Effective Binaural Multi-Channel Processing Algorithm for Improved Environmental Presence

Pei Chee Yong, *Student Member, IEEE*, Sven Nordholm, *Senior Member, IEEE*, and Hai Huyen Dam

***Abstract*—Binaural noise-reduction algorithms based on multi-channel Wiener filter (MWF) are promising techniques to be used in binaural assistive listening devices. The real-time implementation of the existing binaural MWF methods, however, involves challenges to increase the amount of noise reduction without imposing speech distortion, and at the same time preserving the binaural cues of both speech and noise components. Although significant efforts have been made in the literature, most developed methods so far have focused only on either the former or latter problem. This paper proposes an alternative binaural MWF algorithm that incorporates the non-stationarity of the signal components into the framework. The main objective is to design an algorithm that would be able to select the sources that are present in the environment. To achieve this, a modified speech presence probability (SPP) and a single-channel speech enhancement algorithm are utilized in the formulation. The resulting optimal filter also avoids the poor estimation of the second-order clean speech statistics, which is normally done by simple subtraction. Theoretical analysis and performance evaluation using realistic recorded data shows the advantage of the proposed method over the reference MWF solution in terms of the binaural cues preservation, as well as the noise reduction and speech distortion.**

***Index Terms*—Binaural cues, modified sigmoid function, multi-channel wiener filter, single-channel noise reduction, speech enhancement.**

I. Introduction

In an adverse speech communication environment where the characteristics of target speech are distorted by high background noise, speech enhancement systems are required for improving speech quality and intelligibility. Several binaural techniques have been proposed in recent years for future hearing aids, where the full-duplex exchange of microphone signals over wireless link between the two devices become feasible. The aim of binaural noise reduction techniques is to improve the signal-to-noise ratio (SNR) of the signal, while simultaneously preserving the binaural cues of both target speech and residual noise. In terms of the application of hearing protectors where microphones are integrated into the hearing protector adjacent to each ear, the binaural processing algorithms can be readily applied as the microphones can be connected by cables [1].

The binaural noise reduction techniques in literature can be divided into two classes. In the first class, identical real-valued spectral gains are applied to one microphone signal on the left device and one microphone signal on the right device [2]–[6], so that the binaural cues of both speech and noise components are preserved. Although the outputs of a beamformer can be utilized to derive the spectral gain function, these techniques can be viewed as single-channel spectral weighting approaches, which are similar to single-channel speech enhancement techniques. The problem arises as single-channel noise reduction usually introduces speech distortion and noise artefacts known as the musical noise, leading to limited or no speech intelligibility improvements. Another drawback of these techniques is that the interfering sources located in the back direction cannot be suppressed due to the forward-backward ambiguity.

The second class of binaural techniques combines all microphone signals from both ears to perform a true array processing. Some techniques first construct a monaural output and then apply a postprocessing stage to reconstruct the binaural signals with correct binaural cues [7]. Other techniques apply fixed or adaptive beamformers which produce a binaural output, whereby the beamformers are designed or constrained so that the binaural cues are also preserved [7]–[10]. In [8], [9] (adaptive) beamforming is only applied in the higher frequencies, while for the lower frequencies the (low-pass filtered) original microphone signal with correct binaural cues is used. This approach is not suitable in practice, especially for industrial noise that is dominated by its low frequency components. Another method that utilizes coherent processing of all microphone elements is based on the multi-channel Wiener filter (MWF) technique, referred to as speech distortion weighted MWF (SDWMWF) [11], [12]. The drawback of SDWMWF is that it can only preserve binaural cues for speech but not for noise. It changes the noise binaural cues to be the same as those of the speech component. Although the SDWMWF cost function has been extended such that the binaural cues of both the target speech and the residual noise can be preserved, there is a trade-off between binaural cue preservation and noise reduction performance [13]. Another extension for SDWMWF is by employing a constraint on the preservation of the binaural cues of the noise component [14]. This however leads to a trade-off between binaural cues of the speech and noise components.

The focus of this paper is the development of binaural noise reduction algorithms for speech enhancement in assistive listening devices to protect the hearing of those working in loud

Manuscript received February 09, 2014; revised June 27, 2014; accepted September 13, 2014. Date of publication September 22, 2014; date of current version October 01, 2014. This work was supported in part by Sensear Pty Ltd and Linkage Grant LP100100433 by Australian Research Council (ARC). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Thomas Fang Zheng.

P. C. Yong and S. Nordholm are with the Department of Electrical and Computer Engineering, Curtin University, Bentley, WA 6102, Australia (e-mail: peichee.yong@postgrad.curtin.edu.au; s.nordholm@curtin.edu.au).

H. H. Dam is with the Department of Mathematics and Statistics, Curtin University, Bentley, WA 6102, Australia (e-mail: h.dam@curtin.edu.au).

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

Digital Object Identifier 10.1109/TASLP.2014.2359626

2329-9290 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

or noisy conditions. It is therefore crucial for the application of such assistive listening devices to be able to deal with industrial noise coming from different machinery or tools, with sound pressure levels often exceeding 80 dBA, and at the same time maintaining the spatial cues from surroundings. The main consideration in the design of the binaural multi-channel speech processing algorithm is to be able to adapt to changing environments, especially from a very noisy condition to a less noisy one, and vice versa. As such, speech enhancement approaches such as computational auditory scene analysis [15], which require pre-training before use, are not considered especially when we deal with online and adaptive block-wise speech processing.

In this paper, a modified MWF framework which incorporates and integrates non-stationarity assumptions about both speech and noise, is proposed for speech enhancement for assistive listening devices. Due to the fact that speech signals are known to be highly non-stationary, and may not even be present in all time-frames and frequencies during speech segments, a modified conditional speech presence probability (SPP) has been utilized to formulate the proposed method. This results in an optimal filter that prevents the subtraction of noise-only correlation matrix from the speech-plus-noise correlation matrix, and avoids the poor estimate of speech correlation matrix [16]. In addition, it also utilizes techniques from single-channel speech enhancement to adaptively select the sources depending on the conditional SPP from the environment. As such, the proposed method points a beam in the direction of the speech source when speech is active, and points the beam to the noise source when only noise is present. In this case, it is capable of obtaining an improved preservation of binaural cues of both speech sources and noise as well as providing higher noise suppression compared to conventional formulation. This means that the filter is not an actual Wiener filter but is inspired from the Wiener filter. Note that this paper is a modification of [16] in a binaural configuration manner, which emphasizes the ability to maintain spatial cues of both speech and noise by utilizing a modified conditional SPP algorithm from [17] and the single-channel techniques from [18].

The paper is organized as follows. In the second section, the proposed framework is developed, which includes the formulation of an alternative binaural MWF algorithm, the computation of the single-channel speech enhancement algorithm for the reference signals, and the involvement of SPP in different parts of the framework. Section III theoretically compares the proposed method with the conventional SDWMWF. Section IV demonstrates the performance measures used in this work. Section V presents the results and Section VI concludes the paper.

II. A Modified Binaural MWF

A. Configuration and Notation

Consider a configuration of L microphones with the l-th microphone signal on the left and the right sides of the hearing protection device defined in short-time Fourier transform (STFT) domain as

$$\begin{aligned} Y_{\text{left},l}(k,m) &= X_{\text{left},l}(k,m) + V_{\text{left},l}(k,m) \\ Y_{\text{right},l}(k,m) &= X_{\text{right},l}(k,m) + V_{\text{right},l}(k,m) \\ l &= 1,\ldots,L \end{aligned} \tag{1}$$

where $X_{\text{left},l}(k,m)$ and $X_{\text{right},l}(k,m)$ represent the speech components in the microphone signal, while $V_{\text{left},l}(k,m)$ and $V_{\text{right},l}(k,m)$ represent the noise components. Here, k and m denote the frequency bin index and time-frame index, respectively.

The L-dimensional stacked microphone signal vectors $\mathbf{y}_{\text{left}}(k,m)$ and $\mathbf{y}_{\text{right}}(k,m)$, and the $2L$-dimensional signal vector $\mathbf{y}(k,m)$ are given as

$$\mathbf{y}(k,m) = \begin{bmatrix} \mathbf{y}_{\text{left}}(k,m) \\ \mathbf{y}_{\text{right}}(k,m) \end{bmatrix} \tag{2}$$

with

$$\begin{aligned} &\mathbf{y}_{\text{left}}(k,m) \\ &= [Y_{\text{left},1}(k,m)\ Y_{\text{left},2}(k,m)\ \ldots\ Y_{\text{left},L}(k,m)]^T, \\ &\mathbf{y}_{\text{right}}(k,m) \\ &= [Y_{\text{right},1}(k,m)\ Y_{\text{right},2}(k,m)\ \ldots\ Y_{\text{right},L}(k,m)]^T \end{aligned} \tag{3}$$

where T denotes the transpose operator. The correlation matrix of speech plus noise $\mathbf{R}_y(k,m)$, the clean speech correlation matrix $\mathbf{R}_x(k,m)$, and the noise correlation matrix $\mathbf{R}_v(k,m)$ are defined as

$$\begin{aligned} \mathbf{R}_y(k,m) &= E\{\mathbf{y}(k,m)\mathbf{y}^H(k,m)\}, \\ \mathbf{R}_x(k,m) &= E\{\mathbf{x}(k,m)\mathbf{x}^H(k,m)\}, \\ \mathbf{R}_v(k,m) &= E\{\mathbf{v}(k,m)\mathbf{v}^H(k,m)\}, \end{aligned} \tag{4}$$

where H denotes the conjugate transpose operator. The $2L$-dimensional signal vectors $\mathbf{x}(k,m)$ and $\mathbf{v}(k,m)$ are defined similarly as $\mathbf{y}(k,m)$. The speech and the noise components are assumed uncorrelated, such that

$$\mathbf{R}_y(k,m) = \mathbf{R}_x(k,m) + \mathbf{R}_v(k,m). \tag{5}$$

For speech enhancement algorithms, the r_{left}-th(k,m) signal of the left device and the r_{right}-th(k,m) signal of the right device will be used as the so-called reference signals. Typically, the front microphones are used as reference microphones. For conciseness, the reference microphone signals are defined as

$$\begin{aligned} Y_{\text{left}}(k,m) &= \mathbf{e}_{\text{left}}^H \mathbf{y}(k,m) = X_{\text{left}}(k,m) + V_{\text{left}}(k,m) \\ Y_{\text{right}}(k,m) &= \mathbf{e}_{\text{right}}^H \mathbf{y}(k,m) = X_{\text{right}}(k,m) + V_{\text{right}}(k,m) \end{aligned} \tag{6}$$

where $X_{\text{left}}(k,m)$, $X_{\text{right}}(k,m)$, $V_{\text{left}}(k,m)$, and $V_{\text{right}}(k,m)$ are speech and noise components at the respective reference microphones. Here, $\mathbf{e}_{\text{left}}$ and $\mathbf{e}_{\text{right}}$ are $2L$-dimensional vectors with only one element equal to 1 and the other elements equal to 0, i.e., $\mathbf{e}_{\text{left}}(r_{\text{left}}) = 1$ and $\mathbf{e}_{\text{right}}(L + r_{\text{right}}) = 1$.

