Study of Widely Linear Multichannel Wiener Filter for Binaural Noise Reduction

Xin Leng*, Jingdong Chen*, Israel Cohen†, and Jacob Benesty‡
*Northwestern Polytechnical University, CIAIC and School of Marine Science and Technology, Xi'an, China.
†Technion-Israel Institute of Technology, Department of Electrical Engineering, Haifa, Israel.
‡University of Quebec, INRS-EMT, Montreal, Canada.

Abstract—**In this paper, we study the binaural noise-reduction problem using an array of microphones. The widely linear (WL) framework in the short-time-Fourier-transform (STFT) domain is adopted. In such a framework, the microphone array signals and binaural outputs are first merged into complex signals. These complex signals are subsequently transformed into the STFT domain. The WL estimation theory is then applied in STFT subbands with interband correlation to form the optimal WL Wiener filter, which exploits the noncircular properties of the input complex signals to achieve noise reduction and meanwhile to preserve the sound spatial realism. Finally, the time-domain binaural output is reconstructed from the output of the WL Wiener filter using the inverse STFT. The effectiveness of the developed STFT-domain WL Wiener filter for binaural noise reduction is justified using experiments.**

I. Introduction

Binaural noise reduction is an important problem in many applications e.g., hearing aids, virtual/augmented reality, 3D gaming, teleconferencing, etc. It has received tremendous research interest over the last few decades [1]–[10]. Unlike the widely studied subject of monaural noise reduction, which aims only at reducing noise, the objective of binaural noise reduction consists of two aspects: noise reduction (to improve either speech quality or intelligibility [11]) and preservation of sound spatial information. To achieve this objective, a binaural noise-reduction system generally takes multichannel (at least two) inputs from an array of microphones and produces two-channel outputs.

A straightforward way of achieving binaural noise reduction is through the use of some monaural noise reduction techniques to produce two outputs while some constraints between the two outputs are applied to preserve the so-called sound spatial cues [3]–[5]. But this method requires good estimation of the spatial cues and preservation process is in general not optimal. Recently, a widely linear (WL) filtering approach was developed to achieve binaural noise reduction using two microphones [6], [7]. It works in the complex domain by combining both the stereo input and expected binaural output signals into complex signals. Through this, the binaural noise reduction problem is transformed into one of single-channel noise reduction under the WL filtering framework. More recently, this principle was extended to the case of multiple microphones [8]. The WL filtering approach is proven to be effective for binaural noise reduction. However, the time-domain formulation and processing developed in [6]–[8] is in general computationally very expensive. To make the implementation more efficient, the time-domain framework was extended to the short-time-Fourier-transform (STFT) domain in [10], where coefficients from different STFT subbands are assumed to be uncorrelated.

This paper is also concerned with the binaural noise-reduction problem performed in the STFT domain. In contrast with the previous work reported in [10], the contribution of our paper lies in the following two aspects. First, we show that with the WL model in the STFT domain, there exists some relationship between certain subbands. Second, a WL Wiener filter is developed that takes into account the relationship between different subbands to achieve binaural noise reduction. We will show how to derive the optimal WL Wiener filter when interband relationship is taken into account. The performance of the developed STFT-domain WL Wiener filter is verified using experiments and comparison is made to show the advantage of the WL Wiener filter in this paper over its counterpart in [10].

II. Problem Formulation

The signal model adopted in this paper is same as the one used in [8]. Let us consider the scenario where a sound source radiates a signal of interest in a reverberant and noisy acoustic environment. We use a microphone array (with $2M$ sensors) to capture the signal. Then, the output of each microphone is written as

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

where $s(t)$ is the unknown sound source signal, $*$ denotes linear convolution, $g_{\mathrm{r},m}(t)$ denotes the room impulse response from $s(t)$ to the mth channel, and $x_{\mathrm{r},m}(t) = s(t) * g_{\mathrm{r},m}(t)$ and $v_{\mathrm{r},m}(t)$ are the convolved speech and additive noise, respectively, captured by the mth microphone. All the signals $x_{\mathrm{r},m}(t)$ and $v_{\mathrm{r},m}(t)$ are assumed to be real, broadband, and zero mean. Furthermore, it is assumed that the signals $x_{\mathrm{r},m}(t)$ are uncorrelated with $v_{\mathrm{r},m}(t)$. By definition, $x_{\mathrm{r},m}(t)$ are assumed to be coherent across the array, while $v_{\mathrm{r},m}(t)$ may be either partially coherent or incoherent across the array.

