MWF-based speech dereverberation with a local microphone array and an external microphone

Randall Ali[1], Toon van Waterschoot[1,2] and Marc Moonen[1]
[1]KU Leuven, Dept. of Electrical Engineering (ESAT-STADIUS), Leuven, Belgium
[2]KU Leuven, Dept. of Electrical Engineering (ESAT-ETC), e-Media Research Lab, Leuven, Belgium
Email: {randall.ali, toon.vanwaterschoot, marc.moonen}@esat.kuleuven.be

***Abstract*—A method for estimating the relevant quantities in a multi-channel Wiener filter (MWF) for speech dereverberation is proposed for a microphone system consisting of a local microphone array (LMA) and a single external microphone (XM). Typically these MWF quantities can be estimated by considering pre-whitened correlation matrices with a dimension equal to the number of microphones in the system. By following another procedure involving a pre-whitening-transformation operation, it will be demonstrated that when a priori knowledge of the relative transfer function (RTF) vector pertaining to only the LMA is available and when the reverberant component of the signals received by the LMA is uncorrelated with that of the XM, the MWF quantities may be alternatively estimated from a 2×2 matrix. Simulations confirm that using such an estimate results in a similar performance to that obtained by using the higher-dimensional correlation matrix.**

***Index Terms*—Multichannel Wiener Filter, Speech Dereverberation, Microphone Array, External Microphone**

I. Introduction

Speech communication applications incorporating the use of multiple microphones, such as automatic speech recognition, assistive hearing, and hands-free telephony, are compromised in highly reverberant environments, as the excessive reverberation captured by the microphone signals results in a degradation of speech quality and intelligibility. Signal processing techniques for speech dereverberation are therefore necessary in order to restore the optimal functionality for such applications. Throughout this paper, a reverberation suppression approach [1] will be followed, where the reverberant component is modelled as an additive distortion.

In devices equipped with a local microphone array (LMA), a multi-channel Wiener filter (MWF) can be used to suppress this reverberant component, provided that there are estimates of the relevant quantities, namely the speech and reverberant power spectral densities (PSDs), and the relative transfer function (RTF) vector pertaining to all of the microphones [2]–[5]. Recently, microphone systems consisting of an LMA and a single external microphone (XM) have also been considered, (such as a hearing aid that has access to the microphone signal on a mobile device) but for tasks of noise reduction and binaural cue preservation [6]–[9]. This paper therefore investigates how such an additional XM to the LMA can be exploited for estimating the relevant MWF quantities for speech dereverberation.

It should firstly be understood that with an LMA and an XM, the RTF vector required for the MWF would now consist of an RTF vector for the LMA and an additional RTF component for the XM. While a priori knowledge of an RTF vector can be imposed for the LMA as in blocking-based methods for speech dereverberation [3], [4], [10], a priori knowledge of the RTF component for the XM cannot be imposed as its relative position to the LMA is typically unknown. Therefore an estimate is required for this RTF component in order to complete the entire RTF vector for the MWF. Furthermore, as the XM may not always be close to the speech source, it should not be expected that listening to the XM signal alone would be a reliable option.

In [2], the MWF quantities were estimated by considering pre-whitened correlation matrices with a dimension equal to the number of microphones in the system. In the proposed approach, it is assumed that a priori knowledge of the RTF vector for the LMA is available, and that the XM is sufficiently far from the LMA [8], so that the reverberant component of the LMA signals is uncorrelated with the reverberant component of the XM signal. By following a procedure involving a pre-whitening-transformation operation, it is then shown how the relevant MWF quantities can be estimated from the eigenvalue decomposition (EVD) of a 2×2 matrix.

As will be demonstrated by simulations in a noise-free environment, using such an estimate results in a similar performance to that obtained by using the higher-dimensional correlation matrix. Additionally, it is observed that a microphone system consisting of an LMA and an XM is in general, more advantageous for speech dereverberation in comparison to using an LMA alone or an XM alone.

II. Data Model

A reverberant environment consisting of an LMA of M_{a} microphones, one additional XM, and one target speaker is considered as in Figure 1. In the short-time Fourier transform

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of IWT O&O Project nr. 150432 'Advances in Auditory Implants: Signal Processing and Clinical Aspects', KU Leuven C2-16-00449 'Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking', and KU Leuven Internal Funds VES/16/032. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. The scientific responsibility is assumed by its authors.