The output signals are obtained by filtering and summing all microphone signals

$$\begin{aligned} Z_{\text{left}}(k,m) &= \mathbf{w}_{\text{left}}^H(k,m)\mathbf{y}(k,m) \\ Z_{\text{right}}(k,m) &= \mathbf{w}_{\text{right}}^H(k,m)\mathbf{y}(k,m) \end{aligned} \tag{7}$$

where $\mathbf{w}_{\text{left}}(k,m)$ and $\mathbf{w}_{\text{right}}(k,m)$ are $2L$-dimensional complex weight vectors. The output signals can be written as

$$\begin{aligned} Z_{\text{left}}(k,m) &= \mathbf{w}_{\text{left}}^H(k,m)\mathbf{x}(k,m) + \mathbf{w}_{\text{left}}^H(k,m)\mathbf{v}(k,m) \\ &= Z_{x,\text{left}}(k,m) + Z_{v,\text{left}}(k,m) \end{aligned}$$

$$Z_{\text{right}}(k,m) = \mathbf{w}_{\text{right}}^H(k,m)\mathbf{x}(k,m) + \mathbf{w}_{\text{right}}^H(k,m)\mathbf{v}(k,m) = Z_{x,\text{right}}(k,m) + Z_{v,\text{right}}(k,m) \quad (8)$$

where $Z_{x,\text{left}}(k,m)$, $Z_{x,right}(k,m)$ represent the speech component and $Z_{v,\text{left}}(k,m)$, $Z_{v,\text{right}}(k,m)$ represent the noise component of the output signals at respective side.

B. Formulation of Binaural MWF Incorporating SPP

The MWF method has been widely used for binaural speech enhancement given that it produces a minimum mean square error (MMSE) estimate of the speech component in the reference microphone at respective sides, simultaneously reducing noise and limiting speech distortion [11], [14]. The mean square error (MSE) cost function for the filters $\mathbf{w}_{\text{left}}(k,m)$ and $\mathbf{w}_{\text{right}}(k,m)$ is equal to

$$\begin{aligned}\mathcal{J}_{\text{MWF}}&\left(\mathbf{w}(k,m)\right)\\ &= E\left\{\left\|\begin{bmatrix} X_{\text{left}}(k,m) - Z_{\text{left}}(k,m) \\ X_{\text{right}}(k,m) - Z_{\text{right}}(k,m)\end{bmatrix}\right\|^2\right\}\\ &= E\left\{\left\|\begin{bmatrix} X_{\text{left}}(k,m) - \mathbf{w}_{\text{left}}^H(k,m)\mathbf{y}(k,m) \\ X_{\text{right}}(k,m) - \mathbf{w}_{\text{right}}^H(k,m)\mathbf{y}(k,m)\end{bmatrix}\right\|^2\right\}. \quad (9)\end{aligned}$$

The drawback is that some residual noise will still remain in the output signals at both sides, Z_{left} and Z_{right}. In this paper, a bi-criteria optimization problem for binaural MWF is proposed, which consists of a criterion to minimize the error in Eq. (9) and another criterion to minimize the noise power. Furthermore, in order to cope with the fast dynamic changes of non-stationary speech and noise in real environment, a SPP algorithm is fully utilized in the proposed formulation of the binaural speech enhancement method. Let the two-state model for speech events at every frequency bin and frame be defined in this context as

$$\begin{aligned}\mathcal{H}_0(k,m) &: Y_l(k,m) = V_l(k,m)\\ \mathcal{H}_1(k,m) &: Y_l(k,m) = X_l(k,m) + V_l(k,m)\end{aligned} \quad (10)$$

where $\mathcal{H}_0(k,m)$ and $\mathcal{H}_1(k,m)$ denote speech absence and speech presence in the k-th frequency bin of the m-th frame, respectively. Such model can be incorporated directly into the design criterion of the MWF, leading to a weighted average formulation where the first term is weighted by the probability that speech is present, while the second term is weighted by the probability that speech is absent. The cost function for such criteria can be formulated as

$$\begin{aligned}&\mathcal{J}_{\text{MWF}_\lambda-\text{SPP}}\left(\mathbf{w}(k,m)\right)\\ &= p(k,m)\\ &\quad\times E\left\{\left\|\begin{bmatrix} X_{\text{left}}(k,m) - \mathbf{w}_{\text{left}}^H(k,m)\mathbf{y}(k,m) \\ X_{\text{right}}(k,m) - \mathbf{w}_{\text{right}}^H(k,m)\mathbf{y}(k,m)\end{bmatrix}\right\|^2 |\mathcal{H}_1(k,m)\right\}\\ &\quad+(1-p(k,m))\\ &\quad\times E\left\{\left\|\begin{bmatrix} \mathbf{w}_{\text{left}}^H(k,m)\mathbf{y}(k,m) \\ \mathbf{w}_{\text{right}}^H(k,m)\mathbf{y}(k,m)\end{bmatrix}\right\|^2 |\mathcal{H}_0(k,m)\right\}\end{aligned} \quad (11)$$

where $p(k,m)$ is given by

$$p(k,m) = \frac{p_{\text{left},r_{\text{left}}}(k,m) + p_{\text{right},r_{\text{right}}}(k,m)}{2} \quad (12)$$

with $p_{\text{left},r_{\text{left}}}(k,m)$ and $p_{\text{right},r_{\text{right}}}(k,m)$ denote the conditional probability that speech is present obtained from the reference channels respectively at the left and the right, while $1 - p(k,m)$ is the combined conditional probability that speech is absent. The solution of Eq. (11) is then given by

$$\begin{aligned}&\mathbf{w}_{\text{MWF}_\lambda-\text{SPP,left}}(k,m)\\ &= \left(p(k,m)\mathbf{R}_y(k,m) + (1-p(k,m))\mathbf{R}_v(k,m)\right)^{-1}\\ &\quad\times p(k,m)\mathbf{r}_{yx,\text{left}}(k,m)\\ &\mathbf{w}_{\text{MWF}_\lambda-\text{SPP,right}}(k,m)\\ &= \left(p(k,m)\mathbf{R}_y(k,m) + (1-p(k,m))\mathbf{R}_v(k,m)\right)^{-1}\\ &\quad\times p(k,m)\mathbf{r}_{yx,\text{right}}(k,m).\end{aligned} \quad (13)$$

Based on the definition in Eq. (5), the solution of MWF_λ-SPP in Eq. (13) is very similar to the formulation in [19], which minimizes the SDWMWF cost function with SPP. However, the proposed binaural solution in this paper avoids the estimation of the speech correlation matrix $\mathbf{R}_x(k,m)$ and incorporates the cross-correlation vectors $\mathbf{r}_{yx,\text{left}}(k,m)$ and $\mathbf{r}_{yx,\text{right}}(k,m)$ instead of using only the speech correlation vectors, i.e., $\mathbf{R}_{x,\text{left}}(k,m)\mathbf{e}_{\text{left}}$ and $\mathbf{R}_{x,\text{right}}(k,m)\mathbf{e}_{\text{right}}$. Since in practice, $\mathbf{R}_x(k,m)$ is often estimated by

$$\mathbf{R}_x(k,m) = \mathbf{R}_y(k,m) - \mathbf{R}_v(k,m), \quad (14)$$

such estimates can become non-positive semi-definite at low SNR when $\mathbf{R}_v(k,m)$ becomes larger than $\mathbf{R}_y(k,m)$. Thus, instead of using the direct subtraction in the correlation matrices, this paper proposes an estimate of the cross-correlation vectors which follows the dynamic changes in the reference channels.

C. Estimation of Cross-Correlation Vector

The proposed estimates of the cross-correlation vector at both sides $\mathbf{r}_{yx,\text{left}}$ and $\mathbf{r}_{yx,\text{right}}$ are defined as

$$\begin{aligned}\hat{\mathbf{r}}_{yx,\text{left}} &= E\left\{\mathbf{y}(k,m)\mathcal{G}(k,m)Y_{\text{left}}^*(k,m)\right\}\\ \hat{\mathbf{r}}_{yx,\text{right}} &= E\left\{\mathbf{y}(k,m)\mathcal{G}(k,m)Y_{\text{right}}^*(k,m)\right\}\end{aligned} \quad (15)$$

where $\mathcal{G}(k,m)$ is a single-channel weighting gain function. The role of $\mathcal{G}(k,m)$ is to pick up all acoustic source signals from the environment and maintain those with speech while suppressing the background noise. Thus when speech is active, the value of $\mathcal{G}(k,m)$ would be close to one, and when speech is inactive, $\mathcal{G}(k,m)$ would approach the gain floor.

In practice, the mathematical expectations involved in the previous PSD matrices can be estimated by utilizing recursive smoothing. As such, the cross-correlation vectors at both sides are recursively updated by

$$\begin{aligned}\hat{\mathbf{r}}_{yx,\text{left}}(k,m) &= (1-\alpha_x)\hat{\mathbf{r}}_{yx}(k,m-1)\\ &\quad+\alpha_x\mathbf{y}(k,m)\mathcal{G}(k,m)Y_{\text{left}}^*(k,m)\\ \hat{\mathbf{r}}_{yx,\text{right}}(k,m) &= (1-\alpha_x)\hat{\mathbf{r}}_{yx}(k,m-1)\\ &\quad+\alpha_x\mathbf{y}(k,m)\mathcal{G}(k,m)Y_{\text{right}}^*(k,m)\end{aligned} \quad (16)$$

where α_x is the smoothing factor. Here, $\mathcal{G}(k,m)$ is defined as

$$\mathcal{G}(k,m) = \frac{G_{\text{left}}(k,m) + G_{\text{right}}(k,m)}{2} \quad (17)$$

where $G_{\text{left}}(k,m)$ and $G_{\text{right}}(k,m)$ are gain functions obtained for the corresponding reference channels $Y_{\text{left}}(k,m)$

and $Y_{\mathrm{right}}(k,m)$. The arithmetic mean $\mathcal{G}(k,m)$ is employed to ensure that the binaural cues can be preserved for the single-channel approach.

The gain function at each side can be computed by employing the modified sigmoid function (MSIG) developed in [18]. It is a flexible function that is capable of improving speech quality in terms of the trade-off between speech distortion, noise reduction and musical noise. As such, the gain on the left side is given by

$$G_{\mathrm{left}}(k,m) = \left\{ \varepsilon, \frac{1-\exp\left[-a_1\xi_{\mathrm{left}}(k,m)\right]}{1+\exp\left[-a_1\xi_{\mathrm{left}}(k,m)\right]} \times \left(\frac{1}{1+\exp\left(-a_2\left[\xi_{\mathrm{left}}(k,m)-c\right]\right)} \right) \right\} \tag{18}$$

where ε denotes the gain floor, and a_1, a_2, c are fixed MSIG parameters that determine the shape of the gain function. Here, $\xi_{\mathrm{left}}(k,m)$ is the *a priori* SNR in the left reference channel, which is estimated by utilizing the modified decision-directed (MDD) approach defined as [18], [20]

$$\hat{\xi}_{\mathrm{left}}(k,m) = \max\left\{ \frac{\xi_o, \beta|G_{\mathrm{left}}(k,m-1)Y_{\mathrm{left}}(k,m)|^2}{\hat{\lambda}_{v,\mathrm{left}}(k,m)} + (1-\beta)\max[\hat{\gamma}_{\mathrm{left}}(k,m)-1,0] \right\} \tag{19}$$

and $\hat{\lambda}_{v,\mathrm{left}}(k,m)$ and $\hat{X}_{\mathrm{left}}(k,m-1)$ denote, respectively, the estimated noise PSD and the estimated clean speech spectrum from the preceding frame at the left reference channel. The parameters β and ξ_o denote the smoothing factor the SNR floor, respectively. The *a posteriori* estimate is given by

$$\hat{\gamma}_{\mathrm{left}}(k,m) = \frac{\hat{\lambda}_y(k,m)}{\hat{\lambda}_v(k,m)} = \frac{\alpha_y\hat{\lambda}_y(k,m-1)+(1-\alpha_y)|Y(k,m)|^2}{\alpha_v\hat{\lambda}_v(k,m-1)+(1-\alpha_v)|Y(k,m)|^2} \tag{20}$$

where α_y and α_v are smoothing constants for speech and noise, respectively. The gain on the right side $G_{\mathrm{right}}(k,m)$ is derived similar as in Eq. (18).

