x_n(t) + v_n(t),\ n = 1, 2, \ldots, N, \end{aligned} \tag{2}$$

where $j = \sqrt{-1}$,

$$x_n(t) = x_{\mathrm{r},n}(t) + j x_{\mathrm{r},N+n}(t),\ n = 1, 2, \ldots, N \tag{3}$$

is the complex convolved speech signal, and

$$v_n(t) = v_{\mathrm{r},n}(t) + j v_{\mathrm{r},N+n}(t),\ n = 1, 2, \ldots, N \tag{4}$$

is the complex additive noise. Now, our problem can be stated as follows: given the N complex microphone signals, $y_n(t),\ n = 1, 2, \ldots, N$, which are a mixture of the uncorrelated complex signals $x_n(t)$ and $v_n(t)$, our goal is to recover $x_1(t) = x_{\mathrm{r},1}(t) + j x_{\mathrm{r},N+1}(t)$ (i.e., our desired signal) the best way we can, including the phase, which is important for the localization of the source signal.

Effort of the second author is partially supported by the Anhui Science and Technology Project (11010202191).

[1]The generalization to an odd number of microphones is straightforward.

[2]Note that the ideas and algorithms developed in this paper can be applied to any other pair of microphones.

The signal model given in (2) can be put into a vector form as

$$\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{v}(t), \tag{5}$$

where

$$\mathbf{y}(t) = \left[\begin{array}{cccccc} \mathbf{y}_1^T(t) & \mathbf{y}_2^T(t) & \cdots & \mathbf{y}_n^T(t) & \cdots & \mathbf{y}_N^T(t) \end{array} \right]^T,$$

the superscript ^{T} is the transpose operator,

$$\mathbf{y}_n(t) = \left[\begin{array}{cccc} y_n(t) & y_n(t-1) & \cdots & y_n(t-L+1) \end{array} \right]^T$$

is a vector of length L, and $\mathbf{x}(t)$ and $\mathbf{v}(t)$ are defined in a similar way to $\mathbf{y}(t)$. Since the clean and noise signals are uncorrelated by assumption, the correlation matrix (of size $NL \times NL$) of the noisy signal is

$$\mathbf{\Phi}_{\mathbf{y}} = E\left[\mathbf{y}(t)\mathbf{y}^H(t)\right] = \mathbf{\Phi}_{\mathbf{x}} + \mathbf{\Phi}_{\mathbf{v}}, \tag{6}$$

where the superscript ^{H} is the transpose-conjugate operator, and $\mathbf{\Phi}_{\mathbf{x}} = \left[\mathbf{x}(t)\mathbf{x}^H(t)\right]$ and $\mathbf{\Phi}_{\mathbf{v}} = E\left[\mathbf{v}(t)\mathbf{v}^H(t)\right]$ are the correlation matrices of $\mathbf{x}(t)$ and $\mathbf{v}(t)$, respectively.

2. WIDELY LINEAR FILTERING

As seen 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) [4], [5]. 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^* [6], [7], 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 [4] 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 [4] defined as the ratio between the pseudo-variance and the variance, i.e.,

$$\gamma_z = \frac{E\left(z^2\right)}{E\left(|z|^2\right)}. \tag{7}$$

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.

Now, let us examine whether the complex desired signal, $x_1(t) = x_{\mathrm{r},1}(t) + jx_{\mathrm{r},N+1}(t)$, is second-order circular or not. We have

$$\begin{aligned} \gamma_{x_1} &= \frac{E\left[x_1^2(t)\right]}{E\left[|x_1(t)|^2\right]} \\ &= \frac{E\left[x_{\mathrm{r},1}^2(t)\right] - E\left[x_{\mathrm{r},N+1}^2(t)\right] + j2E\left[x_{\mathrm{r},1}(t)x_{\mathrm{r},N+1}(t)\right]}{\phi_{x_1}}, \end{aligned} \tag{8}$$

where $\phi_{x_1} = E\left[|x_1(t)|^2\right]$ is the variance of $x_1(t)$. One can check from (8) that the CRV $x_1(t)$ is second-order circular (i.e., $\gamma_{x_1} = 0$) if and only if

$$E\left[x_{\mathrm{r},1}^2(t)\right] = E\left[x_{\mathrm{r},N+1}^2(t)\right] \text{ and } E\left[x_{\mathrm{r},1}(t)x_{\mathrm{r},N+1}(t)\right] = 0. \tag{9}$$

