A MULTICHANNEL WIDELY LINEAR WIENER FILTER FOR BINAURAL NOISE REDUCTION IN THE SHORT-TIME-FOURIER-TRANSFORM DOMAIN

Liheng Zhao†, Jingdong Chen‡, and Jacob Benesty†

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

‡: Northwestern Polytechnical University
127 Youyi West Road
Xi'an, Shaanxi 710072, China

ABSTRACT

Binaural noise reduction is a very challenging problem since it requires not only to reduce noise, but also to recover the spatial information of the desired speech source so that the listener can localize this source from the binaural outputs. In this paper, we study the problem in the short-time-Fourier-transform (STFT) domain with the use of an array of microphones. Combining the multichannel microphone observations into a number of complex signals and merging the two (binaural) expected output channels into a complex signal, we reformulate the problem with the widely linear (WL) estimation technique. To efficiently achieve the optimal estimation, the complex signals are transformed into the frequency domain via the STFT. We then derive a WL Wiener filter based on the WL estimation theory and the mean-squared-error (MSE) criterion. This WL Wiener filter is shown to be able to exploit the noncircularity of the complex speech signals and the spatial information captured by the microphone array to achieve noise reduction while preserving the sound spatial information.

Index Terms— Binaural noise reduction, microphone array, STFT domain, widely linear, Wiener filter.

1. INTRODUCTION

Binaural noise reduction consists of processing multiple noisy microphone signals to produce two (binaural) outputs with less amount of noise and (hopefully) the same spatial information of the desired sound source. Traditionally, this problem is tackled through the use of either beamforming [1]–[9] or approaches of multichannel noise reduction [10]–[12]. Recently, we developed a WL approach to binaural noise reduction in the time domain [13]–[16]. Basically, we first convert the original multiple input and binaural output system into a complex multiple-input/single-output (MISO) one by merging the real, multichannel input noisy signals and the two (binaural) expected output channels into a number of complex signals. Under the WL filtering framework, the optimal WL filters are then derived that can reduce noise and recover the source spatial information at the same time. Although they have shown great potentials, those time-domain WL filters are computationally expensive to implement. This work is basically an extension of our effort in [13]–[16]. The contribution of this paper is an STFT-domain formulation of the binaural noise reduction problem and a WL Wiener filter. This STFT-domain WL Wiener filter is shown to be able to exploit the noncircularity of the complex speech signals and the spatial information captured by the microphone array to achieve noise reduction and preserve the sound spatial information at the same time. In comparison with its time-domain counterpart, this WL Wiener filter is more efficient to implement in terms of computation thanks to the use of FFT and more flexible to tune the noise reduction performance.

2. SIGNAL MODEL

We consider the scenario where there is a single desired source in an acoustic sound field and we use an array of microphones to pick up the desired signal. For ease of exposition, we assume that the number of microphones is $2N$ with $N \geq 1$ (the generalization to an odd number of microphones is trivial). The signal received at the ith microphone at the discrete-time index t can be 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 (note that the ideas and algorithms developed in this paper can be applied to recovering the clean speech signals at any other pair of microphones). Then, it is clear 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, $s(t)$.

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 one like the monaural noise reduction. Indeed, from the $2N$ real microphone signals given in (1), we can form N complex microphone signals as

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

where $\jmath$ is the imaginary unit with $\jmath^2 = -1$,

$$x_n(t) \triangleq x_{\mathrm{r},n}(t) + \jmath x_{\mathrm{r},N+n}(t) \tag{3}$$

and

$$v_n(t) \triangleq v_{\mathrm{r},n}(t) + \jmath v_{\mathrm{r},N+n}(t) \tag{4}$$

are the complex convolved speech signal and complex additive noise, respectively, at the nth complex microphone. Our noise reduction problem can then be stated as follows: given the N complex

978-1-4799-5403-2/14/$31.00 ©2014 IEEE

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 $x_1(t) = x_{\mathrm{r},1}(t) + \jmath x_{\mathrm{r},N+1}(t)$ (i.e., our desired signal) the best way we can, including the phase, which is important for the localization of the source signal.