D. Estimation of the Conditional SPP

Assuming a complex Gaussian distribution of the STFT coefficients for both the speech and the noise, and by applying Bayes rule, the conditional SPP for each channel, $p_l(k,m) = P(\mathcal{H}_1(k,m)|Y_l(k,m))$ is given for each frequency bin and each frame as [21], [22]

$$p_l(k,m) = \left\{ 1+\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}\left(1+\hat{\xi}_l(k,m)\right)\exp\left(-\frac{\hat{\gamma}_l(k,m)\hat{\xi}_l(k,m)}{1+\hat{\xi}_l(k,m)}\right) \right\}^{-1}. \tag{21}$$

The parameters $\hat{\xi}_l(k,m)$ and $\hat{\gamma}_l(k,m)$ indicate the estimated *a priori* SNR and the *a posteriori* SNR, respectively at l-th channel. However, instead of using the model of conditional SPP expressed in Eq. (21), an improved estimation proposed in [17] has been employed in this work. It is given by

$$p_l(k,m) = \begin{cases} \mathcal{P}_1, & p_l'(k,m) \le 0.3 \\ \mathcal{P}_2, & 0.3 < p_l'(k,m) \le 0.6 \\ \min\left\{\mathcal{P}_3, p_l'(k,m)\right\}, & p_l'(k,m) > 0.6 \end{cases} \tag{22}$$

where $\mathcal{P}_i = \exp(-2.2R)/(t_i f_{\mathrm{s}})$ is the exponential smoothing constant, with t_i, $i=[1,2,3]$ denotes the averaging time constant, with $t_1 < t_2 \ll t_3$. Here, $p_l'(k,m)$ denotes a sigmoid function defined as

$$p_l'(k,m) = \left\{1+\exp\left(-a_{\mathrm{sig}}(k,m)\left(\hat{\gamma}_l(k,m)-c_{\mathrm{sig}}(k,m)\right)\right)\right\}^{-1} \tag{23}$$

where $a_{\mathrm{sig}}(k,m)$ and $c_{\mathrm{sig}}(k,m)$ indicate, respectively, the slope and the mean of the sigmoid function. Both are given by

$$a_{\mathrm{sig}}(k,m) = \frac{\xi_a}{1+\xi_a}, \tag{24}$$

$$c_{\mathrm{sig}}(k,m) = \log\left(\frac{P(\mathcal{H}_0(k,m))}{P(\mathcal{H}_1(k,m))(1+\xi_b)}\right)\frac{1+\xi_b}{\xi_b}. \tag{25}$$

The parameters ξ_a and ξ_b are fixed values that represent the typical $a\,priori$ SNR value when speech is active. Besides providing a better SPP estimate, Eq. (22) also ensures that no stagnation would occur in Eq. (13).

E. Estimation of Speech and Noise Correlation Matrices

By following the two-state model for speech presence or absence from Eq. (10), the correlation matrices $\mathbf{R}_y(k,m)$ and $\mathbf{R}_v(k,m)$ can be recursively updated as follows. In the case when speech is not present, both correlation matrices are given by

$$\begin{aligned} \hat{\mathbf{R}}_v(k,m) &= (1-\alpha_{vv})\hat{\mathbf{R}}_v(k,m-1) \\ &\quad + \alpha_{vv}\mathbf{y}(k,m)\mathbf{y}^H(k,m) \\ \hat{\mathbf{R}}_y(k,m) &= \hat{\mathbf{R}}_y(k,m-1) \end{aligned} \tag{26}$$

and when speech is present,

$$\begin{aligned} \hat{\mathbf{R}}_v(k,m) &= \hat{\mathbf{R}}_v(k,m-1) \\ \hat{\mathbf{R}}_y(k,m) &= (1-\alpha_{yy})\,\hat{\mathbf{R}}_y(k,m-1) \\ &\quad + \alpha_{yy}\mathbf{y}(k,m)\mathbf{y}^H(k,m) \end{aligned} \tag{27}$$

where α_{vv} and α_{yy} denote the fixed smoothing factors for noise correlation matrix and speech plus noise correlation matrix, respectively. By employing the conditional SPP from Eq. (22), the two update formulas can be derived under speech presence uncertainty into the following forms

$$\begin{aligned} &\hat{\mathbf{R}}_v(k,m) \\ &= \left[p(k,m)+(1-\alpha_{vv})\left(1-p(k,m)\right)\right]\hat{\mathbf{R}}_v(k,m-1) \\ &\quad + \alpha_{vv}\left(1-p(k,m)\right)\mathbf{y}(k,m)\mathbf{y}^H(k,m) \\ &= \left(1-\tilde{\alpha}_v(k,m)\right)\hat{\mathbf{R}}_v(k,m-1) \\ &\quad + \tilde{\alpha}_v(k,m)\mathbf{y}(k,m)\mathbf{y}^H(k,m) \end{aligned} \tag{28}$$

$$\begin{aligned}&\hat{\mathbf{R}}_y(k,m)\\&=\left[(1-p(k,m))+p(k,m)\left(1-\alpha_{yy}\right)\right]\hat{\mathbf{R}}_y(k,m-1)\\&\quad+p(k,m)\alpha_{yy}\mathbf{y}(k,m)\mathbf{y}^H(k,m)\\&=\left(1-\tilde{\alpha}_y(k,m)\right)\hat{\mathbf{R}}_y(k,m-1)\\&\quad+\tilde{\alpha}_y(k,m)\mathbf{y}(k,m)\mathbf{y}^H(k,m)\end{aligned}\tag{29}$$

where $\mathbf{R}_y(k,m)$ and $\mathbf{R}_v(k,m)$ can be updated continuously over time. Here, $\tilde{\alpha}_v(k,m)$ and $\tilde{\alpha}_y(k,m)$ denote, respectively, $\tilde{\alpha}_v(k,m)=\alpha_{vv}(1-p(k,m))$ and $\tilde{\alpha}_y(k,m)=\alpha_{yy}p(k,m)$. The values of both α_{vv} and α_{yy} are chosen carefully to reflect the degree of stationarity of speech and noise signals.

III. Theoretical Analysis for Cue Preservation and SNR Improvement

A. Optimal Filters Under Single Speech Source Case

To examine the SNR performance and the ability of binaural MWF_λ-SPP in maintaining binaural cues, firstly we need to observe the behavior of the cross-correlation vectors $\hat{\mathbf{r}}_{yx,\mathrm{left}}(k,m)$ and $\hat{\mathbf{r}}_{yx,\mathrm{right}}(k,m)$. Under speech presence uncertainty, Eq. (15) at the left side becomes

$$\begin{aligned}&\hat{\mathbf{r}}_{yx,\mathrm{left}}(k,m)\\&=\begin{cases}E\left\{\mathbf{y}(k,m)\mathcal{G}(k,m)Y^*_{\mathrm{left}}(k,m)|\mathcal{H}_1(k,m)\right\}, & \mathcal{H}_1(k,m)\\E\left\{\mathbf{y}(k,m)\mathcal{G}(k,m)Y^*_{\mathrm{left}}(k,m)|\mathcal{H}_0(k,m)\right\}, & \mathcal{H}_0(k,m)\end{cases}\end{aligned}\tag{30}$$

where

$$\begin{aligned}&E\left\{\mathbf{y}(k,m)\mathcal{G}(k,m)Y^*_{\mathrm{left}}(k,m)|\mathcal{H}_1(k,m)\right\}\\&=E\left\{\mathcal{G}(k,m)\left(\mathbf{x}(k,m)X^*_{\mathrm{left}}(k,m)+\mathbf{v}(k,m)V^*_{\mathrm{left}}(k,m)\right)\right\}\\&=\hat{\mathbf{r}}_{xx,\mathrm{left}}(k,m)+\hat{\mathbf{r}}_{vv,\mathrm{left}}(k,m),\\&\text{and}\\&E\left\{\mathbf{y}(k,m)\mathcal{G}(k,m)Y^*_{\mathrm{left}}(k,m)|\mathcal{H}_0(k,m)\right\}\\&=\varepsilon\hat{\mathbf{r}}_{vv,\mathrm{left}}(k,m).\end{aligned}\tag{31}$$

This implies that when speech is present, the solution of binaural MWF_λ-SPP in Eq. (13) becomes

$$\begin{aligned}&\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,left}}(k,m)\\&=\left(p(k,m)\hat{\mathbf{R}}_y(k,m)+(1-p(k,m))\hat{\mathbf{R}}_v(k,m)\right)^{-1}\\&p(k,m)\left(\hat{\mathbf{r}}_{xx,\mathrm{left}}(k,m)+\hat{\mathbf{r}}_{vv,\mathrm{left}}(k,m)\right)\\&\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,right}}(k,m)\\&=\left(p(k,m)\hat{\mathbf{R}}_y(k,m)+(1-p(k,m))\hat{\mathbf{R}}_v(k,m)\right)^{-1}\\&p(k,m)\left(\hat{\mathbf{r}}_{xx,\mathrm{right}}(k,m)+\hat{\mathbf{r}}_{vv,\mathrm{right}}(k,m)\right)\end{aligned}\tag{32}$$

while when speech is absent, $R_y(k,m)=R_v(k,m)$ and the solution of binaural MWF_λ-SPP is given by

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,left}}(k,m)&=\varepsilon p(k,m)\hat{\mathbf{R}}_v^{-1}(k,m)\hat{\mathbf{r}}_{vv,\mathrm{left}}(k,m)\\\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,right}}(k,m)&=\varepsilon p(k,m)\hat{\mathbf{R}}_v^{-1}(k,m)\hat{\mathbf{r}}_{vv,\mathrm{right}}(k,m)\end{aligned}\tag{33}$$

where ε is the floor used in Eq. (18).

In the case of a single target speech source, the speech signal vector can be modeled as

$$\mathbf{x}(k,m)=\mathbf{a}(k,m)S(k,m)\tag{34}$$

where S is the target speech signal. The L-dimensional stacked vector $\mathbf{a}(k,m)$ is given by

$$\mathbf{a}(k,m)=\begin{bmatrix}\mathbf{a}_{\mathrm{left}}(k,m)\\\mathbf{a}_{\mathrm{right}}(k,m)\end{bmatrix},\tag{35}$$

with

$$\begin{aligned}&\mathbf{a}_{\mathrm{left}}(k,m)\\&=\left[A_{\mathrm{left}},1(k,m)\quad A_{\mathrm{left},2}(k,m)\ldots A_{\mathrm{left},L}(k,m)\right]^T,\\&\mathbf{a}_{\mathrm{right}}(k,m)\\&=\left[A_{\mathrm{right},1}(k,m)A_{\mathrm{right},2}(k,m)\ldots A_{\mathrm{right},L}(k,m)\right]^T.\end{aligned}\tag{36}$$

The speech correlation matrix is a rank-one matrix, i.e.,

$$\mathbf{R}_x(k,m)=\Phi_{\mathrm{ss}}(k,m)\mathbf{a}(k,m)\mathbf{a}^H(k,m)\tag{37}$$

with $\Phi_{\mathrm{ss}}(k,m)=E\{|S(k,m)|^2\}$ representing the PSD of the speech signal.