Since the signals $x_{\mathrm{r},1}(t)$ and $x_{\mathrm{r},N+1}(t)$ come from the same source, they are in general correlated. As a result, the second condition in (9) should not be true. Therefore, we can safely state that the complex desired signal, $x_1(t)$, is noncircular, and so is the complex microphone signal, $y_1(t)$. If the noise terms at the two microphones are assumed to be uncorrelated and have the same power, then $\gamma_{v_1} = 0$ [i.e., $v(t)$ is a second-order CCRV].

Since we deal with noncircular CRVs as demonstrated above, the classical linear estimation technique [8], which is developed for processing real signals or CCRVs, cannot be applied. Instead, an estimate of $x_1(t)$ should be obtained using the widely linear (WL) estimation theory as [5], [9]

$$\widehat{x}_1(t) = \mathbf{h}^H\mathbf{y}(t) + \mathbf{h}'^H\mathbf{y}^*(t) = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{y}}(t), \tag{10}$$

where $\mathbf{h}$ and $\mathbf{h}'$ are two complex FIR filters of length NL and

$$\widetilde{\mathbf{h}} = \left[\begin{array}{c} \mathbf{h} \\ \mathbf{h}' \end{array} \right] \tag{11}$$

$$\widetilde{\mathbf{y}}(t) = \left[\begin{array}{c} \mathbf{y}(t) \\ \mathbf{y}^*(t) \end{array} \right] \tag{12}$$

are the augmented WL filter and observation vector, respectively, both of length $2NL$. We can rewrite (10) as

$$\widehat{x}_1(t) = \widetilde{\mathbf{h}}^H\left[\widetilde{\mathbf{x}}(t) + \widetilde{\mathbf{v}}(t)\right] = x_{\mathrm{f}}(t) + v_{\mathrm{rn}}(t), \tag{13}$$

where $\widetilde{\mathbf{x}}(t)$ and $\widetilde{\mathbf{v}}(t)$ are defined in a similar way to $\widetilde{\mathbf{y}}(t)$,

$$x_{\mathrm{f}}(t) = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{x}}(t) \tag{14}$$

is a filtered version of the desired signal vector and its conjugate, and

$$v_{\mathrm{rn}}(t) = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{v}}(t) \tag{15}$$

is the residual noise.

From (13), we see that $\widehat{x}_1(t)$ depends on the vector $\widetilde{\mathbf{x}}(t)$. However, our desired signal at time t is not the whole vector $\widetilde{\mathbf{x}}(t)$ but only the sample $x_1(t)$; so we should decompose the vector $\widetilde{\mathbf{x}}(t)$ into two orthogonal vectors: one correlated and the other uncorrelated with the desired signal sample, $x_1(t)$. This decomposition is given as

$$\widetilde{\mathbf{x}}(t) = x_1(t)\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} + \widetilde{\mathbf{x}}_{\mathrm{i}}(t) = \widetilde{\mathbf{x}}_{\mathrm{d}}(t) + \widetilde{\mathbf{x}}_{\mathrm{i}}(t), \tag{16}$$

where

$$\widetilde{\mathbf{x}}_{\mathrm{d}}(t) = x_1(t)\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \tag{17}$$

is the desired signal vector,

$$\widetilde{\mathbf{x}}_{\mathrm{i}}(t) = \widetilde{\mathbf{x}}(t) - \widetilde{\mathbf{x}}_{\mathrm{d}}(t) \tag{18}$$

is the interference signal vector,

$$\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} = \frac{E\left[\widetilde{\mathbf{x}}(t)x_1^*(t)\right]}{E\left[|x_1(t)|^2\right]} \tag{19}$$

is the normalized [with respect to $x_1(t)$] correlation vector between $\widetilde{\mathbf{x}}(t)$ and $x_1(t)$, and

$$E\left[\widetilde{\mathbf{x}}_{\mathrm{i}}(t)x_1^*(t)\right] = \mathbf{0}_{2NL\times 1}. \tag{20}$$