To achieve binaural noise reduction, we need to simultaneously recover the speech signals at two of the $2M$ microphones. Without loss of generality, we choose to recover $x_{\mathrm{r},1}(t)$ and $x_{\mathrm{r},M+1}(t)$. Following the principle in [6], [8], we choose to work in the complex domain by merging the real array outputs into complex signals so that the original problem is converted to one of multiple-input-single-output noise reduction. With the real signal model given in (1), the complex signals used in this paper are formed as

$$\begin{aligned} y_i(t) &\triangleq y_{\mathrm{r},i}(t) + jy_{\mathrm{r},M+i}(t) \\ &= s(t) * g_i(t) + v_i(t) \\ &= x_i(t) + v_i(t),\ i = 1, 2, \ldots, M, \end{aligned} \tag{2}$$

where $j = \sqrt{-1}$ denotes the imaginary unit, $g_i(t) = g_{\mathrm{r},i}(t) + jg_{\mathrm{r},M+i}(t)$ is the complex acoustic impulse response for the ith complex channel, $x_i(t) = x_{\mathrm{r},i}(t) + jx_{\mathrm{r},M+i}(t)$ is the complex clean signal, and $v_i(t) = v_{\mathrm{r},i}(t) + jv_{\mathrm{r},M+i}(t)$ is the complex additive noise. With the above complex signal model, the binaural noise-reduction problem can now be restated as: minimizing the effect of the noise term, $v_i(t)$, thereby recovering the complex signal $x_1(t)$, including the spatial information embedded in it.

As demonstrated in [6], [7], all the signals $y_i(t)$ are noncircular complex random variables (CRVs). So, the WL filtering theory needs to be used in order to recover $x_1(t)$ from the M complex noisy signals $y_i(t)$.

In the STFT domain, we can rewrite (2) as

$$Y_i(k,n) = X_i(k,n) + V_i(k,n), \tag{3}$$

where $Y_i(k,n)$, $X_i(k,n)$, and $V_i(k,n)$ are respectively the STFT coefficients of the complex signals $y_i(t)$, $x_i(t)$, and $v_i(t)$ at frequency-bin k (with $k = 0, 1, \ldots, K-1$ and K being the total frequency bins) and time-frame n. Putting $Y_i(k,n)$, $i = 1, 2, \cdots, M$, into a vector notation, we get

$$\mathbf{y}(k,n) = \mathbf{x}(k,n) + \mathbf{v}(k,n), \tag{4}$$

where $\mathbf{y}(k,n) \triangleq \left[Y_1(k,n)\ Y_2(k,n)\ \cdots\ Y_M(k,n)\right]^T$ with ^{T} standing for the transpose operator, and $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ are defined analogously to $\mathbf{y}(k,n)$.

III. Correlation between Different STFT Subbands

In monaural noise reduction in the STFT domain, coefficients from different STFT subbands are assumed uncorrelated either implicitly or explicitly and noise reduction at different bands are typically processed independently. This is generally true for real signals if the length of the fast Fourier transform (FFT) is sufficiently large. The same assumption was adopted in [10] for binaural noise reduction in the STFT domain with the WL framework. However, with the signal model given in (3), there exists certain relationship between the STFT coefficients at the k and $(K-k)$th subbands [12], [13]. As a matter of fact, it can be checked from (3) that

$$X_i^*(K-k,n) = \frac{G_i^*(K-k)}{G_i(k)} X_i(k,n), \tag{5}$$

where the superscript * stands for complex conjugation, and $G_i(k)$ is the STFT coefficient of $g_i(t)$. Therefore, both the coefficients from the k and $(K-k)$th subbands should be considered together in order to recover the clean speech at the kth subband. To explore this relationship, let us define the following signal vector:

$$\begin{aligned} \underline{\mathbf{y}}(k,n) &\triangleq \begin{bmatrix} \mathbf{y}(k,n) \\ \mathbf{y}^*(K-k,n) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{x}(k,n) \\ \mathbf{x}^*(K-k,n) \end{bmatrix} + \begin{bmatrix} \mathbf{v}(k,n) \\ \mathbf{v}^*(K-k,n) \end{bmatrix} \\ &= \underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n), \end{aligned} \tag{6}$$

where $\underline{\mathbf{x}}(k,n)$ and $\underline{\mathbf{v}}(k,n)$ are defined analogously to $\underline{\mathbf{y}}(k,n)$, respectively. It follows then that

$$\mathbf{x}(k,n) = \mathbf{d}(k)X_1(k,n), \tag{7}$$

$$\mathbf{x}^*(K-k,n) = \mathbf{d}'(K-k)X_1(k,n), \tag{8}$$

where

$$\mathbf{d}(k) = \left[1\ \frac{G_2(k)}{G_1(k)}\ \cdots\ \frac{G_M(k)}{G_1(k)}\right]^T, \tag{9}$$

$$\mathbf{d}'(K-k) = \left[\frac{G_1^*(K-k)}{G_1(k)}\,\frac{G_2^*(K-k)}{G_1(k)}\cdots\frac{G_M^*(K-k)}{G_1(k)}\right]^T. \tag{10}$$

Combining (6), (7), and (8), we obtain

$$\underline{\mathbf{y}}(k,n) = \underline{\mathbf{d}}(k)X_1(k,n) + \underline{\mathbf{v}}(k,n), \tag{11}$$

where

$$\underline{\mathbf{d}}(k) \triangleq \begin{bmatrix} \mathbf{d}(k) \\ \mathbf{d}'(K-k) \end{bmatrix}. \tag{12}$$

From the signal model (11), one can see that the binaural noise-reduction problem now is changed into one of estimating $X_1(k,n)$ from the complex signal vector $\underline{\mathbf{y}}(k,n)$.

IV. STFT-Domain Widely Linear Filtering for Binaural Noise Reduction

The estimation of $X_1(k,n)$ from the complex signal vector $\underline{\mathbf{y}}(k,n)$ can be accomplished using the WL estimation theory [14]–[16] as

$$\begin{aligned} \widehat{X}_1(k,n) &= \mathbf{h}^H(k,n)\underline{\mathbf{y}}(k,n) + \mathbf{h}'^H(k,n)\underline{\mathbf{y}}^*(k,n) \\ &= \widetilde{\mathbf{h}}^H(k,n)\widetilde{\mathbf{y}}(k,n) \\ &= \widetilde{\mathbf{h}}^H(k,n)\widetilde{\mathbf{x}}(k,n) + \widetilde{\mathbf{h}}^H(k,n)\widetilde{\mathbf{v}}(k,n), \end{aligned} \tag{13}$$

where the superscript ^{H} denotes the conjugate-transpose operator, $\mathbf{h}(k,n)$ and $\mathbf{h}'(k,n)$ are two complex finite-impulse-response (FIR) filters both of length $2M$,

$$\widetilde{\mathbf{h}}(k,n) \triangleq \begin{bmatrix} \mathbf{h}(k,n) \\ \mathbf{h}'(k,n) \end{bmatrix} \tag{14}$$

is a vector of length $4M$, named as the augmented WL filter,

$$\widetilde{\mathbf{y}}(k,n) \triangleq \begin{bmatrix} \underline{\mathbf{y}}(k,n) \\ \underline{\mathbf{y}}^*(k,n) \end{bmatrix} \tag{15}$$

is the augmented noisy signal vector, also with a length of $4M$, and $\widetilde{\mathbf{x}}(k,n)$ and $\widetilde{\mathbf{v}}(k,n)$ are defined analogously to $\widetilde{\mathbf{y}}(k,n)$.

If we set $\mathbf{h}'(k,n) = \mathbf{0}_{2M}$ (where $\mathbf{0}_{2M}$ is a $2M \times 1$ vector consisting of all zero elements) for any k and n, (13) degenerates to the classical linear filtering framework [17], [18]; however, this classical filtering process is not optimal for noncircular signals [14].