Using Eq. (32), in the case of a single speech source, where both speech and noise are present, the optimal filters of the binaural MWF_λ-SPP are obtained by applying the matrix inversion lemma as

$$\begin{aligned}&\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,left}}(k,m)\\&=\left(1-\frac{1}{\psi(k,m)}\right)\frac{\Phi_{\mathrm{ss}}(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)A^*_{\mathrm{left}}(k,m)}{\psi(k,m)+\varrho(k,m)}\\&\quad+\frac{1}{\psi(k,m)}\mathbf{e}_{\mathrm{left}}\\&\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,right}}(k,m)\\&=\left(1-\frac{1}{\psi(k,m)}\right)\frac{\Phi_{\mathrm{ss}}(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)A^*_{\mathrm{right}}(k,m)}{\psi(k,m)+\varrho(k,m)}\\&\quad+\frac{1}{\psi(k,m)}\mathbf{e}_{\mathrm{right}}.\end{aligned}\tag{38}$$

where

$$\psi(k,m)=\frac{1}{p(k,m)}\tag{39}$$

$$\begin{aligned}A^*_{\mathrm{left}}(k,m)&=\mathbf{a}^H(k,m)\mathbf{e}_{\mathrm{left}}\\A^*_{\mathrm{right}}(k,m)&=\mathbf{a}^H(k,m)\mathbf{e}_{\mathrm{right}}\end{aligned}\tag{40}$$

$$\begin{aligned}\varrho(k,m)&=\Phi_{\mathrm{ss}}(k,m)\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\\&\quad\times\mathbf{a}(k,m).\end{aligned}\tag{41}$$

Meanwhile, when only noise is present, the optimal filters is derived from Eq. (33) into

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,left}}(k,m)&=\frac{\varepsilon}{\psi(k,m)}\mathbf{e}_{\mathrm{left}}\\\mathbf{w}_{\mathrm{MWF}_\lambda-\mathrm{SPP,right}}(k,m)&=\frac{\varepsilon}{\psi(k,m)}\mathbf{e}_{\mathrm{right}}.\end{aligned}\tag{42}$$

B. Theoretical Performance Measures

Here, definitions of both SNR improvement and interaural transfer functions for cues preservation are given for the purpose of the theoretical performance measurement [23]. The narrowband input SNR is defined as the power ratio of speech and noise component in the reference microphones at both sides, as

$$\begin{aligned}\mathrm{SNR}^{\mathrm{in}}_{\mathrm{left}}(k,m) &= \frac{E|X_{\mathrm{left}}(k,m)|^2}{E|V_{\mathrm{left}}(k,m)|^2} = \frac{\mathbf{e}^H_{\mathrm{left}}\mathbf{R}_x(k,m)\mathbf{e}_{\mathrm{left}}}{\mathbf{e}^H_{\mathrm{left}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{left}}}\\ \mathrm{SNR}^{\mathrm{in}}_{\mathrm{right}}(k,m) &= \frac{E|X_{\mathrm{right}}(k,m)|^2}{E|V_{\mathrm{right}}(k,m)|^2} = \frac{\mathbf{e}^H_{\mathrm{right}}\mathbf{R}_x(k,m)\mathbf{e}_{\mathrm{right}}}{\mathbf{e}^H_{\mathrm{right}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{right}}}\end{aligned} \tag{43}$$

and the narrowband output SNR is given by the power ratio of speech and noise component in the output signals

$$\begin{aligned}&\mathrm{SNR}^{\mathrm{out}}_{\mathrm{left}}(k,m)\\ &\quad= \frac{E|Z_{x,\mathrm{left}}(k,m)|^2}{E|Z_{v,\mathrm{left}}(k,m)|^2}\\ &\quad= \frac{\mathbf{w}^H_{\mathrm{left}}(k,m)\mathbf{R}_x(k,m)\mathbf{w}_{\mathrm{left}}(k,m)}{\mathbf{w}^H_{\mathrm{left}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{\mathrm{left}}(k,m)}\\ &\mathrm{SNR}^{\mathrm{out}}_{\mathrm{right}}(k,m)\\ &\quad= \frac{E|Z_{x,\mathrm{right}}(k,m)|^2}{E|Z_{v,\mathrm{right}}(k,m)|^2}\\ &\quad= \frac{\mathbf{w}^H_{\mathrm{right}}(k,m)\mathbf{R}_x(k,m)\mathbf{w}_{\mathrm{right}}(k,m)}{\mathbf{w}^H_{\mathrm{right}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{\mathrm{right}}(k,m)}.\end{aligned} \tag{44}$$

Thus, the SNR improvement at each side is given by

$$\begin{aligned}\Delta\mathrm{SNR}_{\mathrm{left}}(k,m) &= \frac{\mathrm{SNR}^{\mathrm{out}}_{\mathrm{left}}(k,m)}{\mathrm{SNR}^{\mathrm{in}}_{\mathrm{left}}(k,m)}\\ \Delta\mathrm{SNR}_{\mathrm{right}}(k,m) &= \frac{\mathrm{SNR}^{\mathrm{out}}_{\mathrm{right}}(k,m)}{\mathrm{SNR}^{\mathrm{in}}_{\mathrm{right}}(k,m)}.\end{aligned} \tag{45}$$

The input and output ITFs of the speech component are defined as the ratio of the component in the reference microphones. In the case of single speech source, they are given by

$$\begin{aligned}\mathrm{ITF}^{\mathrm{in}}_x(k,m) &= \frac{X_{\mathrm{left}}(k,m)}{X_{\mathrm{right}}(k,m)}\\ &= \frac{\mathbf{e}^H_{\mathrm{left}}\mathbf{a}(k,m)}{\mathbf{e}^H_{\mathrm{right}}\mathbf{a}(k,m)} = \frac{A_{\mathrm{left}}(k,m)}{A_{\mathrm{right}}(k,m)},\end{aligned} \tag{46}$$

$$\begin{aligned}\mathrm{ITF}^{\mathrm{out}}_x(k,m) &= \frac{Z_{x,\mathrm{left}}(k,m)}{Z_{x,\mathrm{right}}(k,m)}\\ &= \frac{\mathbf{w}^H_{\mathrm{left}}(k,m)\mathbf{a}(k,m)}{\mathbf{w}^H_{\mathrm{right}}(k,m)\mathbf{a}(k,m)}\end{aligned} \tag{47}$$

while the input and output ITFs of the noise component are defined as

$$\begin{aligned}\mathrm{ITF}^{\mathrm{in}}_v(k,m) &= \frac{V_{\mathrm{left}}(k,m)}{V_{\mathrm{right}}(k,m)}\\ &= \frac{\mathbf{e}^H_{\mathrm{left}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{right}}}{\mathbf{e}^H_{\mathrm{right}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{right}}}\\ &= \frac{\mathbf{e}^H_{\mathrm{left}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{left}}}{\mathbf{e}^H_{\mathrm{right}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{left}}}\end{aligned} \tag{48}$$

$$\begin{aligned}\mathrm{ITF}^{\mathrm{out}}_v(k,m) &= \frac{Z_{v,\mathrm{left}}(k,m)}{Z_{v,\mathrm{right}}(k,m)}\\ &= \frac{\mathbf{w}^H_{\mathrm{left}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{\mathrm{right}}(k,m)}{\mathbf{w}^H_{\mathrm{right}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{\mathrm{right}}(k,m)}\\ &= \frac{\mathbf{w}^H_{\mathrm{left}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{left}(k,m)}{\mathbf{w}^H_{\mathrm{right}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{\mathrm{left}}(k,m)}.\end{aligned} \tag{49}$$

C. Comparison Between Proposed Method and SDWMWF

The binaural SDWMWF method is included for performance comparison. The SDWMWF minimizes a weighted sum of the residual noise energy and the speech distortion energy in order to provide a trade-off between speech distortion and noise reduction [14]. The binaural SDWMWF[1] cost function is equal to

$$\begin{aligned}&\mathcal{J}_{\mathrm{MWF}_\mu}\left(\mathbf{w}(k,m)\right)\\ &= E\left\{\left\|\begin{bmatrix} X_{\mathrm{left}}(k,m) - \mathbf{w}^H_{\mathrm{left}}(k,m)\mathbf{x}(k,m)\\ X_{\mathrm{right}}(k,m) - \mathbf{w}^H_{\mathrm{right}}(k,m)\mathbf{x}(k,m)\end{bmatrix}\right\|^2\right.\\ &\quad\left. - \mu\left\|\begin{bmatrix}\mathbf{w}^H_{\mathrm{left}}(k,m)\mathbf{v}(k,m)\\ \mathbf{w}^H_{\mathrm{right}}(k,m)\mathbf{v}(k,m)\end{bmatrix}\right\|^2\right\}\end{aligned} \tag{50}$$

where μ provides a trade-off between reduction and speech distortion. The optimal MWF_μ filters for the respective sides are equal to

$$\begin{aligned}&\mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{left}}(k,m)\\ &= \left(\mathbf{R}_x(k,m) + \mu\mathbf{R}_v(k,m)\right)^{-1}\mathbf{R}_x(k,m)\mathbf{e}_{\mathrm{left}},\\ &\mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{right}}(k,m)\\ &= \left(\mathbf{R}_x(k,m) + \mu\mathbf{R}_v(k,m)\right)^{-1}\mathbf{R}_x(k,m)\mathbf{e}_{\mathrm{right}}.\end{aligned} \tag{51}$$

By assuming a single speech source and by applying the matrix inversion lemma, it has been shown in [14] that Eq. (51) can be reduced to the following optimal filter at each side:

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{left}}(k,m) &= \mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)\\ &\quad\cdot\frac{\Phi_{\mathrm{ss}}(k,m)A^*_{left}(k,m)}{\mu+\varrho(k,m)}\\ \mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{right}}(k,m) &= \mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)\\ &\quad\cdot\frac{\Phi_{\mathrm{ss}}(k,m)A^*_{\mathrm{right}}(k,m)}{\mu+\varrho(k,m)}.\end{aligned} \tag{52}$$

To examine the ability of binaural MWF_μ in maintaining binaural cues, the single target source optimal filters from Eq. (52) are utilized. From there, the binaural MWF_μ vectors for the left and the right sides of the devices are found to be parallel, such that

$$\mathbf{w}_{\mathrm{MWF}_\lambda,\mathrm{left}}(k,m) = \mathrm{ITF}^{\mathrm{in},*}_x(k,m)\mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{right}}(k,m) \tag{53}$$

where $\mathrm{ITF}^{\mathrm{in}}_x(k,m)$ is given in Eq. (46). Hence, the ITFs of the output speech and noise components are both equal to $\mathrm{ITF}^{\mathrm{in}}_x(k,m)$, implying that the binaural speech cues are perfectly preserved, but the binaural noise cues are distorted. As all output components are perceived as coming from the speech direction, the auditory perception of the acoustic scene is therefore not preserved by the binaural MWF_μ.

[1]For conciseness, SDWMWF is abbreviated to MWF_μ in the equations.

Since the solutions of binaural MWF_μ are parallel, by using the definition in Eq. (44) the narrowband output SNR for each side will be the same, such that

$$\begin{aligned}\mathrm{SNR}^{\mathrm{out}}_{\mathrm{left}}(k,m) &= \mathrm{SNR}^{\mathrm{out}}_{\mathrm{right}}(k,m)\\ &= \frac{E|Z_{x,\mathrm{left}}(k,m)|^2}{E|Z_{v,\mathrm{left}}(k,m)|^2}\\ &= \frac{|\mathbf{w}^H_{\mathrm{MWF}_\mu,\mathrm{left}}(k,m)\mathbf{a}(k,m)|^2\Phi_{ss}}{\mathbf{w}^H_{\mathrm{MWF}_\mu,\mathrm{left}}(k,m)\mathbf{R}_v\mathbf{w}_{\mathrm{MWF}_\mu,\mathrm{left}}(k,m)}\\ &= \Phi_{ss}(k,m)\mathbf{a}^H(k,m)\mathbf{R}_v^{-1}(k,m)\mathbf{a}(k,m)\\ &= \varrho(k,m). \end{aligned} \tag{54}$$

From Eq. (45), it follows that the SNR improvement at each respective side can be obtained as

$$\begin{aligned}\Delta\mathrm{SNR}_{\mathrm{left}}(k,m) &= \frac{\varrho(k,m)\mathbf{e}^H_{\mathrm{left}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{left}}}{\Phi_{ss}|A_{\mathrm{left}}|^2}\\ \Delta\mathrm{SNR}_{\mathrm{right}}(k,m) &= \frac{\varrho(k,m)\mathbf{e}^H_{\mathrm{right}}\mathbf{R}_v(k,m)\mathbf{e}_{\mathrm{right}}}{\Phi_{ss}|A_{\mathrm{right}}|^2}. \end{aligned} \tag{55}$$

This implies that the SNR improvements are directly related to the noise correlation matrix $\mathbf{R}_v(k,m)$ and speech correlation matrix $\mathbf{R}_x(k,m)$. As a matter of fact, they are related to their estimates and the accuracy of the model. Hence it is important that the estimates of speech and noise correlation matrices provide an accurate reflection of the true noise correlation, speech power and the ATF of the target speech signal.