Substituting (16) into (13), we obtain

$$\begin{aligned} \widehat{x}_1(t) &= \widetilde{\mathbf{h}}^H\left[\widetilde{\mathbf{x}}_{\mathrm{d}}(t) + \widetilde{\mathbf{x}}_{\mathrm{i}}(t) + \widetilde{\mathbf{v}}(t)\right] \\ &= x_{\mathrm{fd}}(t) + x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t), \end{aligned} \tag{21}$$

where

$$x_{\mathrm{fd}}(t) = \widetilde{\mathbf{h}}^H\widetilde{\mathbf{x}}_{\mathrm{d}}(t) = x_1(t)\widetilde{\mathbf{h}}^H\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1} \tag{22}$$

is the filtered desired signal and

$$x_{\mathrm{ri}}(t) = \widetilde{\mathbf{h}}^H \widetilde{\mathbf{x}}_{\mathrm{i}}(t) \tag{23}$$

is the residual interference. We observe that the estimate of the desired signal at time t is the sum of three terms that are mutually uncorrelated. Therefore, the variance of $\widehat{x}_1(t)$ is

$$\phi_{\widehat{x}_1} = \phi_{x_{\mathrm{fd}}} + \phi_{x_{\mathrm{ri}}} + \phi_{v_{\mathrm{rn}}}, \tag{24}$$

where

$$\phi_{x_{\mathrm{fd}}} = \phi_{x_1} \left| \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \right|^2 = \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{d}}} \widetilde{\mathbf{h}}, \tag{25}$$

$$\phi_{x_{\mathrm{ri}}} = \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{i}}} \widetilde{\mathbf{h}} = \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}} \widetilde{\mathbf{h}} - \phi_{x_1} \left| \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \right|^2, \tag{26}$$

$$\phi_{v_{\mathrm{rn}}} = \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} \widetilde{\mathbf{h}}, \tag{27}$$

$\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{d}}} = \phi_{x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H$ is the correlation matrix (whose rank is equal to 1) of $\widetilde{\mathbf{x}}_{\mathrm{d}}(t)$, and $\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{i}}} = E\left[\widetilde{\mathbf{x}}_{\mathrm{i}}(t) \widetilde{\mathbf{x}}_{\mathrm{i}}^H(t)\right]$, $\boldsymbol{\Phi}_{\widetilde{\mathbf{x}}} = E\left[\widetilde{\mathbf{x}}(t) \widetilde{\mathbf{x}}^H(t)\right]$, and $\boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} = E\left[\widetilde{\mathbf{v}}(t) \widetilde{\mathbf{v}}^H(t)\right]$ are the correlation matrices of $\widetilde{\mathbf{x}}_{\mathrm{i}}(t)$, $\widetilde{\mathbf{x}}(t)$, and $\widetilde{\mathbf{v}}(t)$, respectively.

It is clear from (21) that the objective of our noise reduction problem is to find optimal filters that can minimize the effect of $x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t)$ while preserving the desired signal, $x_1(t)$.

3. OPTIMAL FILTERS

Before deriving the optimal filters, we first define the mean-square-error (MSE) criterion.

The error signal between the estimated and desired signals is defined as

$$e(t) = \widehat{x}_1(t) - x_1(t) = x_{\mathrm{fd}}(t) + x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t) - x_1(t). \tag{28}$$

We can rewrite (28) as

$$e(t) = e_{\mathrm{d}}(t) + e_{\mathrm{r}}(t), \tag{29}$$

where

$$e_{\mathrm{d}}(t) = x_{\mathrm{fd}}(t) - x_1(t) = \left(\widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} - 1 \right) x_1(t) \tag{30}$$

is the speech distortion due to the WL filter, and

$$e_{\mathrm{r}}(t) = x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t) = \widetilde{\mathbf{h}}^H \widetilde{\mathbf{x}}_{\mathrm{i}}(t) + \widetilde{\mathbf{h}}^H \widetilde{\mathbf{v}}(t) \tag{31}$$

represents the residual interference-plus-noise. The two error signals $e_{\mathrm{d}}(t)$ and $e_{\mathrm{r}}(t)$ are clearly uncorrelated.