From (13), one can see that $\widehat{X}_1(k,n)$ depends on the signal vector $\widetilde{\mathbf{x}}(k,n)$; but the desired signal at frequency-bin k and time-frame n is $X_1(k,n)$ instead of the whole vector $\widetilde{\mathbf{x}}(k,n)$. To see how each element in $\widetilde{\mathbf{x}}(k,n)$ contributes to the estimate $\widehat{X}_1(k,n)$, let us first decompose $X_1^*(k,n)$ as

$$X_1^*(k,n) = \gamma_{X_1}^*(k,n)X_1(k,n) + X_1'(k,n), \tag{16}$$

where

$$X_1'(k,n) = X_1^*(k,n) - \gamma_{X_1}^*(k,n)X_1(k,n), \tag{17}$$

$$\gamma_{X_1}(k,n) = \frac{E[X_1^2(k,n)]}{E[|X_1(k,n)|^2]} \tag{18}$$

is the second-order circularity quotient [19] of $X_1(k,n)$, and $E[\cdot]$ denotes the mathematical expectation. If $\gamma_{X_1}(k,n) = 0$, $X_1(k,n)$ is second-order circular; otherwise, $X_1(k,n)$ is noncircular. The absolute value of $\gamma_{X_1}(k,n)$, which is between 0 and 1, quantifies the degree of noncircularity of $X_1(k,n)$; a larger value of $|\gamma_{X_1}(k,n)|$ indicates that $X_1(k,n)$ more noncircular. From (16), it can checked that

$$E[X_1'(k,n)X_1^*(k,n)] = 0. \tag{19}$$

Using (16), we can write $\widetilde{\mathbf{x}}(k,n)$ as

$$\begin{aligned}\widetilde{\mathbf{x}}(k,n) &= \mathbf{d}_{X_1}(k,n)X_1(k,n) + \widetilde{\mathbf{x}}'(k,n)\\ &= \mathbf{x}_{\rm d}(k,n) + \widetilde{\mathbf{x}}'(k,n),\end{aligned} \tag{20}$$

where

$$\begin{aligned}\mathbf{d}_{X_1}(k,n) &\triangleq \begin{bmatrix} \underline{\mathbf{d}}(k) \\ \gamma_{X_1}^*(k,n)\underline{\mathbf{d}}^*(k)\end{bmatrix}\\ &= \frac{E[\widetilde{\mathbf{x}}(k,n)X_1^*(k,n)]}{E[|X_1(k,n)|^2]},\end{aligned} \tag{21}$$

$$\mathbf{x}_{\rm d}(k,n) \triangleq \mathbf{d}_{X_1}(k,n)X_1(k,n), \tag{22}$$

$$\widetilde{\mathbf{x}}'(k,n) \triangleq \begin{bmatrix} \mathbf{0}_{2M} \\ X_1'(k,n)\underline{\mathbf{d}}^*(k)\end{bmatrix}. \tag{23}$$

Now, substituting (20) into (13), we get

$$\begin{aligned}\widehat{X}_1(k,n) &= \widetilde{\mathbf{h}}^H(k,n)[\mathbf{x}_{\rm d}(k,n) + \widetilde{\mathbf{x}}'(k,n) + \widetilde{\mathbf{v}}(k,n)]\\ &= X_{\rm fd}(k,n) + X_{\rm ri}(k,n) + V_{\rm rn}(k,n),\end{aligned} \tag{24}$$

where $X_{\rm fd}(k,n) \triangleq X_1(k,n)\widetilde{\mathbf{h}}^H(k,n)\mathbf{d}_{X_1}(k,n)$ is the filtered desired signal, $X_{\rm ri}(k,n) \triangleq \widetilde{\mathbf{h}}^H(k,n)\widetilde{\mathbf{x}}'(k,n)$ is called the residual interference, and $V_{\rm rn}(k,n) \triangleq \widetilde{\mathbf{h}}^H(k,n)\widetilde{\mathbf{v}}(k,n)$ is called the residual noise.