For the proposed binaural MWF_λ-SPP, the narrowband output SNR for each side can be obtained by using the definition in Eq. (43) and Eq. (44) on the optimal filters obtained under the case when both speech and noise are present, i.e. Eq. (38). It is given by

$$\begin{aligned}\mathrm{SNR}^{\mathrm{out}}_{\mathrm{left}}(k,m) &= \frac{\psi(k,m)^2(\varrho(k,m)+1)^2\mathrm{SNR}^{\mathrm{in}}_{\mathrm{left}}(k,m)}{\zeta(k,m)\mathrm{SNR}^{\mathrm{in}}_{\mathrm{left}}(k,m)+(\psi(k,m)+\varrho(k,m))^2}\\ \mathrm{SNR}^{\mathrm{out}}_{\mathrm{right}}(k,m) &= \frac{\psi(k,m)^2(\varrho(k,m)+1)^2\mathrm{SNR}^{\mathrm{in}}_{\mathrm{right}}(k,m)}{\zeta(k,m)\mathrm{SNR}^{\mathrm{in}}_{\mathrm{right}}(k,m)+(\psi(k,m)+\varrho(k,m))^2} \end{aligned} \tag{56}$$

where

$$\zeta(k,m) = \psi^2(k,m)\varrho(k,m)+2\psi^2(k,m)-\psi(k,m)-\varrho(k,m). \tag{57}$$

By using Eq. (45), the SNR improvement on the left side is defined as

$$\begin{aligned}&\Delta\mathrm{SNR}_{\mathrm{left}}(k,m)\\ &= \frac{\left(\frac{\varrho(k,m)+1}{\psi(k,m)+\varrho(k,m)}\right)^2\Delta\mathrm{SNR}^{\mu}_{\mathrm{left}}(k,m)}{\left(\frac{\varrho(k,m)+1}{\psi(k,m)+\varrho(k,m)}\right)^2+\frac{1}{\psi^2(k,m)}\left(\Delta\mathrm{SNR}^{\mu}_{\mathrm{left}}(k,m)-1\right)}\end{aligned} \tag{58}$$

where $\Delta\mathrm{SNR}^{\mu}_{\mathrm{left}}(k,m)$ is the SNR improvement of the binaural SDWMWF on the left side (Eq. (55)). The SNR improvement on the right side is defined similarly to Eq. (58). Since the conditional SPP $p(k,m)$ has only values in between 0 and 1, for Eq. (58), if $p(k,m) = 1/\psi(k,m) = 0$, the SNR improvement is equal to $\Delta\mathrm{SNR}^{\mu}_{\mathrm{left}}(k,m)$. If $p(k,m) = 1/\psi(k,m) = 1$, the SNR improvement is equal to 1, which simply means no improvement. Since a large value of $p(k,m)$ generally indicates speech component is dominant over noise, no SNR improvement at $p(k,m) = 1$ indicates that the proposed algorithm offers less speech distortion in the output signals. By substituting Eq. (38) to Eq. (47), the ITF of the output speech component for the binaural MWF_λ-SPP is equal to

$$\begin{aligned}&\mathrm{ITF}^{\mathrm{out}}_x(k,m)\\ &= \frac{\mathbf{w}^H_{\mathrm{MWF}_\lambda-\mathrm{SPP},\mathrm{left}}(k,m)\mathbf{a}(k,m)}{\mathbf{w}^H_{\mathrm{MWF}_\lambda-\mathrm{SPP},\mathrm{right}}(k,m)\mathbf{a}(k,m)}\\ &= \frac{A_{\mathrm{left}}(k,m)\left((1-\frac{1}{\psi(k,m)})\frac{\varrho(k,m)}{\psi(k,m)+\varrho(k,m)}+\frac{1}{\psi(k,m)}\right)}{A_{\mathrm{right}}(k,m)\left((1-\frac{1}{\psi(k,m)})\frac{\varrho(k,m)}{\psi(k,m)+\varrho(k,m)}+\frac{1}{\psi(k,m)}\right)}\\ &= \frac{A_{\mathrm{left}}(k,m)}{A_{\mathrm{right}}(k,m)}\end{aligned} \tag{59}$$

thus $\mathrm{ITF}^{\mathrm{out}}_x(k,m) = \mathrm{ITF}^{\mathrm{in}}_x(k,m)$, which means the ITF of speech component is preserved when speech is present. In order to find the ITF of the output noise component, both Eq. (38) and Eq. (42) are utilized. By substituting them into Eq. (49), the noise ITFs are defined as

$$\begin{aligned}&\mathrm{ITF}^{\mathrm{out}}_v(k,m)\\ &= \frac{\mathbf{w}^H_{\mathrm{MWF}_\lambda\text{–}\mathrm{SPP},\mathrm{left}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{\mathrm{MWF}_\lambda\text{–}\mathrm{SPP},\mathrm{right}}(k,m)}{\mathbf{w}^H_{\mathrm{MWF}_\lambda\text{–}\mathrm{SPP},\mathrm{right}}(k,m)\mathbf{R}_v(k,m)\mathbf{w}_{\mathrm{MWF}_\lambda\text{–}\mathrm{SPP},\mathrm{right}}(k,m)}\\ &= \frac{\varsigma(k,m)SNR^{\mathrm{in}}_{\mathrm{right}}(k,m)}{\varsigma(k,m)SNR^{\mathrm{in}}_{\mathrm{right}}(k,m)+1}ITF^{\mathrm{in}}_x(k,m)\\ &\quad+\frac{1}{\varsigma(k,m)SNR^{\mathrm{in}}_{\mathrm{right}}(k,m)+1}ITF^{\mathrm{in}}_v(k,m)\end{aligned} \tag{60}$$

where

$$\varsigma(k,m)=\begin{cases}\frac{(\psi(k,m)-1)^2\varrho(k,m)}{(\psi(k,m)+\varrho(k,m))^2}+\frac{2(\psi(k,m)-1)}{\psi(k,m)+\varrho(k,m)} & \mathcal{H}_1(k,m)\\ 0 & \mathcal{H}_0(k,m)\end{cases} \tag{61}$$

Eq. (60) shows that when speech is present, $\mathrm{ITF}^{\mathrm{out}}_v(k,m)$ is a weighted sum of $\mathrm{ITF}^{\mathrm{in}}_x(k,m)$ and $\mathrm{ITF}^{\mathrm{in}}_v(k,m)$, which relies on the input SNR. When the input SNR is significantly large, $\mathrm{ITF}^{\mathrm{in}}_v(k,m)$ will be distorted by $\mathrm{ITF}^{\mathrm{in}}_x(k,m)$. However, if the input SNR is sufficient small, i.e. $SNR^{\mathrm{in}}_{\mathrm{left}}(k,m) = SNR^{\mathrm{in}}_{\mathrm{right}}(k,m) = 0$, $\mathrm{ITF}^{\mathrm{out}}_v(k,m)$ is equal to $\mathrm{ITF}^{\mathrm{in}}_v(k,m)$, which means the noise ITF is preserved. In the case when only noise is present, the noise cue is also preserved.

IV. Performance Measures Under Real Environment

A hearing protection device, as shown in Fig. 1 with two-microphone array (with inter-microphone space of 1 cm) being mounted on each side, was utilized for processing. Performance evaluation for speech quality includes the comparison of the binaural cues and the noise reduction performance. The noise reduction performance is measured by the intelligibility frequency weighted segmental SNR (IFWSNRseg) measure [24], [25]

$$\mathrm{IFWSNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{k=0}^{K-1}B_k\log_{10}\frac{\mathcal{A}^2(k,m)}{\mathcal{A}^2(k,m)-\check{\mathcal{A}}^2(k,m)}}{\sum_{k=0}^{K-1}B_k} \tag{62}$$

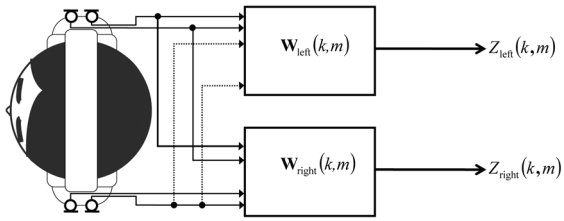

Fig. 1. A hearing protection device with $L = 2$ microphones (with inter-element space of 1 cm) embedded on each side of a pair of earmuffs. The microphone signals were processed with the reference and the proposed MWF approaches to produce the output signals.

where B_k is the ANSI SII weight placed on the k-th frequency bin [26], K is the number of bands, M is the number of frames, $\mathcal{A}(k,m)$ and $\hat{\mathcal{A}}(k,m)$ are spectrum amplitudes of the clean speech signal and enhanced speech signal, respectively. Each frame has a threshold of -10 dB lower bound and a 35 dB upper bound to discard non-speech frames.

In addition, the segmental noise attenuation (NATTseg) and the segmental speech preservation (SPREseg) measures are utilized to study if a difference in IFWSNRseg is due to more noise reduction or less speech distortion. Both are given, respectively, by [27]

$$\text{NATTseg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{\|\mathbf{v}_t(m)\|^2}{\|\tilde{\mathbf{v}}_t(m)\|^2}, \tag{63}$$

and

$$\text{SPREseg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{\|\mathbf{x}_t(m)\|^2}{\|\mathbf{x}_t(m)-\tilde{\mathbf{x}}_t(m)\|^2}. \tag{64}$$

Here, $\mathbf{v}_t(m)$ and $\mathbf{x}_t(m)$ are m-th frame time-domain vectors for the noise and the clean speech signal, respectively. The signals $\tilde{\mathbf{v}}_t(m)$ and $\tilde{\mathbf{x}}_t(m)$ indicate both noise and the clean signals processed with the same corresponding filters as used to enhance the noisy signal. The widely-used perceptual evaluation of speech quality (PESQ) measure has also been included for performance comparison [25]. For all measurements, results of the reference channels from the left and right were averaged to obtain a single value.