The classical MSE criterion is then

$$\begin{aligned} J\left(\widetilde{\mathbf{h}}\right) &= E\left[|e(t)|^2\right] = E\left[|e_{\mathrm{d}}(t)|^2\right] + E\left[|e_{\mathrm{r}}(t)|^2\right] \\ &= \phi_{x_1} \left| \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} - 1 \right|^2 + \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{i}}} \widetilde{\mathbf{h}} + \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} \widetilde{\mathbf{h}} \\ &= J_{\mathrm{d}}\left(\widetilde{\mathbf{h}}\right) + J_{\mathrm{r}}\left(\widetilde{\mathbf{h}}\right), \end{aligned} \tag{32}$$

where

$$J_{\mathrm{d}}\left(\widetilde{\mathbf{h}}\right) = E\left[|e_{\mathrm{d}}(t)|^2\right] = \phi_{x_1} \left| \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} - 1 \right|^2, \tag{33}$$

$$J_{\mathrm{r}}\left(\widetilde{\mathbf{h}}\right) = E\left[|e_{\mathrm{r}}(t)|^2\right] = \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}}, \tag{34}$$

and

$$\boldsymbol{\Phi}_{\mathrm{in}} = \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}_{\mathrm{i}}} + \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} = \boldsymbol{\Phi}_{\widetilde{\mathbf{y}}} - \phi_{x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H \tag{35}$$

is the correlation matrix of the interference-plus-noise.

Given the above MSE criteria, the filters that can achieve noise reduction can be derived by either directly minimizing $J\left(\widetilde{\mathbf{h}}\right)$ or minimizing $J_{\mathrm{d}}\left(\widetilde{\mathbf{h}}\right)$ or $J_{\mathrm{r}}\left(\widetilde{\mathbf{h}}\right)$ with some constraints.

3.1. WL Wiener Filter

In our problem, the WL Wiener filter can be obtained by taking the gradient of (32) with respect to $\widetilde{\mathbf{h}}$ and equating the result to zero. We easily get

$$\widetilde{\mathbf{h}}_{\mathrm{W}} = \boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}^{-1} \boldsymbol{\Phi}_{\widetilde{\mathbf{x}}} \widetilde{\mathbf{i}} = \left[\mathbf{I}_{2NL} - \boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}^{-1} \boldsymbol{\Phi}_{\widetilde{\mathbf{v}}} \right] \widetilde{\mathbf{i}}, \tag{36}$$

where $\widetilde{\mathbf{i}}$ is the first column of the identity matrix $\mathbf{I}_{2NL}$ of size $2NL \times 2NL$. From (35), we can write the inverse of $\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}$ according to the Woodbury's identity:

$$\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}^{-1} = \boldsymbol{\Phi}_{\mathrm{in}}^{-1} - \frac{\boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H \boldsymbol{\Phi}_{\mathrm{in}}^{-1}}{\phi_{x_1}^{-1} + \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H \boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}}. \tag{37}$$

Substituting (37) into (36), we get an alternative form of the WL Wiener filter, i.e.,

$$\tilde{\mathbf{h}}_{\mathrm{W}} = \frac{\boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}}{\phi_{x_1}^{-1} + \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H \boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}}, \tag{38}$$

which can be more useful for practical implementation as the matrix $\boldsymbol{\Phi}_{\mathrm{in}}$ is generally better conditioned than $\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}$ in real-world applications.

3.2. WL MVDR Filter

The WL MVDR filter can be derived by minimizing either $E\left[|\widehat{x}_1(t)|^2\right]$, $J\left(\widetilde{\mathbf{h}}\right)$, or $J_{\mathrm{r}}\left(\widetilde{\mathbf{h}}\right)$ subject to the constraint that $J_{\mathrm{d}}\left(\widetilde{\mathbf{h}}\right) = 0$. Using $J\left(\widetilde{\mathbf{h}}\right)$, the problem can be mathematically written as

$$\min_{\widetilde{\mathbf{h}}} J\left(\widetilde{\mathbf{h}}\right) \quad \text{subject to} \quad \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} = 1. \tag{39}$$

Using a Lagrange multiplier to adjoin the constraint to the cost function, taking the gradient with respect to $\widetilde{\mathbf{h}}$, and then equating the result to zero, we obtain

$$\widetilde{\mathbf{h}}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}}{\boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H \boldsymbol{\Phi}_{\widetilde{\mathbf{y}}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}}. \tag{40}$$