To make the noise reduction process efficient, let us work in the short-time-Fourier-transform (STFT) domain, in which the signal model in (2) is rewritten as

$$Y_n(k,m) = X_n(k,m) + V_n(k,m),\ n = 1, 2, \ldots, N, \quad (5)$$

where $Y_n(k,m)$, $X_n(k,m)$, and $V_n(k,m)$ are the STFTs of the complex signals $y_n(t)$, $x_n(t)$, and $v_n(t)$, respectively, at frequency bin $k \in \{0, 1, \ldots, K\}$ and time frame m. Putting the N complex signals $Y_n(k,m)$, $n = 1, 2, \ldots, N$, in (5) into a vector, we have

$$\begin{aligned}\mathbf{y}(k,m) &\triangleq \begin{bmatrix} Y_1(k,m) & Y_2(k,m) & \cdots & Y_N(k,m) \end{bmatrix}^T \\ &= \mathbf{x}(k,m) + \mathbf{v}(k,m), \end{aligned} \quad (6)$$

where $\mathbf{x}(k,m)$ and $\mathbf{v}(k,m)$ are defined in a similar way to $\mathbf{y}(k,m)$, and ^{T} denotes the transpose of a vector or a matrix. Now, the problem of binaural noise reduction in the STFT domain becomes one of estimating $X_1(k,m)$ given $\mathbf{y}(k,m)$.

Since $x_n(t)$ and $v_n(t)$ are uncorrelated by assumption, the variance of $Y_n(k,m)$ can be written as

$$\phi_{Y_n}(k,m) \triangleq E\left[|Y_n(k,m)|^2\right] = \phi_{X_n}(k,m) + \phi_{V_n}(k,m), \quad (7)$$

where $E[\cdot]$ stands for mathematical expectation, and $\phi_{X_n}(k,m) \triangleq E\left[|X_n(k,m)|^2\right]$ and $\phi_{V_n}(k,m) \triangleq E\left[|V_n(k,m)|^2\right]$ are the variances of $X_n(k,m)$ and $V_n(k,m)$, respectively.

3. WIDELY LINEAR (WL) MODEL

We deal with complex random variables (CRVs) in this study as seen in the signal models given in (2) and (5). It was shown in [13]–[16] that the CRVs in (2) are noncircular[1], and so are the signals in (5). Since we deal with noncircular CRVs, we have to employ the so-called widely linear (WL) estimation theory [13]–[18] to estimate the desired signal, $X_1(k,m)$, thereby achieving noise reduction. Specifically, an estimate of $X_1(k,m)$ should be obtained from the noisy signal vector, $\mathbf{y}(k,m)$, according to

$$\begin{aligned}\widehat{X}_1(k,m) &= \mathbf{h}^H(k,m)\mathbf{y}(k,m) + \mathbf{h}'^H(k,m)\mathbf{y}^*(k,m) \\ &= \widetilde{\mathbf{h}}^H(k,m)\widetilde{\mathbf{y}}(k,m), \end{aligned} \quad (8)$$

where the superscripts ^{H} and * are the conjugate-transpose and conjugate operators, respectively, $\mathbf{h}(k,m)$ and $\mathbf{h}'(k,m)$ are two complex finite-impulse-response (FIR) filters of length N, and

$$\widetilde{\mathbf{h}}(k,m) \triangleq \begin{bmatrix} \mathbf{h}(k,m) \\ \mathbf{h}'(k,m) \end{bmatrix} \quad (9)$$

and

$$\widetilde{\mathbf{y}}(k,m) \triangleq \begin{bmatrix} \mathbf{y}(k,m) \\ \mathbf{y}^*(k,m) \end{bmatrix} \quad (10)$$

are the augmented WL filter and noisy signal vector, respectively, both of length $2N$. Substituting (6) into (8) gives

$$\begin{aligned}\widehat{X}_1(k,m) &= \widetilde{\mathbf{h}}^H(k,m)\left[\widetilde{\mathbf{x}}(k,m) + \widetilde{\mathbf{v}}(k,m)\right] \\ &= X_{\mathrm{f}}(k,m) + V_{\mathrm{rn}}(k,m), \end{aligned} \quad (11)$$

where $\widetilde{\mathbf{x}}(k,m)$ and $\widetilde{\mathbf{v}}(k,m)$ are defined in a similar way to $\widetilde{\mathbf{y}}(k,m)$, $X_{\mathrm{f}}(k,m) \triangleq \widetilde{\mathbf{h}}^H(k,m)\widetilde{\mathbf{x}}(k,m)$ is the filtered clean speech, and $V_{\mathrm{rn}}(k,m) \triangleq \widetilde{\mathbf{h}}^H(k,m)\widetilde{\mathbf{v}}(k,m)$ is the residual noise, which is uncorrelated with $X_{\mathrm{f}}(k,m)$.