The binaural cues were evaluated using the ITD and the ILD measures. The ITDs here are computed using the cross-correlation, which is commonly used to estimate time delays, as defined by

$$\begin{aligned}
R_{x_{\text{left}}x_{\text{right}}}(\eta) &= E\left(x_{\text{left}}(n)x_{\text{right}}(n-\eta)\right)\\
R_{v_{\text{left}}v_{\text{right}}}(\eta) &= E\left(v_{\text{left}}(n)v_{\text{right}}(n-\eta)\right)\\
R_{\tilde{x}_{\text{left}}\tilde{x}_{\text{right}}}(\eta) &= E\left(\tilde{x}_{\text{left}}(n)\tilde{x}_{\text{right}}(n-\eta)\right)\\
R_{\tilde{v}_{\text{left}}\tilde{v}_{\text{right}}}(\eta) &= E\left(\tilde{v}_{\text{left}}(n)\tilde{v}_{\text{right}}(n-\eta)\right).
\end{aligned} \tag{65}$$

where $x_{\text{left}}(n)$, $\tilde{x}_{\text{left}}(n)$, $v_{\text{left}}(n)$, $\tilde{v}_{\text{left}}(n)$ denote, respectively, the reference signals and the output signals for speech and noise in discrete time domain on the left side. Similar notations are used for the right side. Note that the front microphones (microphones located nearer to the direction where the user is facing, as shown in Fig. 1) at respective sides are used as the reference signals. This is the same for all the performance measures that require the access of the clean signals and/or the observed signals. The delay is then given by the argument $\eta = \eta_0$, which yields the maximum absolute value of Eq. (65), as

$$\begin{aligned}
\text{ITD}_x^{\text{in}} &= \arg\max_{\eta}\left[R_{x_{\text{left}}x_{\text{right}}}(\eta)\right]\\
\text{ITD}_v^{\text{in}} &= \arg\max_{\eta}\left[R_{v_{\text{left}}v_{\text{right}}}(\eta)\right]\\
\text{ITD}_x^{\text{out}} &= \arg\max_{\eta}\left[R_{\tilde{x}_{\text{left}}\tilde{x}_{\text{right}}}(\eta)\right]\\
\text{ITD}_v^{\text{out}} &= \arg\max_{\eta}\left[R_{\tilde{v}_{\text{left}}\tilde{v}_{\text{right}}}(\eta)\right].
\end{aligned} \tag{66}$$

Since η_0 has integer values only and the delay is usually fractional, the cross-correlation function needs to be interpolated. After that, the delay τ_0 in seconds is obtained by dividing η_0 by the sampling frequency f_{s}. The absolute ITD errors are then given by [28]

$$\begin{aligned}
\Delta\text{ITD}_x &= |\text{ITD}_x^{\text{in}} - \text{ITD}_x^{\text{out}}|\\
\Delta\text{ITD}_v &= |\text{ITD}_v^{\text{in}} - \text{ITD}_v^{\text{out}}|.
\end{aligned} \tag{67}$$

The ILDs are obtained by evaluating the logarithm of the power ratio between the respective signals of the left and right side. The ILD errors of speech and noise are given as [28]

$$\begin{aligned}
&\Delta\text{ILD}_x\\
&= \frac{1}{I}\sum\nolimits_{i=1}^{I} 10\log_{10}\left(\frac{|X_{\text{left}}(k,m)|^2}{|X_{\text{right}}(k,m)|^2} - \frac{|Z_{x,\text{left}}(k,m)|^2}{|Z_{x,\text{right}}(k,m)|^2}\right)^2\\
&\Delta\text{ILD}_v\\
&= \frac{1}{I}\sum\nolimits_{i=1}^{I} 10\log_{10}\left(\frac{|V_{\text{left}}(k,m)|^2}{|V_{\text{right}}(k,m)|^2} - \frac{|Z_{v,\text{left}}(k,m)|^2}{|Z_{v,\text{right}}(k,m)|^2}\right)^2
\end{aligned} \tag{68}$$

where I is the signal length in samples.

V. Performance Evaluation

For a fair performance comparison, the conditional SPP is also applied to replace the trade-off parameter of the conventional SDWMWF function in Eq. (51), such that $\mu(k,m) = 1/p(k,m)$. In addition, the minimum variance distortionless response (MVDR) beamformer is included in this section for comparison [29], [30]. Here, the binaural MVDR filter in the form without explicit dependence on all ATFs is derived in a binaural configuration, such that

$$\begin{aligned}
\mathbf{w}_{\text{MVDR,left}}(k,m) &= \frac{\mathbf{R}_v(k,m)^{-1}\mathbf{R}_x(k,m)\mathbf{e}_{\text{left}}}{\text{Tr}\left(\mathbf{R}_v(k,m)^{-1}\mathbf{R}_x(k,m)\right)},\\
\mathbf{w}_{\text{MVDR,right}}(k,m) &= \frac{\mathbf{R}_v(k,m)^{-1}\mathbf{R}_x(k,m)\mathbf{e}_{\text{right}}}{\text{Tr}\left(\mathbf{R}_v(k,m)^{-1}\mathbf{R}_x(k,m)\right)}.
\end{aligned} \tag{69}$$

where Tr denotes the trace operator. As such, the overall performance of the binaural MWF_μ-SPP and the binaural MWF_λ-SPP, together with the binaural MSIG function and the binaural MVDR, is evaluated and compared. The setup for the underlying measurements is depicted in Fig. 2. A manikin with put-on earmuffs as depicted in Fig. 1 is placed close to the center of a room with dimensions 3.05 m $\times$ 3.05 m, with a reverberation time T_{60} of approximately 0.2 s. The loudspeakers are positioned 1 m from the center of the head, with speech and noise rendered at different positions around the head to create point source sounds. An additional four loudspeakers

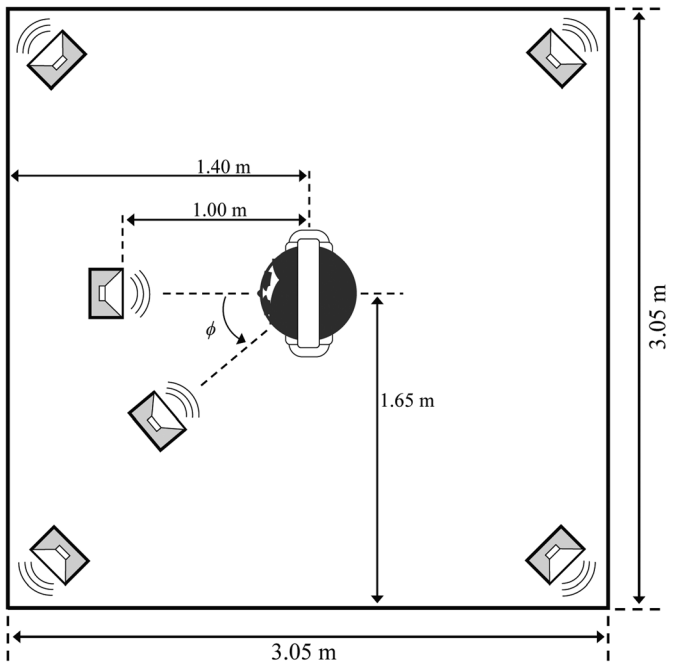

Fig. 2. Measurement setup for the evaluation of the binaural noise reduction techniques.

TABLE I
PARAMETER SETTINGS

Parameters	Values
Smoothing constant for speech PSD $\hat{\lambda}_y$	$\alpha_y = 0$
Smoothing constant for noise PSD $\hat{\lambda}_v$	$\alpha_v = 0.8$
Smoothing constant for ξ_{MDD}	$\beta = 0.9$
MSIG parameters	$a_1 = 3$, $a_2 = 1$, $c = 0.7$
Spectral noise floor	$\epsilon = -15$ dB
Smoothing constant for $\hat{\mathbf{R}}_v$	$\alpha_{vv} = 0.98$
Smoothing constant for $\hat{\mathbf{R}}_y$ and $\hat{\mathbf{r}}_{yx}$	$\alpha_{yy} = \alpha_x = 0.17$
SVAD parameters for noise estimate [17]	$P(\mathcal{H}_0) = 0.3$
	$P(\mathcal{H}_1) = 0.7$
	$\xi_a = \xi_b = 12$ dB
	$t_1 = 0.05$
	$t_2 = 0.08$
	$t_3 = 240$

are placed in each corner of the room facing the walls to create diffuse-like background noise.

The speech signals consist of 7 (4 male and 3 female) sentences with lengths ranging from 11 s to 22 s, with the noise sources being industrial noises. For evaluation purposes, the speech and noise signals were recorded separately. The processing was done with a sampling frequency of 16 kHz using an STFT with the square root of a Hann window, both for analysis and synthesis, frame length $K = 512$, and 50% overlap, i.e., $R = 256$. The parameters of both frameworks are given in Table I, which were determined empirically from previous works. The algorithms were not adjusted to obtain the largest amount of noise suppression, but rather to achieve a good trade-off among the amount of noise suppression, speech distortion, and musical noise generated from the processing.

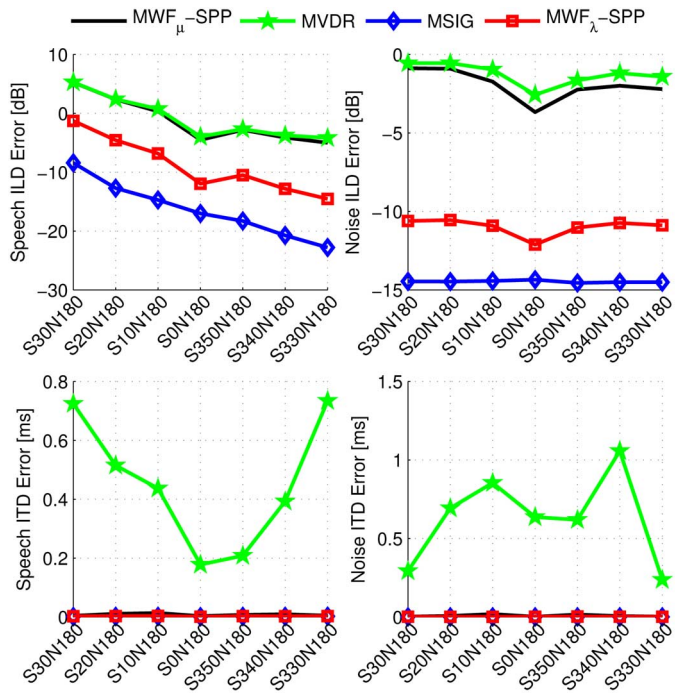

Fig. 3. ITD and ILD results for MWF when direction of noise was fixed with speech source coming from different directions.

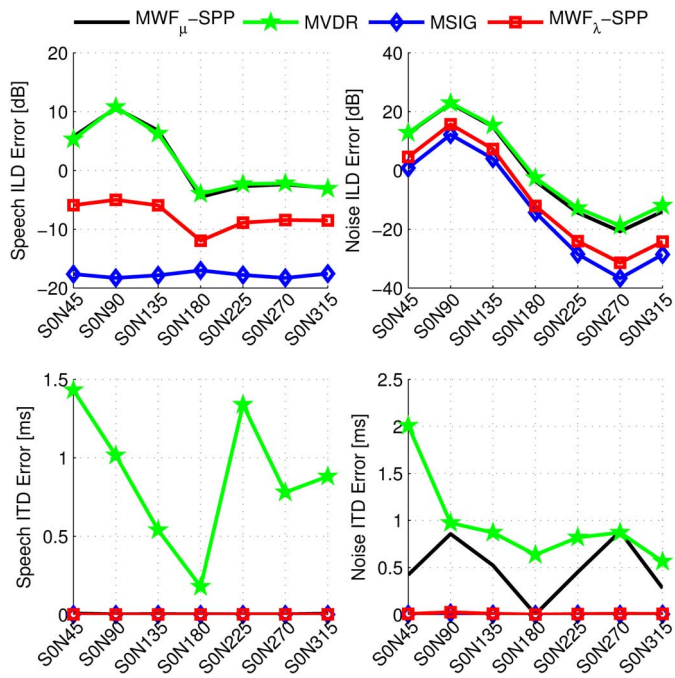

Fig. 4. ITD and ILD results for MWF when direction of the target speech was fixed with noise source coming from different directions.

For evaluation of the ITD and ILD, two scenarios were considered. One is a speech source from the front of the head with several noise configurations delivered at 45°, 90°, 135°, 180°, 225°, 270° and 315° with respect to the left of the head. The latter scenario had the noise source originated from behind the head with several speech configurations at 10°, 20°, 30°, 330°, 340°, and 350°. The reason that only 6 configurations of speech source were recorded is due to the assumption that the target speech is often coming from the frontal directions.