Now, using $J_{\mathrm{r}}\left(\widetilde{\mathbf{h}}\right)$, the problem can be written into the following form:

$$\min_{\widetilde{\mathbf{h}}} \widetilde{\mathbf{h}}^H \boldsymbol{\Phi}_{\mathrm{in}} \widetilde{\mathbf{h}} \quad \text{subject to} \quad \widetilde{\mathbf{h}}^H \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1} = 1, \tag{41}$$

for which the solution is

$$\tilde{\mathbf{h}}_{\mathrm{MVDR}} = \frac{\boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}}{\boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}^H \boldsymbol{\Phi}_{\mathrm{in}}^{-1} \boldsymbol{\rho}_{\widetilde{\mathbf{x}} x_1}}. \tag{42}$$

It is easy to check that the two forms of the MVDR filter in (40) and (42) are identical.

4. EXPERIMENTAL EVALUATION

Experiments were conducted with the speech source recorded in a reverberant but quiet room. The room is 6 m long and 5 m wide. For ease of exposition, positions in the room are designated by (x, y) coordinates with reference to one corner of the room, $0 \leq x \leq 6$ and $0 \leq y \leq 5$. An equispaced linear array with six omnidirectional microphones is configured where the first and last microphones are, respectively, at (3.4, 0.5) and (3.9, 0.5), and spacing between two neighboring microphones is 0.1 m. A female talker reads a story

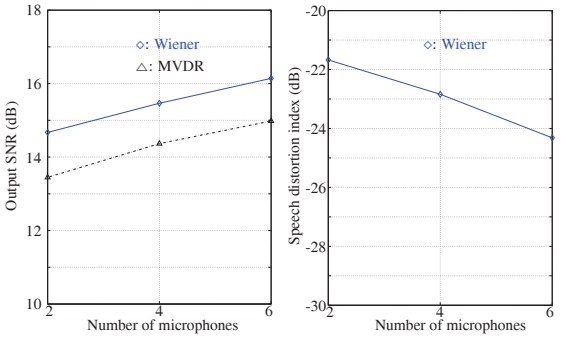

Fig. 1. The output SNR and speech distortion index as a function of the number of microphones for the Wiener and MVDR filters in white noise: $L = 40$ and input SNR of 10 dB. Note that the values of the speech distortion index for the MVDR filter are not displayed because they are very small.

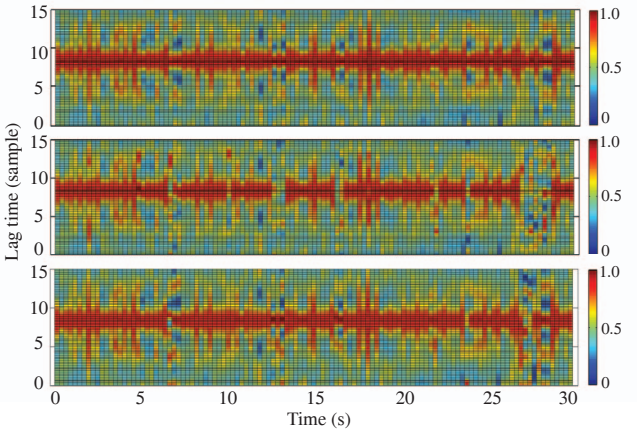

Fig. 2. The contours of the short-time cross-correlation coefficients between the left and right channels. Upper trace: clean speech; middle trace: noisy speech (with an input SNR of 10 dB); lower trace: enhanced speech using the MVDR filter with 6 microphones.

while standing at (1.3, 3.0) and her voice is recorded with a sampling rate of 8 kHz. The recorded multichannel signals are treated as the convolved speech. White Gaussian noise is then added to the signals so that the input SNR is equal to 10 dB. The overall length of the signal is 30 s. We set L to 40 based on the results from a previous study [3].