The variance of $\widehat{X}_1(k,m)$ is

$$\begin{aligned}\phi_{\widehat{X}_1}(k,m) &\triangleq E\left[\left|\widehat{X}_1(k,m)\right|^2\right] \\ &= \widetilde{\mathbf{h}}^H(k,m)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(k,m)\widetilde{\mathbf{h}}(k,m) \\ &= \widetilde{\mathbf{h}}^H(k,m)\left[\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k,m) + \mathbf{\Phi}_{\widetilde{\mathbf{v}}}(k,m)\right]\widetilde{\mathbf{h}}(k,m) \\ &= \phi_{X_{\mathrm{f}}}(k,m) + \phi_{V_{\mathrm{rn}}}(k,m), \end{aligned} \quad (12)$$

where $\phi_{X_{\mathrm{f}}}(k,m) \triangleq E\left[|X_{\mathrm{f}}(k,m)|^2\right]$, $\phi_{V_{\mathrm{rn}}}(k,m) \triangleq E\left[|V_{\mathrm{rn}}(k,m)|^2\right]$, and $\mathbf{\Phi}_{\widetilde{\mathbf{a}}}(k,m)$ is the correlation matrix of $\widetilde{\mathbf{a}}(k,m)$ with $\widetilde{\mathbf{a}} \in \{\widetilde{\mathbf{y}}, \widetilde{\mathbf{x}}, \widetilde{\mathbf{v}}\}$ and $\widetilde{\mathbf{a}}(k,m) \triangleq [A_1(k,m) \ \cdots \ A_N(k,m) \ A_1^*(k,m) \ \cdots \ A_N^*(k,m)]^T$, which can be written into the following form:

$$\mathbf{\Phi}_{\widetilde{\mathbf{a}}}(k,m) = E\left[\widetilde{\mathbf{a}}(k,m)\widetilde{\mathbf{a}}^H(k,m)\right] \quad (13)$$

$$= \mathrm{diag}\left[\phi_{A_1}(k,m), \ldots, \phi_{A_N}(k,m), \phi_{A_1}(k,m), \ldots, \phi_{A_N}(k,m)\right] \times$$

$$\begin{bmatrix} 1 & \rho_{A_1A_2} & \cdots & \rho_{A_1A_N} & \gamma_{A_1A_1} & \gamma_{A_1A_2} & \cdots & \gamma_{A_1A_N} \\ \vdots & \cdots & \ddots & \vdots & \vdots & \cdots & \ddots & \vdots \\ \rho_{A_NA_1} & \rho_{A_NA_2} & \cdots & 1 & \gamma_{A_NA_1} & \gamma_{A_NA_2} & \cdots & \gamma_{A_NA_N} \\ \gamma_{A_1^*A_1^*} & \gamma_{A_1^*A_2^*} & \cdots & \gamma_{A_1^*A_N^*} & 1 & \rho_{A_1^*A_2^*} & \cdots & \rho_{A_1^*A_N^*} \\ \vdots & \cdots & \ddots & \vdots & \vdots & \cdots & \ddots & \vdots \\ \gamma_{A_N^*A_1^*} & \gamma_{A_N^*A_2^*} & \cdots & \gamma_{A_N^*A_N^*} & \rho_{A_N^*A_1^*} & \rho_{A_N^*A_2^*} & \cdots & 1 \end{bmatrix},$$

where

$$\rho_{A_iA_j} \triangleq \frac{E\left[A_i(k,m)A_j^*(k,m)\right]}{\phi_{A_i}(k,m)} \quad (14)$$