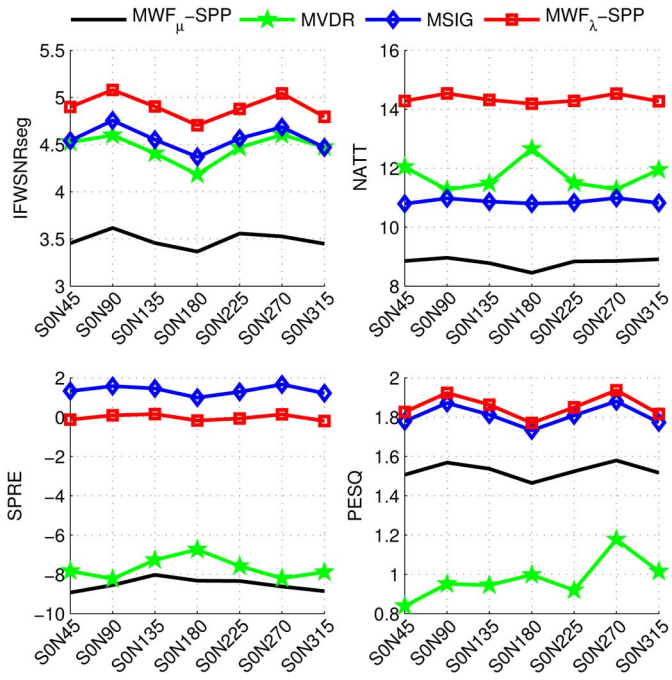

Fig. 5. Noise reduction performance for MWF at 0 dB SNR.

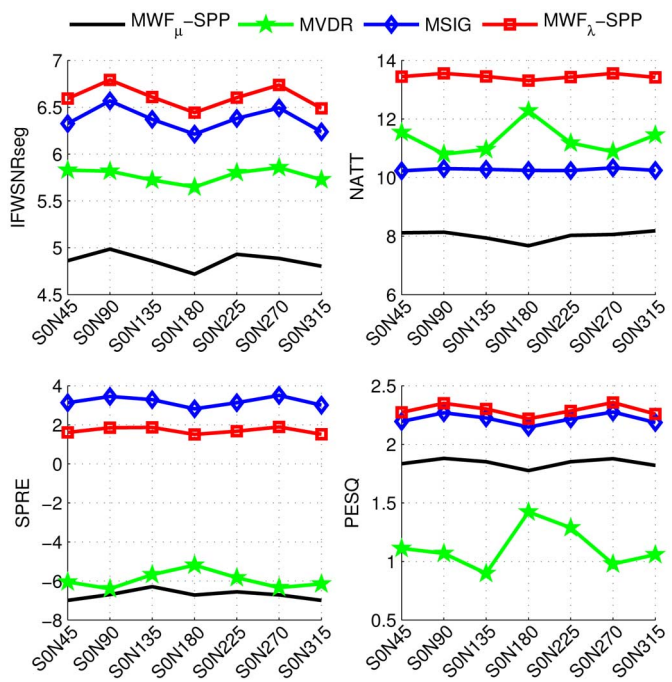

Fig. 6. Noise reduction performance for MWF at 5 dB SNR.

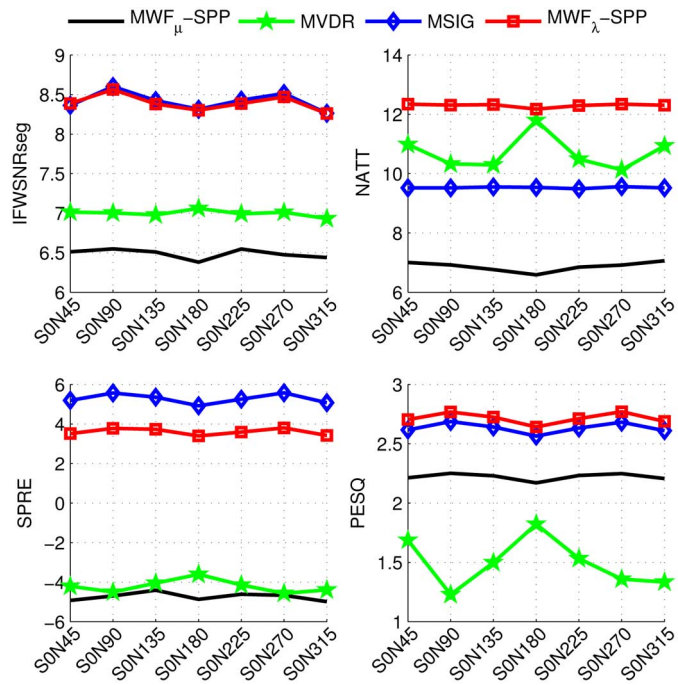

Fig. 7. Noise reduction performance for MWF at 10 dB SNR.

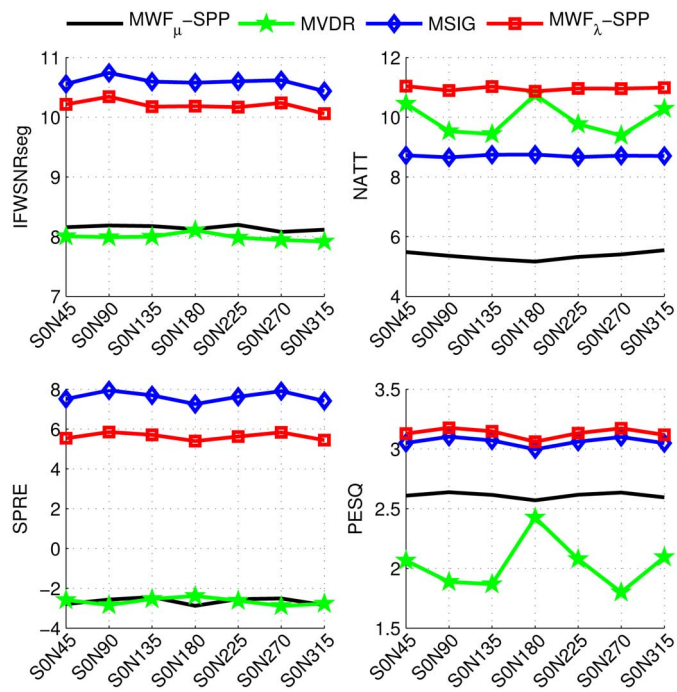

Fig. 8. Noise reduction performance for MWF at 15 dB SNR.

For noise reduction performance, scenarios with speech sources positioned at the head anterior and several point source noise configurations rendered at 45°, 90°, 135°, 180°, 225°, 270° and 315° with respect to the left of the head, were used. The results for different input SNRs will be plotted to show the robustness of the proposed methods ranging from extremely noisy to quiet environments. Note that for all of the performance evaluations, only the average scores obtained from the evaluated seven sentences will be shown rather than the measurement results from every speech sequence. Due to the limitation of the laboratory equipment, the experiment with longer reverberation time has not been conducted in this paper.

Figs. 3 and 4 portray the ILD and ITD errors for the evaluated algorithms averaged across 7 speech sequences and 4 different input SNRs (0 dB, 5 dB, 10 dB and 15 dB). Fig. 3 shows the results of the speech source coming from different angles, with the direction of noise fixed, while Fig. 4 depicts the results of the noise from different positions, with the direction of speech fixed. As predicted, all the evaluated methods successfully preserved speech cues at all configurations, except for the binaural MVDR approach. This is because the output speech signals produced by this filter has been greatly distorted, which can be seen in the results of the noise reduction performance.

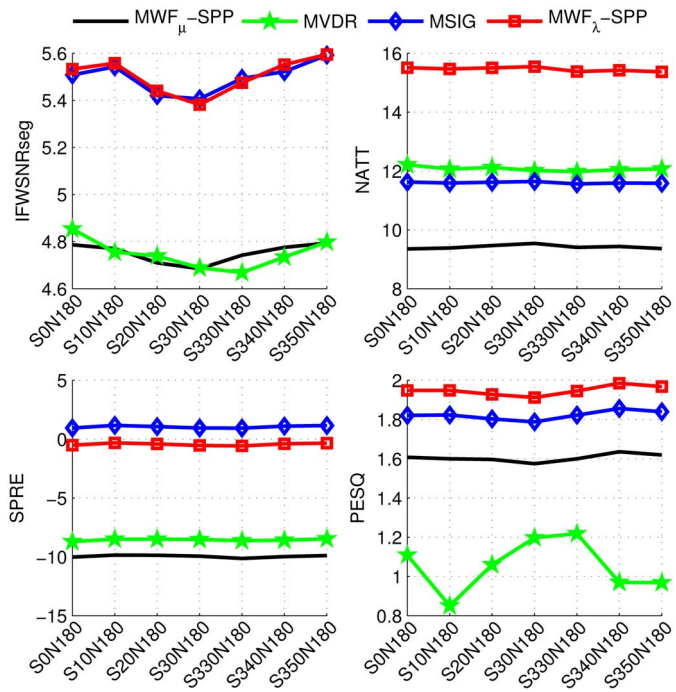

Fig. 9. Noise reduction performance for MWF in diffuse-like jack-hammer noise at 0 dB SNR.

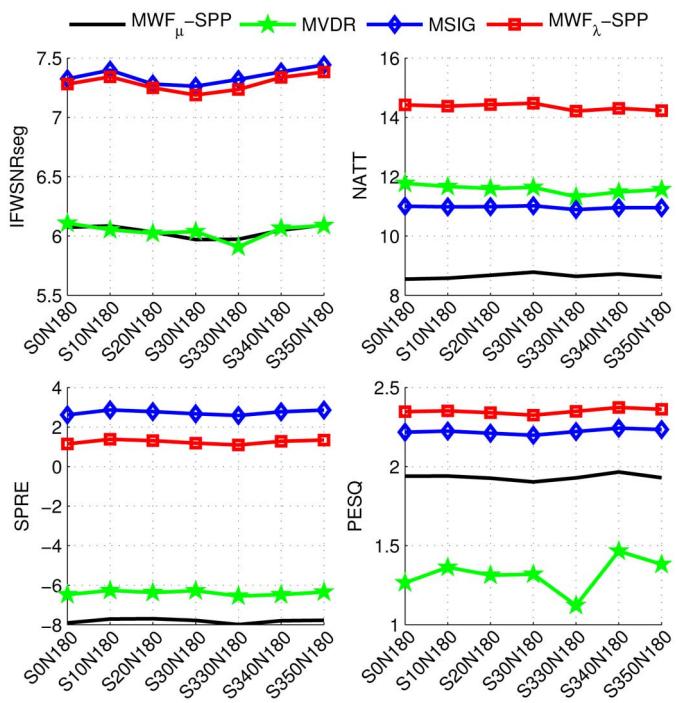

Fig. 10. Noise reduction performance for MWF in diffuse-like jack-hammer noise at 5 dB SNR.

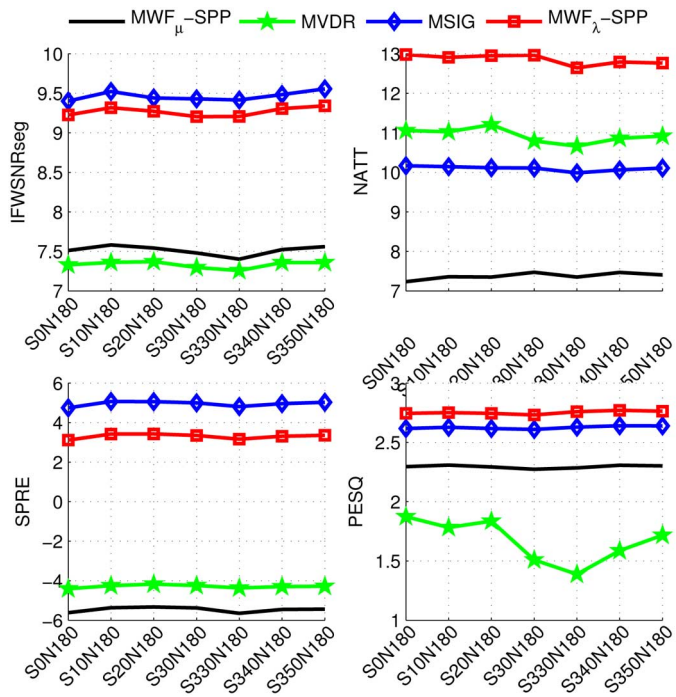

Fig. 11. Noise reduction performance for MWF in diffuse-like jack-hammer noise at 10 dB SNR.