We choose to implement the WL Wiener and MVDR filters according to (38) and (42), respectively, for which we need to know the parameters $\mathbf{\Phi}_{\text{in}}$, ϕ_{x_1}, and $\boldsymbol{\rho}_{\widetilde{\mathbf{x}}x_1}$. In this paper, we first compute the two covariance matrices $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$ and $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}$ directly from the noisy and noise signals using a short-time average. Specifically, at each time instant t, an estimate of $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}$, i.e., $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{y}}}(t)$, is computed using the most recent 480 $\widetilde{\mathbf{y}}$ vectors, and $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t)$ is computed in the same manner but from the noise signal. An estimate of $\mathbf{\Phi}_{\widetilde{\mathbf{x}}}$ is calculated as $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}}(t) = \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{y}}}(t) - \hat{\mathbf{\Phi}}_{\widetilde{\mathbf{v}}}(t)$. Then $\hat{\phi}_{x_1}(t)$ is taken as the first element of $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}}(t)$ and $\hat{\boldsymbol{\rho}}_{\widetilde{\mathbf{x}}x_1}(t)$ is the first column of $\hat{\mathbf{\Phi}}_{\widetilde{\mathbf{x}}}(t)$ normalized by $\hat{\phi}_{x_1}(t)$. The matrix $\hat{\mathbf{\Phi}}_{\text{in}}(t)$ can then be computed according to (35).

Substituting $\hat{\mathbf{\Phi}}_{\text{in}}(t)$, $\hat{\phi}_{x_1}(t)$, and $\hat{\boldsymbol{\rho}}_{\widetilde{\mathbf{x}}x_1}(t)$ into (38) and (42), we implemented the WL Wiener and MVDR filters. To evaluate the noise reduction performance of these two filters, we used the output SNR and speech distortion index, which are defined as [3]:

$$\text{oSNR}\left(\tilde{\mathbf{h}}\right) = \frac{\phi_{x_{\text{fd}}}}{\phi_{x_{\text{ri}}} + \phi_{v_{\text{rn}}}}, \tag{43}$$

$$v_{\text{sd}}\left(\tilde{\mathbf{h}}\right) = \frac{E\left[\left|x_{\text{fd}}(t) - x_1(t)\right|^2\right]}{\phi_{x_1}}. \tag{44}$$

These two measures are computed using a long-time average. The results, as a function of the number of microphones, are plotted in Fig. 1. (Note that for a fair comparison, in the two-microphone case, we used the 1st and 4th microphones, and in the four-microphone case, we used the 1st, 2nd, 4th, and 5th microphones.) For both the WL Wiener and MVDR filters, we see that the output SNR increases while the speech distortion index decreases with the number of microphones. This clearly demonstrates the advantage of using multiple microphones for binaural noise reduction. It is also seen that the WL MVDR can improve SNR without adding speech distortion, though its SNR improvement is less than that of the WL Wiener filter.

To visualize the spatial sound effect, we computed the cross-correlation function between the two real output channels [i.e., the real and imaginary parts of $\widehat{x}_1(t)$] every 64 ms using a short-time average with a frame size of 64 ms. The location of the maximum value of this function indicates the position of the talker at time instant t. Figure 2 shows the contours of the time-varying cross-correlation function of the clean and noisy signals as well as that of the enhanced signal (due to space limitation, we only show the result with the MVDR filter using 6 microphones). One can notice that the noise has dramatically modified the sound spatial effect. It is clearly seen that the MVDR filter with 6 microphones has significantly recovered the spatial effect.

5. SUMMARY

In this paper, we presented a multichannel widely linear (WL) approach to the problem of binaural noise reduction using a microphone array. We first formed a number of complex input signals from the multiple and real microphone observations. We also merged the two expected real outputs into a complex output signal. By doing so, the binaural noise reduction problem is converted into one of multiple-input/single-output WL filtering. The WL estimation theory was then used to derive the optimal WL Wiener and MVDR filters that can achieve noise reduction while preserving the desired signal (speech) and its spatial information. Experimental results have justified the effectiveness of these filters.

6. REFERENCES

[1] 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.

[2] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener fitler based on equalization-cancellation model," in *Proc. IEEE WASPAA*, 2009, pp. 133–136.

[3] J. Benesty, J. Chen, and Y. Huang, "Binaural noise reduction in the time domain with a stereo setup," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2260–2272, Nov. 2011.

[4] E. Ollila, "On the circularity of a complex random variable," *IEEE Signal Process. Lett.*, vol. 15, pp. 841–844, 2008.

[5] D. P. Mandic and S. L. Goh, *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. Wiley, 2009.

[6] 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.

[7] 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.

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

[9] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Process.*, vol. 43, pp. 2030–2033, Aug. 1995.