is the inter-channel correlation coefficient and

$$\gamma_{A_iA_j} \triangleq \frac{E[A_i(k,m)A_j(k,m)]}{\phi_{A_i}(k,m)} \quad (15)$$

is the pseudo inter-channel correlation coefficient. If $i = j$, we call $\gamma_{A_iA_i}$ the (second-order) circularity quotient [19], which satisfies

$$0 \le |\gamma_{A_iA_i}| \le 1. \quad (16)$$

The circularity coefficient measures the degree of noncircularity of $A_i(k,m)$. If $A_i(k,m)$ is circular then $|\gamma_{A_iA_i}| = 0$. A larger value of $|\gamma_{A_iA_i}|$ means that the signal $A_i(k,m)$ is more noncircular.

4. WL WIENER FILTER

In order to derive the WL Wiener filter, we first define the mean-squared-error (MSE) criterion.

The error between the desired signal, $X_1(k,m)$, and its estimate is defined as

$$\begin{aligned}\mathcal{E}(k,m) &\triangleq \widehat{X}_1(k,m) - X_1(k,m) \\ &= \widetilde{\mathbf{h}}^H(k,m)\widetilde{\mathbf{y}}(k,m) - X_1(k,m). \end{aligned} \quad (17)$$

[1]For the definition of circularity and how to measure it, see [19]–[21].

The subband MSE criterion is then given by

$$J\left[\widetilde{\mathbf{h}}(k,m)\right] \triangleq E\left[|\mathcal{E}(k,m)|^2\right] = E\left[\left|\widetilde{\mathbf{h}}^H(k,m)\widetilde{\mathbf{y}}(k,m) - X_1(k,m)\right|^2\right]. \quad (18)$$

Now taking the gradient of $J\left[\widetilde{\mathbf{h}}(k,m)\right]$ with respect to $\widetilde{\mathbf{h}}(k,m)$ and equating the result to zero, we obtain the optimal WL Wiener filter:

$$\widetilde{\mathbf{h}}_{\mathrm{W}}(k,m) = \mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,m)\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k,m)\mathbf{i}_1 = \left[\mathbf{I} - \mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,m)\mathbf{\Phi}_{\widetilde{\mathbf{v}}}(k,m)\right]\mathbf{i}_1, \quad (19)$$

where $\mathbf{i}_1$ is the first column of the identity matrix $\mathbf{I}$ of size $2N \times 2N$.

5. PERFORMANCE MEASURES

In this section, we give two performance metrics that will be applied to evaluate the performance of the WL Wiener filter, i.e., the signal-to-noise ratio (SNR) and the speech distortion index. The former measures the amount of noise reduction while the latter quantifies the amount of speech distortion.

With the signal model given in (2), we define the input SNR (i.e., the SNR at the first microphone pair) as

$$\mathrm{iSNR} \triangleq \frac{E\left[|x_1(t)|^2\right]}{E\left[|v_1(t)|^2\right]}. \quad (20)$$

After noise reduction, the enhanced signal consists of two components: the filtered desired signal, $x_{\mathrm{f}}(t)$, which is reconstructed from $X_{\mathrm{f}}(k,m)$ and the residual noise, $v_{\mathrm{rn}}(t)$, which is synthesized from $V_{\mathrm{rn}}(k,m)$. So, the output SNR can be defined as

$$\mathrm{oSNR} \triangleq \frac{E\left[|x_{\mathrm{f}}(t)|^2\right]}{E\left[|v_{\mathrm{rn}}(t)|^2\right]}. \quad (21)$$

A noise reduction filter is expected to improve the SNR; so the output SNR should be larger than the input SNR.

Since the noise reduction filter may distort the desired speech, we adopt the speech distortion index to quantify the distortion level of the desired signal [22], [23], i.e.,

$$\upsilon_{\mathrm{sd}} \triangleq \frac{E\left[|x_{\mathrm{f}}(t) - x_1(t)|^2\right]}{E\left[|x_1(t)|^2\right]}. \quad (22)$$

A larger value of υ_{sd} corresponds to a higher level of speech distortion.

Note that in our evaluation, we compute the SNRs and the speech distortion index using a long-time average, i.e., replacing the expectation operator with a time average. We should point out that the subband SNRs and the subband speech distortion index can also be defined to evaluate the WL Wiener filter; but we will not present them here due to space limitation.