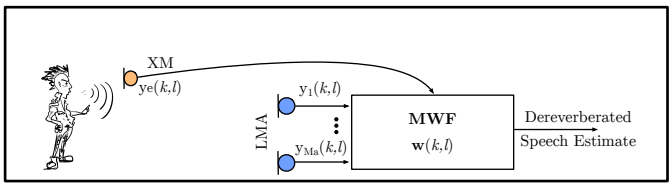

Fig. 1: Acoustic scenario consisting of a target speaker, an LMA, and an XM.

(STFT) domain, the stacked vector of microphone signals at frequency bin k and time frame l, are modelled as:

$$\mathbf{y}(k,l) = \underbrace{\mathbf{d}(k,l)\mathrm{s}_1(k,l)}_{\mathbf{x_d}(k,l)} + \mathbf{x_r}(k,l) \tag{1}$$

where $\mathbf{y}(k,l) = [\mathbf{y_a}^T(k,l)\ \mathrm{y_e}(k,l)]^T$, consists of the stacked LMA signals, $\mathbf{y_a}(k,l) = [\mathrm{y}_1(k,l)\ \mathrm{y}_2(k,l)\ \ldots \mathrm{y}_{M_a}(k,l)]^T$, and the XM signal, $\mathrm{y_e}(k,l)$. $\mathbf{x_d}(k,l)$ is the contribution from the direct path component of speech, represented by $\mathrm{s}_1(k,l)$, the speech signal in the first microphone of the LMA (i.e. the target signal of interest), filtered with the direct path RTF vector, $\mathbf{d}(k,l) = [\mathbf{d_a}^T(k,l)\ \mathrm{d_e}(k,l)]^T$, consisting of the direct path RTF vector for the LMA, $\mathbf{d_a}(k,l)$ (with the first microphone used as the reference, i.e. the first component of $\mathbf{d_a}(k,l)$ equal to 1), and the direct path RTF component for the XM, $\mathrm{d_e}(k,l)$. Finally, $\mathbf{x_r}(k,l)$ is the reverberant component. Throughout this paper, variables with the subscript "$\mathbf{a}$" refer to the LMA and those with the subscript "e" refer to the XM.

In the following, the early reflections of the reverberant component of the signals are deliberately excluded. However, such a model is not uncommon [1] and the proposed dereverberation procedure will be evaluated using signals that contain all direct, early, and reverberant components. Assuming that all frequency bins can be treated independently, only the dependence on time in the following derivations will be retained (where necessary) in order to simplify the notation.

With the consideration of a single speaker in a fixed position, and modelling the reverberant field as spatially homogeneous, the corresponding $(M_a+1)\times(M_a+1)$ correlation matrices for the microphone signals, $\mathbf{\Phi_y}(l)$, direct path component of speech, $\mathbf{\Phi_{x_d}}(l)$, and the reverberant component, $\mathbf{\Phi_{x_r}}(l)$, can be given respectively as:

$$\mathbf{\Phi_y}(l) = \mathbb{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\} \tag{2}$$

$$\mathbf{\Phi_{x_d}}(l) = \mathbb{E}\{\mathbf{x_d}(l)\mathbf{x_d}^H(l)\} = \Phi_\mathrm{s}(l)\mathbf{d}\mathbf{d}^H \tag{3}$$

$$\mathbf{\Phi_{x_r}}(l) = \mathbb{E}\{\mathbf{x_r}(l)\mathbf{x_r}^H(l)\} = \Phi_\mathrm{r}(l)\mathbf{\Gamma} \tag{4}$$

where $\mathbb{E}\{.\}$ is the expectation operator and $\{.\}^H$ is the Hermitian transpose. $\mathbf{\Phi_{x_d}}(l)$ is a rank-1 matrix, $\Phi_\mathrm{s}(l) = \mathbb{E}\{|\mathrm{s}_1(l)|^2\}$ is the time-varying power spectral density (PSD) of the target signal, $\Phi_\mathrm{r}(l)$ is the time-varying PSD of the reverberation, and $\mathbf{\Gamma}$ is a time-invariant spatial coherence matrix. $\mathbf{d}$ and particularly, $\mathrm{d_e}$ is assumed time-invariant, however, the position of the XM still remains unknown with respect to the LMA. Assuming that the direct path component of the speech is uncorrelated with the reverberant component, $\mathbf{\Phi_y}(l)$ can be expressed as:

$$\mathbf{\Phi_y}(l) = \mathbf{\Phi_{x_d}}(l) + \mathbf{\Phi_{x_r}}(l) = \Phi_\mathrm{s}(l)\mathbf{d}\mathbf{d}^H + \Phi_\mathrm{r}(l)\mathbf{\Gamma} \tag{5}$$