One can verify that the two vectors $\widetilde{\mathbf{y}}(k,n)$ and $\widetilde{\mathbf{y}}(K-k,n)$ satisfy the following relation:

$$\widetilde{\mathbf{y}}(K-k,n) = \mathbf{P}\,\widetilde{\mathbf{y}}(k,n), \tag{25}$$

where

$$\mathbf{P} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_M \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_M & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_M & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_M & \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix} \tag{26}$$

is the anti-diagonal matrix which has the properties of $\mathbf{P}^T = \mathbf{P}$ and $\mathbf{P}^2 = \mathbf{I}_{4M}$, $\mathbf{I}_M$ denotes the identity matrix of size $M \times M$. Therefore, $\widetilde{\mathbf{y}}(K-k,n)$ is simply a permutation of $\widetilde{\mathbf{y}}(k,n)$. It follows then that

$$\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(K-k,n) = \mathbf{P}\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(k,n)\mathbf{P}, \tag{27}$$

where $\mathbf{\Phi}_{\widetilde{\mathbf{y}}}(k,n) \triangleq E[\widetilde{\mathbf{y}}(k,n)\widetilde{\mathbf{y}}^H(k,n)]$ is the covariance matrix of the noisy signal vector. The above relationship can be used to reduce the complexity of the WL noise reduction filter, which will become clear in the next section.

V. Widely Linear Wiener Filter

Before deriving the optimal STFT-domain WL Wiener filter, let us first define the subband mean-square error (MSE) between the estimated and clean signals at the frequency-bin k and time-frame n:

$$\begin{aligned}J(k,n) &\triangleq E\big[|\widehat{X}_1(k,n) - X_1(k,n)|^2\big]\\ &= E\big[|\widetilde{\mathbf{h}}^H(k,n)\widetilde{\mathbf{y}}(k,n) - X_1(k,n)|^2\big].\end{aligned} \tag{28}$$

The WL Wiener filter is derived by taking the gradient of the subband MSE, $J(k,n)$, with respect to $\widetilde{\mathbf{h}}^H(k,n)$ and forcing the result equal to zero. The solution is

$$\begin{aligned}\widetilde{\mathbf{h}}_{\rm W}(k,n) &= \mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,n)\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k,n)\mathbf{i}_{4M,1}\\ &= \big[\mathbf{I}_{4M} - \mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,n)\mathbf{\Phi}_{\widetilde{\mathbf{v}}}(k,n)\big]\mathbf{i}_{4M,1},\end{aligned} \tag{29}$$

where $\mathbf{i}_{4M,1}$ is the first column of $\mathbf{I}_{4M}$, and $\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k,n) \triangleq E[\widetilde{\mathbf{x}}(k,n)\widetilde{\mathbf{x}}^H(k,n)]$ and $\mathbf{\Phi}_{\widetilde{\mathbf{v}}}(k,n) \triangleq E[\widetilde{\mathbf{v}}(k,n)\widetilde{\mathbf{v}}^H(k,n)]$ are the covariance matrices of $\widetilde{\mathbf{x}}(k,n)$ and $\widetilde{\mathbf{v}}(k,n)$, respectively.

According to (21), we have

$$\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k,n)\mathbf{i}_{4M,1} = \phi_{X_1}(k,n)\mathbf{d}_{X_1}(k,n), \tag{30}$$

where $\phi_{X_1}(k,n) = E[|X_1(k,n)|^2]$ is the variance of $X_1(k,n)$. So, we can also write the WL Wiener filter as

$$\begin{aligned}\widetilde{\mathbf{h}}_{\rm W}(k,n) &= \phi_{X_1}(k,n)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,n)\mathbf{d}_{X_1}(k,n)\\ &= \mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,n)\big[\phi_{Y_1}(k,n)\mathbf{d}_{Y_1}(k,n)\\ &\quad - \phi_{V_1}(k,n)\mathbf{d}_{V_1}(k,n)\big],\end{aligned} \tag{31}$$

where $\phi_{Y_1}(k,n)$ and $\phi_{V_1}(k,n)$ are, respectively, the variances of $Y_1(k,n)$ and $V_1(k,n)$, and $\mathbf{d}_{Y_1}(k,n)$ and $\mathbf{d}_{V_1}(k,n)$ are defined analogously to $\mathbf{d}_{X_1}(k,n)$ in (21).