6. EXPERIMENT

6.1. Experimental Setup

We conduct experiments in a reverberant room that is simulated using Lehmann's image-source method [24], [25]. The room is a rectangular one with size of 6700 mm long by 6100 mm wide by 2900 mm high. The floor layout of the room 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. An equispaced linear microphone array consisting of six omnidirectional microphones is employed in the measurement. The first microphone is at M1 (3237, 500) and the last one at M6 (3737, 500). A loudspeaker, simulating the source signal, is placed at one of the four positions from P1 (337, 1938) to P4 (3337, 1938). The four positions are uniformly spaced along the line connecting P1 and P4. The elevation of the microphone array is 1400, while it is 1600 for the source positions. Using Lehmann's model, we generate the impulse responses from the source positions to the microphones in a condition where the reverberation time T_{60} is 0.2 s.

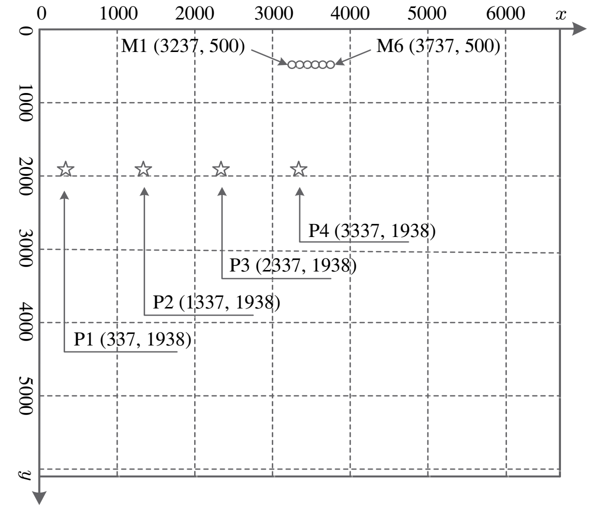

Fig. 1. Floor layout of the room setup (coordinate values measured in millimeters).

Table 1. Reconstruction error (i.e., the error between the original signal and the reconstructed one) with the overlap add technique (the overlap between two neighboring frames is 75%).

Window size (sample)	64	128	256	512
Error (dB)	−90.2	−87.4	−86.5	−86.1

6.2. Algorithm Implementation

The WL Wiener filter is implemented in the STFT domain. We first divide the noisy signals into short frames with an overlap of 75% between two consecutive frames. Then, each frame of the noisy signals is transformed into the frequency domain using the fast Fourier transform (FFT). Subsequently, a WL Wiener filter is designed and applied to the noisy STFT coefficients in each subband. Finally, the enhanced binaural signal is reconstructed in the time domain using the inverse FFT (IFFT) and overlap-add technique. Note that we deal with complex signals in both the time and frequency domains, the overlap add reconstruction and the aliasing problem due to circular convolution are slightly different from those in traditional noise reduction in the STFT domain. We will not discuss the aliasing problem in this paper due to space limitation. But just as in the traditional approach, we apply a real-valued Kaiser window in both the analysis and synthesis steps to reduce the aliasing effect. Table 1 presents the reconstruction error (i.e., the error between the original signal and the reconstructed one without applying any filter) for different frame lengths. One can see that the overlap add technique achieves a good reconstruction accuracy with the use of a Kaiser window.

To implement the WL Wiener filter in (19), we need to know the $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(k,m)$ and $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}(k,m)$ matrices. In our experiments, we compute $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(k,m)$ and $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}(k,m)$ using a short-time average on the P most recent frames of the corresponding signals. In other words, the noise