It will also be assumed that there is a perfect communication link between the LMA and XM with no bandwidth constraints and synchronous sampling, so that the signal correlations can be estimated as if all signals were available in a centralised processor. The estimate of the target signal, $\hat{\mathrm{s}}_1(l)$, i.e. the direct path component of the speech in the first microphone of the LMA, is then obtained through the linear filtering of the microphone signals, such that $\hat{\mathrm{s}}_1(l) = \mathbf{w}^H(l)\mathbf{y}(l)$, where $\mathbf{w}(l) = [\mathbf{w_a}^T(l)\ \mathrm{w_e}(l)]^T$. As discussed, an MWF will be used, which consists of a minimum variance distortionless response (MVDR) beamformer, followed by a single-channel post-filter:

$$\mathbf{w}(l) = \underbrace{\frac{\mathbf{\Gamma}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d}}}_{\text{MVDR}}\ \underbrace{\frac{\Phi_\mathrm{s}(l)}{\Phi_\mathrm{s}(l)+\Phi_\mathrm{r}(l)(\mathbf{d}^H\mathbf{\Gamma}^{-1}\mathbf{d})^{-1}}}_{\text{Single-Channel Post-Filter}} \tag{6}$$

Consequently, estimates are required for the quantities $\mathbf{d}$, $\mathbf{\Gamma}$, $\Phi_\mathrm{s}(l)$, and $\Phi_\mathrm{r}(l)$ in order to compute the MWF filter.

III. Estimation of the MWF quantities

This section summarises the state-of-the-art methods for estimating the MWF quantities, $\mathbf{d}$, $\mathbf{\Gamma}$, $\Phi_\mathrm{s}(l)$, and $\Phi_\mathrm{r}(l)$. As such methods have only considered an LMA, they are also extended to include an XM.

Firstly, $\mathbf{\Gamma}$ can be modelled as a spherically diffuse coherence matrix, so that each element, $\gamma_{p,q}$, in the matrix can be computed as $\gamma_{p,q} = \mathrm{sinc}\,(\omega\, r_{pq}/c)$ [11], where ω is the angular frequency (rad/s), c is the speed of sound (m/s), and r_{pq} is the distance (m) between the p-th and q-th microphone. Although the distance between the microphones in the LMA and the XM are unknown in practice, it can be assumed that the XM is far enough away from the LMA [8] so that the reverberant component of the XM signal is uncorrelated with the reverberant component of the LMA signal. An estimate for $\mathbf{\Gamma}$ in block matrix representation can then be given as:

$$\hat{\mathbf{\Gamma}} = \left[\begin{array}{c|c} \hat{\mathbf{\Gamma}}_\mathbf{a} & \mathbf{0}_{M_\mathrm{a}\times 1} \\ \hline \mathbf{0}_{1\times M_\mathrm{a}} & 1 \end{array}\right] \tag{7}$$

where $\hat{\mathbf{\Gamma}}_\mathbf{a}$ is the $(M_\mathrm{a}\times M_\mathrm{a})$ diffuse field coherence matrix for the LMA, whose elements can be computed as the inter-microphone distances in an LMA are typically known.

A spatial pre-whitening operation can then be defined by using the Cholesky decomposition:

$$\hat{\mathbf{\Gamma}} = \hat{\mathbf{\Gamma}}^{1/2}\,\hat{\mathbf{\Gamma}}^{H/2} \tag{8}$$

where $\hat{\mathbf{\Gamma}}^{1/2}$ is a lower triangular matrix. The MWF quantities can be estimated by using the pre-whitened correlation matrices in the optimization problem:

$$\min_{\Phi_\mathrm{r}(l),\Phi_\mathrm{s}(l),\,\mathbf{d}} ||\hat{\mathbf{\Gamma}}^{-1/2}(\hat{\mathbf{\Phi}}_\mathbf{y}(l) - \Phi_\mathrm{r}(l)\hat{\mathbf{\Gamma}} - \Phi_\mathrm{s}(l)\mathbf{d}\mathbf{d}^H)\hat{\mathbf{\Gamma}}^{-H/2}||_F^2 \tag{9}$$

where $||.||_F$ is the Frobenius norm and $\hat{\mathbf{\Phi}}_\mathbf{y}(l)$ is the estimate of $\mathbf{\Phi_y}(l)$ (for instance with recursive averaging [12]). Performing an eigenvalue decomposition (EVD) on the pre-whitened microphone signal PSD matrix, results in:

$$\hat{\mathbf{\Gamma}}^{-1/2}\hat{\mathbf{\Phi}}_\mathbf{y}(l)\hat{\mathbf{\Gamma}}^{-H/2} = \mathbf{U}\mathbf{\Lambda}(l)\mathbf{U}^H \tag{10}$$

where $\mathbf{U}$ is a unitary matrix of eigenvectors and $\mathbf{\Lambda}(l) = \text{diag}\{\lambda_1(l), \lambda_2(l), \dots \lambda_{M_\text{a}+1}(l)\}$ is a diagonal matrix of eigenvalues arranged in descending order. As the Frobenius norm is invariant under a unitary transformation [13], substituting (10) in (9) results in:

$$\min_{\Phi_\text{r}(l),\Phi_\text{s}(l),\,\mathbf{d}} ||\mathbf{\Lambda}(l) - \Phi_\text{r}(l)\mathbf{I}_{M_\text{a}+1} - \mathbf{U}^H \hat{\mathbf{\Gamma}}^{-1/2} \mathbf{\Phi}_{\mathbf{x_d}}(l) \hat{\mathbf{\Gamma}}^{-H/2} \mathbf{U}||_F^2 \tag{11}$$

where $\mathbf{I}_{M_\text{a}+1}$ is the $(M_\text{a}+1)\times(M_\text{a}+1)$ identity matrix (in general $\mathbf{I}_\vartheta$ will denote the $\vartheta \times \vartheta$ identity matrix). The solution to (11) can be interpreted as the best approximation of $\mathbf{\Lambda}(l)$ by means of a sum of a scaled identity matrix and a rank-1 $(M_\text{a}+1)\times(M_\text{a}+1)$ matrix. Firstly an estimate for $\mathbf{d}$ can be computed from the principal eigenvector of $\mathbf{U}$ [14] as:

$$\boxed{\hat{\mathbf{d}}^\text{pw} = \frac{1}{\rho}\hat{\mathbf{\Gamma}}^{1/2}\mathbf{U}\mathbf{e_1}} \tag{12}$$

where $\rho = \mathbf{e_1}^T \hat{\mathbf{\Gamma}}^{1/2}\mathbf{U}\mathbf{e_1}$, and the $(M_\text{a}+1)$ selection vector, $\mathbf{e_1} = [1\ 0 \dots 0]^T$. On replacing $\mathbf{d}$ with $\hat{\mathbf{d}}^\text{pw}$ in (11) then gives:

$$\min_{\Phi_\text{r}(l),\Phi_\text{s}(l)} ||\mathbf{\Lambda}(l) - \Phi_\text{r}(l)\mathbf{I}_{M_\text{a}+1} - \Phi_\text{s}(l)\mathbf{\Lambda_x}||_F^2 \tag{13}$$

where $\mathbf{\Lambda_x} = \text{diag}\{\frac{1}{|\rho|^2}, 0, \dots 0\}$. An estimate for $\Phi_\text{r}(l)$ follows by averaging the last M_a eigenvalues of $\mathbf{\Lambda}(l)$ [2]:

$$\boxed{\hat{\Phi}_\text{r}^\text{pw}(l) = \frac{1}{M_\text{a}}\big(\text{trace}\{\mathbf{\Lambda}(l)\} - \lambda_1(l)\big)} \tag{14}$$

Finally, on replacing $\Phi_\text{r}(l)$ with $\hat{\Phi}_\text{r}^\text{pw}(l)$ in (13) an estimate for $\Phi_\text{s}(l)$ can be computed as [12]:

$$\boxed{\hat{\Phi}_\text{s}^\text{pw}(l) = \big(\lambda_1(l) - \hat{\Phi}_\text{r}^\text{pw}(l)\big)\,|\rho|^2} \tag{15}$$

IV. Estimation of MWF quantities with a priori knowledge of the RTF vector for the LMA

With a priori knowledge of the direct path RTF vector for the LMA, the direct path speech component and the associated correlation matrix can be re-defined respectively as:

$$\widetilde{\mathbf{d}} = [\widetilde{\mathbf{d}}_\mathbf{a}^T \ \ \text{d}_\text{e}]^T; \quad \widetilde{\mathbf{\Phi}}_{\mathbf{x_d}}(l) = \Phi_\text{s}(l)\widetilde{\mathbf{d}}\widetilde{\mathbf{d}}^H \tag{16}$$

where the $\widetilde{\mathbf{d}}_\mathbf{a}$ is the known direct path RTF vector for the LMA. Therefore only an estimate is required for d_e as opposed to the entire RTF vector as in Section III.