With the derived WL Wiener filter, the resulting signal estimate at (k,n) is

$$\begin{aligned}\widehat{X}_1(k,n) &= \widetilde{\mathbf{h}}_{\rm W}^H(k,n)\widetilde{\mathbf{y}}(k,n)\\ &= \mathbf{i}_{4M,1}^T\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k,n)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,n)\widetilde{\mathbf{y}}(k,n).\end{aligned} \tag{32}$$

Now using the relationship in (27), the estimate of $X_1(K-k,n)$ can be obtained as

$$\widehat{X}_1(K-k,n) = \mathbf{i}_{4M,3M+1}^T\mathbf{\Phi}_{\widetilde{\mathbf{x}}}(k,n)\mathbf{\Phi}_{\widetilde{\mathbf{y}}}^{-1}(k,n)\widetilde{\mathbf{y}}(k,n), \tag{33}$$

where $\mathbf{i}_{4M,3M+1}$ is the $(3M+1)$th column of $\mathbf{I}_{4M}$. Inspecting (32) and (33), one can see that we only need to estimate the WL Wiener filter for half of the total STFT subbands, which is similar to the case of monaural noise reduction with real input signals.

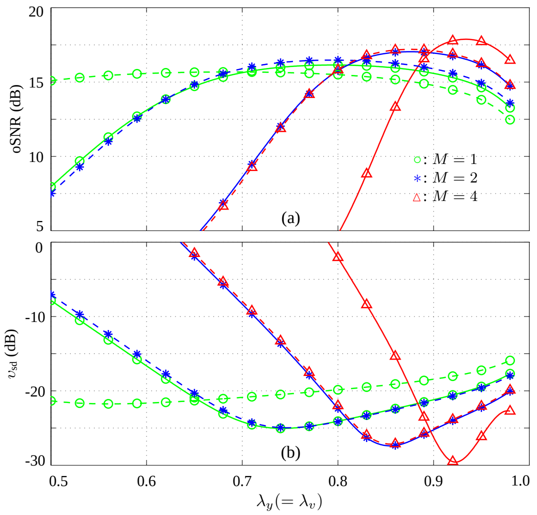

Fig. 1. Performance of the developed WL Wiener filter (solid) and the filter in [10] (dashed) as a function of the forgetting factor in white Gaussian noise: (a) the fullband output SNR and (b) the fullband speech distortion index.

VI. Experiments

Now, we briefly evaluate the performance of the developed STFT-domain WL Wiener filer using experiments. For comparison, the filter developed in [10] is also evaluated. The experiments are configured using the room impulse responses measured at Bell Labs Varechoic Chamber [20], [21]. We consider a moderate reverberation condition with the reverberation time T_{60} of approximately $0.24\,$s. An equispaced linear microphone array with 8 omnidirectional microphones is configured: the first sensor is located at the position (3.037, 0.500, 1.400) (in meters) and the last sensor is place at (3.737, 0.500, 1.400), the spacing is $0.1\,$m. To simulate a moving source, we play back some speech signals from the TIMIT database [22] and change the position of the source every 4 seconds among positions (1.337:1.000:4.337, 1.938, 1.600) (forth and back). The microphone signals are generated by convolving the source signal with the corresponding impulse responses and white Gaussian noise is then added to the convolution results to control the input signal-to-noise ratio (SNR) to be $5\,$dB. All the signals are resampled from the original sampling rate to $8\,$kHz. Note that in this paper we put aside the influence of noise estimation on performance and compute the covariance matrices directly from the noisy and noise signals using a recursive method with the two forgetting factors $\lambda_y = \lambda_v$ [23].