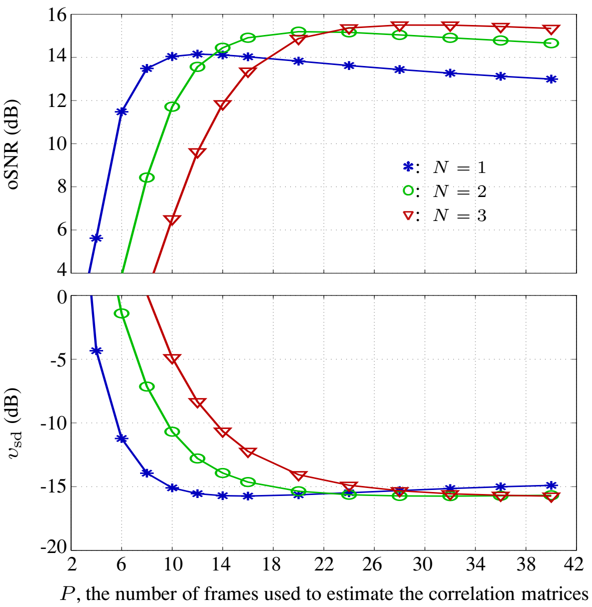

Fig. 2. Noise reduction performance of the WL Wiener filter as a function of P for three different values of N. The input SNR is 5 dB, the frame length is 128 (16 ms), and the overlap is 75%.

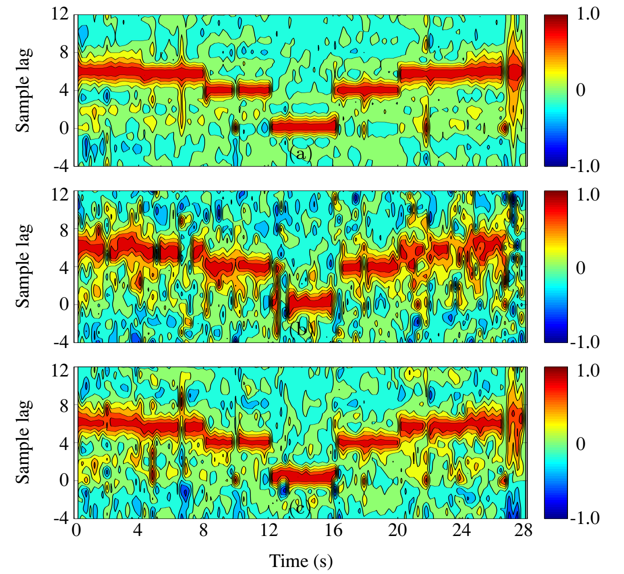

Fig. 3. The contours of the cross-correlation functions between the 1st and the 4th channels: (a) clean speech; (b) noisy speech (iSNR $=$ 5 dB); and (c) the enhanced speech by the WL Wiener filter with 6 microphones (i.e., $N = 3$).

estimation problem is put aside.

6.3. Experimental Result

In our experiments, the source is a speech signal recorded from a female speaker in a quiet office environment, which is sampled at 8 kHz. To simulate a moving speech source, we change its position every 4 seconds, first from P1 to P4, and then back. The microphone signals are generated by convolving the source signal with the corresponding impulse responses. White Gaussian noise is then added to control the input SNR.

We study the performance of the WL Wiener filter as a function of the number of pairs of microphones, i.e., $N = 1, 2, 3$. To make the results comparable, we use the 1st and 4th microphones in the array when $N = 1$, and the 1st, 2nd, 4th, and 5th microphones when $N = 2$. The input SNR is 5 dB, the frame length (same as the FFT size) is 128 (16 ms), and the number of frames that is used to compute the correlation matrices varies between 2 and 40, i.e., $P \in [2, 40]$. To evaluate the performance of the WL Wiener filter, we compute the output SNR, i.e., oSNR and the speech distortion index υ_{sd} using a long-time average. Figure 2 plots oSNR and υ_{sd} for different number of microphones as a function of P.

When the number of frames used to compute the signal correlation matrices is small (e.g., $P < 6$), the estimated correlation matrices are not accurate and the statistics estimation error affects the performance of the WL Wiener filter. As more frames are used, the correlation matrix estimates are more accurate. Therefore, one can see that the performance of the WL Wiener filter first increases with the value of P. However, when P is large (e.g., $P > 14$), if we continue to increase its value, the noise reduction performance does no longer improve, or sometimes degrades slightly. This is due to the fact that the estimated correlation matrices cannot follow the time-varying statistics of the speech signal if P is too large. One can see that the best performance of the WL Wiener filter for each studied value of N is achieved with a moderate value of P. Comparing the best performance of the WL Wiener filter for different values of N, one can see that the WL Wiener filter with more microphones (corresponding to a larger N) performs better. However, as the value of N increases, we need a larger P to achieve the optimal performance. This result can be explained as follows. With a larger value of N, the WL Wiener filter can exploit the spatial information from more microphones to produce better performance. However, with a larger N, the signal correlation matrices are larger in size and, therefore, require more data samples for accurate estimation.