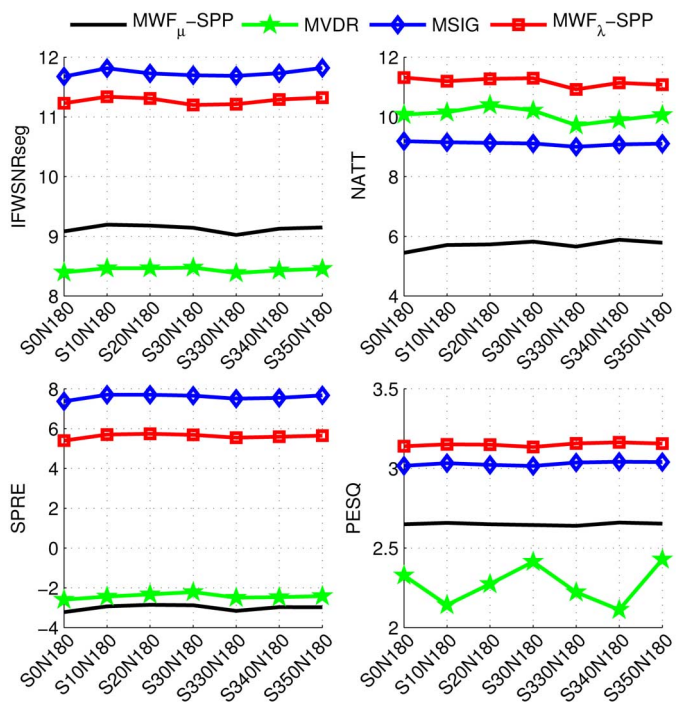

Fig. 12. Noise reduction performance for MWF in diffuse-like jack-hammer noise at 15 dB SNR.

For noise cues, an interesting finding is that ITDs of noise can be preserved with binaural MSIG, binaural MWF_λ-SPP, as well as binaural MWF_μ-SPP when noise source was fixed at 180° with speech source rendered from -30° to 30°. However, when speech was fixed in front of the head with noise coming from the side (45° to 135°), the ITDs of noise were not preserved by binaural MWF_μ-SPP. In general, the binaural MSIG has the least ILD and ITD errors when compared to both the proposed method and SDWMWF. The binaural MWF_λ-SPP has comparable performance with the binaural MSIG, but with slightly higher ILD errors. As for the binaural MWF_μ-SPP and the binaural MVDR, they recorded poor results among all four evaluated algorithms, particularly in the measured ITD errors. Thus, it has been shown that MWF_λ-SPP is a more preferable formulation compared to MWF_μ-SPP and MVDR for a binaural assistive listening device.

As for the noise reduction performance, Figs. 5 to 8 show the results are consistent with the results obtained from the monaural MWF formulations, as reported in [16]. It can be observed that the proposed binaural MWF_{λ}-SPP approach has the best performance recorded among all evaluated algorithms under lower input SNRs, which means having the largest IFWSNRseg, NATTseg and PESQ scores. This indicates that the binaural MWF_{λ}-SPP has better speech quality compared to the binaural MWF_{μ}-SPP and MSIG methods. However, it has consistently lower SPREseg when compared to the binaural MSIG function. This is mainly because the smoothing factors for the MDD approach were chosen as low as possible to reduce the amount of speech distortion (refer to Table I). As a result, for higher input SNRs, i.e., 15 dB in Fig. 8, where the impact of the speech distortion towards speech quality is larger, MWF_{λ}-SPP produced slightly lower IFWSNRseg results when compared to the binaural MSIG method. As for the performance of the binaural MWF_{μ}-SPP approach, it was totally outperformed by the proposed method in terms of both the SNR gains and the overall perceptual speech quality. For the binaural MVDR approach, it has the poorest performance among all the evaluated methods as it generates a large amount of speech distortion while having a high noise reduction, as can be observed in SPRE and NATT scores. It is also worth mentioning that when target speech is directed from the front, all algorithms except MVDR show a better performance when the noise is coming from the left or the right side of the head, and contrary, a poorer performance when noise is coming from behind due to the front-back ambiguity.

The performance of both frameworks were again examined in realistic scenarios with diffuse background noise. Evaluations were done using a diffuse-like jack-hammer noise, with the results depicted in Figs. 9 to 12. The results are similar when compared to the results from Fig. 5 to 8. The advantage of the binaural MWF_{λ}-SPP over the binaural MSIG method in terms of the SNR gains in IFWSNRseg results is less observable under diffuse-like noise conditions. However, in terms of the amount of noise reduction and overall speech quality, as depicted in NATT and PESQ results, the binaural MWF_{λ}-SPP method still performs better than the binaural MSIG. As for the other two reference methods, the binaural MVDR has much higher NATT scores when compared to the binaural MWF_{μ}-SPP. Since both methods have similar results for IFWSNRseg and SPRE, the higher noise reduction ratio in the binaural MVDR leads to undesired suppression of the speech components. This phenomenon can also be observed in PESQ scores.

VI. Conclusions

The use of the binaural multi-channel Wiener Filter (MWF) technique in hearing protection devices is challenging as it is unable to preserve the hearing impression of noise, which is critical for industrial workers in extremely noisy environments. There are more issues in regards to the implementation of MWF, particularly the estimation of the second order statistics, which often requires the aid of a voice activity detection (VAD) to detect speech presence and absence. Therefore, the prospective solution presented in this work incorporates the binaural MWF with the single-channel noise reduction approach. As such, the speech and noise components in the framework were proposed to be continuously estimated by utilizing a single-channel conditional speech presence probability (SPP) approach and a single-channel spectral weighing gain function. Experimental results show that the proposed MWF formulation performs better than the SDWMWF method in maintaining the spatial cues by having smaller ILD errors for both speech and noise. The proposed algorithm has also outperformed traditional methods in terms of speech quality, with a larger SNR improvement recorded without introducing a higher speech distortion.

References

[1] S. E. Nordholm and K. Fynn, "Apparatus and method for protecting hearing from noise while enhancing a sound signal of interest," U.S. 8,229,740, Jul. 24, 2012.

[2] V. Hamacher, "Comparison of advanced monaural and binaural noise reduction algorithms for hearing aids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, Orlando, FL, USA, May 2002, vol. 4, pp. IV–4008.

[3] A. 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.

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

[5] K. Reindl, Y. Zheng, and W. Kellermann, "Analysis of two generic wiener filtering concepts for binaural speech enhancement in hearing aids," in *Proc. 18th Eur. Signal Process. Conf. (EUSIPCO'10)*, Aalborg, Denmark, Feb. 2010, pp. 988–993.

[6] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Commun.*, vol. 39, no. 1, pp. 111–138, Jan. 2003.

[7] R. Aichner, H. Buchner, M. Zourub, and W. Kellermann, "Multi-channel source separation preserving spatial information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'07)*, Honolulu, HI, USA, Apr. 2007, vol. 1, pp. I–5.

[8] J. G. Desloge, W. M. Rabinowitz, and P. M. Zurek, "Microphone-array hearing aids with binaural output. I. Fixed-processing systems," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 529–542, Nov. 1997.

[9] D. P. Welker, J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[10] R. Nishimura, Y. Suzuki, and F. Asano, "A new adaptive binaural microphone array system using a weighted least squares algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, Orlando, FL, USA, May 2002, vol. 2, pp. II–1925.

[11] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, no. 7, pp. 636–656, Jul. 2007.

[12] A. Spriet, M. Moonen, and J. Wouters, "Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids," *IEEE Trans. Signal Process.*, vol. 53, no. 3, pp. 911–925, Mar. 2005.

[13] 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. 4, pp. 1579–1585, Apr. 2007.

[14] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC'06)*, Paris, France, Sep. 2006.

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

[16] P. C. Yong, S. Nordholm, H. H. Dam, Y. H. Leung, and C. C. Lai, "Incorporating multi-channel Wiener filter with single-channel speech enhancement algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'13)*, Vancouver, BC, Canada, May 2013, pp. 7284–7288.

[17] P. C. Yong, S. Nordholm, and H. H. Dam, "Noise estimation based on soft decisions and conditional smoothing for speech enhancement," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC'12)*, Aachen, Germany, Sep. 2012, pp. 4640–4643.
[18] P. C. Yong, S. Nordholm, and H. H. Dam, "Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement," *Speech Commun.*, vol. 55, no. 2, pp. 358–376, Feb. 2013.
[19] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter based noise reduction in hearing aids," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, p. 930625, Jul. 2009.
[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[21] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[22] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[23] 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.
[24] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.
[25] P. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL, USA: CRC, 2007.
[26] A. S. of America, "ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index," Jun. 1997.
[27] R. C. Hendriks and R. Martin, "MAP estimators for speech enhancement under normal and rayleigh inverse gaussian distributions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 918–927, Mar. 2007.
[28] T. J. Klasen, S. Doclo, T. den Bogaert, M. Moonen, and J. Wouters, "Binaural multi-channel wiener filtering for hearing aids: Preserving interaural time and level differences," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, Toulouse, France, May 2006, vol. 5, pp. V145–V148.
[29] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. New York, NY, USA: Springer, 2008, vol. 1.
[30] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

Pei Chee Yong (S'08) received the B.E (first class honors) in electronic and communication engineering, and the Ph.D. degree from Curtin University, Perth, Australia, in 2010 and 2014, respectively. He was awarded the Curtin International Postgraduate Research Scholarship for his Ph.D. study. He is currently a research fellow at the Department of Electrical and Computer Engineering, Curtin University. His research interests include speech enhancement, noise estimation, microphone arrays, and adaptive signal processing.

Sven Nordholm (M'90–SM'05) received his Ph.D. in signal processing in 1992 Licentiate of engineering in 1989 and MscEE (Civilingenjr) in 1983 all from Lund University, Sweden. He is Professor at Curtin University, Perth, Australia.

He was one of the founding members of the Department of Signal Processing, Blekinge Institute of Technology, BTH, Sweden in 1990 At BTH he held positions as Lecturer, Senior Lecturer, Associate Professor and Professor. Since 1999, he has been at Curtin University of Technology in Perth, Western Australia. From 1999-2002, he was director of ATRI and Professor at Curtin University of Technology. From 2002–2009, he was director Signal Processing Laboratory, WATRI, Western Australian Telecommunication Research Institute, a joint institute between The University of Western Australia and Curtin University. From 2009, he has been a Professor in Electrical and Computer Engineering in Curtin University. During 2012–2013, he served as Head of Department. He has also been Chief Scientist and co-founder of a start-up company Sensear, which provide voice communication in extreme noise conditions. He is also founder and director of a hearing aid company Hearmore. He is an associate editor *EURASIP Journal on Advances in Signal Processing* and *Journal of Franklin Institute*. He is a Senior Member of IEEE and a member of IEEE TC AASP.

His main research efforts have been spent in the fields of speech enhancement, adaptive and optimum microphone arrays, acoustic echo cancellation, adaptive signal processing, sub-band adaptive filtering and filter design.

Hai Huyen Dam received the bachelor (with first-class honors) and Ph.D. degrees (with distinction) from Curtin University of Technology, Perth, Australia, in 1996 and 2001, respectively. From 1999 to 2000, she was at the Blekinge Institute of Technology, Karlskrona, Sweden, as a Visiting Researcher. From 2001–2005, she was a Research Fellow/Senior Research Fellow with the Western Australian Telecommunications Research Institute (WATRI), Curtin University of Technology, Australia. Since 2016, she is a Senior Lecturer with the Department of Mathematics and Statistics, Curtin University of Technology. Her research interests are adaptive array processing, speech enhancement, optimization methods, equalization, and filter design.