Both the fullband output SNR and speech distortion index [6] of the developed WL Wiener filter and the filter in [10] are plotted in Fig. 1. We observe that both filters are able to improve the output SNR considerably, but with some distortion being added into the speech. Comparatively, the WL Wiener filter developed in this paper can yield better performance, i.e., higher output SNR and smaller value of the speech distortion index when the forgetting factors are properly chosen. It is interesting to notice that the developed WL filter requires only half the number of microphones for obtaining a similar performance achieved with the method in [10].

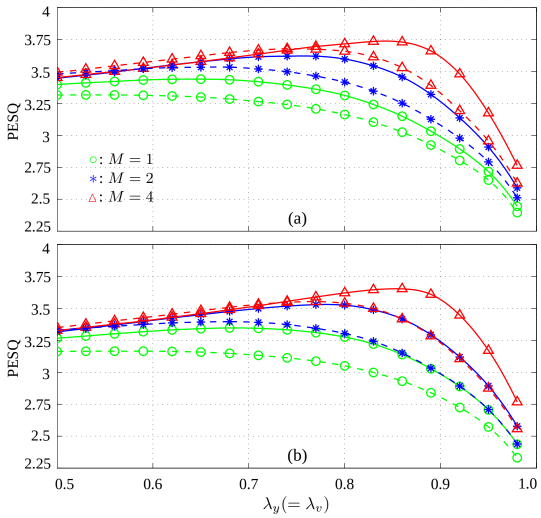

Fig. 2. The PESQ score of the developed WL Wiener filter (solid) and the filter in [10] (dashed) as a function of the forgetting factor in white Gaussian noise. (a) The PESQ score for the left-channel enhanced speech. (b) The PESQ score for the right-channel enhanced speech.

Fig. 2 plots the perceptual evaluation of speech quality (PESQ) [24] scores of both the developed WL Wiener filter and the filter in [10] as a function of the forgetting factor, λ_y. Since the PESQ standard does not support complex signals, we take the left- and right-channel outputs from the enhanced complex speech signals and compute the PESQ scores separately. It can be observed from Fig. 2 that the PESQ score first increases with λ_y and then decreases. Comparatively, the WL Wiener filter developed in this work achieves a higher PESQ score than the method in [10]. Based on the results in Fig. 2, Table I gives the difference between the maximum PESQ scores that are achieved with the two WL Wiener filters with properly chosen forgetting factors.

TABLE I
Difference between the maximum PESQ scores of the developed WL Wiener filter and the filter in [10].

M	1	2	4
Left	0.12	0.09	0.06
Right	0.18	0.13	0.10

To visualize the preservation of the sound spatial information, we computed the cross-correlation function (CCF) between the signals at the two output channels (estimating the signal of interest from the first and 5th microphones) every $128\,$ms. The CCFs are computed using a short-time average method as in [6]. The contours of the time-varying CCFs of the clean, noisy, and two enhanced signals are plotted in Fig. 3, where 8 microphones are used, i.e., $M = 4$, and value of the forgetting factors for the method in [10] is 0.89 and that of the developed Wiener filter is 0.92 (the value is chosen according to the maximum output SNR that the respective filter can achieve as in the previous simulation). In Fig. 3, the maximal value of the CCF at each time can be seen as the current position of the moving speech source. At the presence of noise, one can note that the sound spatial effect has been dramatically modified. From the third and bottom traces in Fig. 3, One can

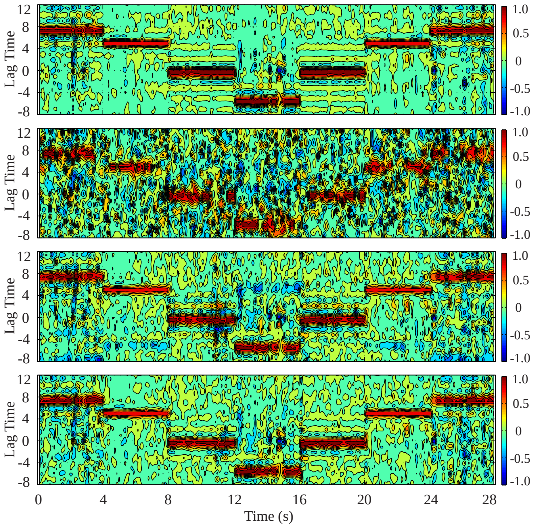

Fig. 3. Computed CCF Contours for the clean speech (top trace), noisy speech (second trace), the enhanced speech by the method [10] (third trace), and the enhanced speech using (29) (bottom trace).

see that both the method in [10] and the developed WL Wiener filter recover the sound spatial information very well.