To visualize the spatial information recovery with the WL Wiener filter, we compute the cross-correlation function between the enhanced binaural (the 1st and 4th microphones) signals and compare it to those of the clean and noisy speech. The cross-correlation functions are computed every 256 ms using a short-time average. The contours of the computed cross-correlation functions are plotted in Fig. 3, where the lag time corresponding to the maximal value of the cross-correlation function shows the position of the speech source. One can see that the noise has considerably changed the spatial information of the speech source. The WL Wiener filter does not only mitigate the noise effect, but also recovers the source spatial information.

7. CONCLUSIONS

In this paper, we studied the problem of binaural noise reduction with an array of microphones in the STFT domain. We first modeled the problem into a complex MISO system by merging the multichannel observation signals and the two (binaural) expected output channels into complex signals. The complex signals are then transformed into the frequency domain via the STFT. Under the WL framework, a WL Wiener filter based on the WL estimation theory and the MSE criterion is designed and applied to the STFT coefficients to achieve noise reduction. It is shown that this WL Wiener filter is able to exploit the noncircularity of the complex speech signals and the spatial information captured by the microphone array to achieve noise reduction and preserve the sound spatial information at the same time. The more the number of microphones, the better is the performance of binaural noise reduction.

8. REFERENCES

[1] J. E. Greenberg and P. M. Zurek, “Evaluation of an adaptive beamforming method for hearing aids,” *J. Acoust. Soc. Am.*, vol. 93, pp. 1662–1676, 1992.

[2] M. Kompis and N. Dillier, “Noise reduction for hearing aids: combining directional microphones with an adpative beamformer,” *J. Acoust. Soc. Am.*, vol. 96, pp. 1910–1913, 1994.

[3] 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.

[4] 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, pp. 543–551, Nov. 1997.

[5] J. V. Berghe and J. Wouters, “An adaptive noise canceller for hearing aids using two nearby microphones,” *J. Acoust. Soc. Am.*, vol. 103, pp. 3621–3626, 1998.

[6] 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.

[7] T. Lotter, B. Sauert, and P. Vary, “A stereo input-output superdrective beamformer for dual channel noise reduction,” in *Proc. EUROSPEECH*, 2005, pp. 2285–2288.

[8] K. U. Simmer, J. Bitzer, and C. Marro, “Post-filtering techniques,”in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds., chapter 3, pp. 39–60, Berlin, Germany:Springer, 2001.

[9] T. Lotter and P. Vary, “Dual-channel speech enhancement by superdirective beamforming,”*EURASIP J. Applied Signal Process.*, vol. 2006, pp. 175–175, Jan. 2006.

[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] 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.

[12] J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, “Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication,” *Speech Communication*, Vol. 53, No. 5, pp. 677–689, 2011.

[13] 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.

[14] J. Benesty and J. Chen, “A multichannel widely linear approach to binaural noise reduction using an array of microphones,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Mar. 2012, pp. 313–316.

[15] J. Chen and J. Benesty, “Binaural noise reduction based on widely linear filtering with multiple microphones,” in *Proc. InterNoise*, Aug. 2012.

[16] J. Chen and J. Benesty, “On the time-domain widely linear LCMV filter for noise reduction with a stereo system,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1343–1354, July 2013.

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

[18] B. Picinbono and P. Chevalier, “Widely linear estimation with complex data,” *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 2030–2033, Aug. 1995.

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

[20] 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.

[21] 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.

[22] 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, Chapter 2, pp. 9–41.

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

[24] E. Lehmann and A. Johansson, “Prediction of energy decay in room impulse responses simulated with an image-source model,” *J. Acoust. Soc. Am.*, vol. 124, pp. 269–277, 2008.

[25] Image-source method: Matlab code implementation. http://www.eric-lehmann.com/