To quantitatively compare the performance of the developed WL Wiener filter and the filter in [10] in terms of noise reduction and spatial information preservation, we compute both the output SNR and the Euclidean distance between the clean speech CCF (the CCF between the clean speech at the first microphone and that at the 5th microphone) and that of the enhanced signals. With our experimental setup, the output SNR of the WL Wiener filter developed in this paper is 17.77 dB while that of the method in [10] is 17.16 dB. The distance between the clean CCF and the CCF of the enhanced signals by the WL Wiener filter developed in this paper is 9.13 while that by the method in [10] is 11.88. These results clearly indicate that the developed STFT-domain WL Wiener filter outperforms the method in [10].

VII. Conclusion

In this paper, we investigated the binaural noise-reduction problem based on the use of microphone arrays. We adopted the WL filtering framework in which both the multiple inputs and binaural outputs were merged into complex signals, which were subsequently transformed into the STFT domain to achieve binaural noise reduction. The noncircularity property of the complex signals and the interband relationship were subsequently exploited and a WL multichannel Wiener filter was developed. Experiments showed that this WL Wiener filter did not only enhance the noisy speech dramatically, but also recovered the spatial information of the clean speech source. In comparison with a method developed recently, the WL Wiener filter derived in this work yielded higher output SNR, larger PESQ score, smaller value of the speech distortion index, and better preservation of the source spatial information. It was observed that the developed WL Wiener filter only requires half of the number of microphones for obtaining a similar performance of a recently developed method.

References

[1] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[2] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Feb. 2015.

[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, no. 6, pp. 529–542, Nov. 1997.

[4] S. Doclo, R. Dong, T. J. Klasen, J. Wouters, S. Haykin, and M. Moonen, "Extension of the multi-channel Wiener filter with ITD cues for noise reduction in binaural hearing aids," in *Proc. IEEE WASPAA*, 2005, pp. 70–73.

[5] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication," *Speech Commun.*, vol. 53, no. 5, pp. 677–689, 2011.

[6] 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, no. 8, pp. 2260–2272, Nov. 2011.

[7] 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, no. 7, pp. 1343–1354, Jul. 2013.

[8] J. Benesty and J. Chen, "A multichannel widely linear approach to binaural noise reduction using an array of microphones," in *Proc. IEEE ICASSP*, 2012, pp. 313–316.

[9] J. Szurley, A. Bertrand, and M. Moonen, "On the use of time-domain widely linear filtering for binaural speech enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 649–652, Jul. 2013.

[10] L. Zhao, J. Chen, and J. Benesty, "A multichannel widely linear Wiener filter for binaural noise reduction in the short-time-Fourier-transform domain," in *Proc. IEEE ChinaSIP*, 2014, pp. 227–231.

[11] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[12] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.

[13] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

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

[15] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5101–5125, Nov. 2011.

[16] P. Chevalier, J.-P. Delmas, and A. Oukaci, "Properties, performance and practical interest of the widely linear MMSE beamformer for nonrectilinear signals," *Signal Process.*, vol. 97, pp. 269–281, Apr. 2014.

[17] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Berlin, Germany: Springer-Verlag, 2009, vol. 2.

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

[19] J. Benesty, J. Chen, and Y. Huang, "A widely linear distortionless filter for single-channel noise reduction," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 469–472, May 2010.

[20] W. C. Ward, G. Elko, R. Kubli, and W. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallace Clement Sabine Centenn. Symp.*, 1994.

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

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT acoustic-phonetic continuous speech corpus*. Linguistic Data Consort., 1993.

[23] J. Chen, J. Benesty, and Y. Huang, "Study of the noise-reduction problem in the Karhunen–Loève expansion domain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 787–802, May 2009.

[24] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*. ITU-T Rec. P. 862, 2001.