Following from the approach outlined in [9], a transformation matrix can then be defined such that:

$$\mathbf{\Upsilon_1} = \left[\begin{array}{c|c} [\mathbf{C_a}\ \ \mathbf{f_a}] & \mathbf{0}_{M_\text{a}\times 1} \\ \hline \mathbf{0}_{1\times M_\text{a}} & 1 \end{array}\right] \tag{17}$$

where the $M_\text{a}\times(M_\text{a}-1)$ blocking matrix, $\mathbf{C_a}$, and $M_\text{a}\times 1$ fixed beamformer, $\mathbf{f_a}$, are defined such that:

$$\mathbf{C_a}^H\widetilde{\mathbf{d}}_\mathbf{a} = \mathbf{0}_{(M_\text{a}-1)\times 1}; \qquad \mathbf{f_a}^H\widetilde{\mathbf{d}}_\mathbf{a} = 1 \tag{18}$$

A transformed version of the microphone signals is therefore:

$$\mathbf{\Upsilon_1}^H\mathbf{y}(l) = \big[(\mathbf{C_a}^H\mathbf{y_a}(l))^T \ \ \mathbf{f_a}^H\mathbf{y_a}(l) \ \ \text{y}_\text{e}(l)\big]^T \tag{19}$$

consisting of the blocking matrix signals from the LMA, the fixed beamformer output signal, and the XM signal. A new spatial pre-whitening operator, $\mathbf{L}$, can then be defined from the transformed spatial coherence matrix:

$$\mathbf{\Upsilon_1}^H\hat{\mathbf{\Gamma}}\mathbf{\Upsilon_1} = \mathbf{L}\mathbf{L}^H \tag{20}$$

where $\mathbf{L}$ is lower triangular and can be factorised as $\mathbf{L} = \mathbf{\Upsilon_1}^H\hat{\mathbf{\Gamma}}^{1/2}\Theta$, for some unitary matrix, Θ. In fact, since the reverberant component of the XM signal is assumed to be uncorrelated with the reverberant component of the LMA signals, the last row $\mathbf{L}$ consists of only zeros except for a one in the last entry. After some rearranging [9], (9) can eventually be re-written as:

$$\min_{\Phi_\text{r}(l),\Phi_\text{s}(l),\,\text{d}_\text{e}} ||\mathbf{\Omega}^H\hat{\mathbf{\Phi}}_\mathbf{y}(l)\mathbf{\Omega} - \Phi_\text{r}(l)\mathbf{I}_{M_\text{a}+1} - \mathbf{\Omega}^H\Phi_\text{s}(l)\widetilde{\mathbf{d}}\widetilde{\mathbf{d}}^H\mathbf{\Omega}||_F^2 \tag{21}$$

where $\mathbf{\Omega}^H = \mathbf{L}^{-1}\mathbf{\Upsilon_1}^H$ is the pre-whitening-transformation operation. As a consequence of this operation, the last term in (21) is all zeros except for the bottom-right 2×2 block. Hence the EVD of the 2×2 matrix is considered:

$$\mathbf{J}^T\mathbf{\Omega}^H\hat{\mathbf{\Phi}}_\mathbf{y}(l)\mathbf{\Omega}\mathbf{J} = \underline{\underline{\mathbf{U}}}\,\underline{\underline{\mathbf{\Lambda}}}(l)\underline{\underline{\mathbf{U}}}^H \tag{22}$$

where $\mathbf{J} = [\ \mathbf{0}_{2\times(M_\text{a}-1)} \mid \mathbf{I}_2\ ]^T$, $\underline{\underline{\mathbf{U}}}$ is a 2×2 unitary matrix of eigenvectors and $\underline{\underline{\mathbf{\Lambda}}}(l) = \text{diag}\{\underline{\lambda}_1(l), \underline{\lambda}_2(l)\}$ is a diagonal matrix of eigenvalues arranged in descending order.

Applying a unitary transform to (21) with the block diagonal matrix, $\mathbf{G} = \text{blkdiag}\{\mathbf{I}_{M_\text{a}-1}, \underline{\underline{\mathbf{U}}}\}$ (blkdiag$\{.\}$ is an operator that creates a block diagonal matrix from its arguments), then results in (23), where only the bottom 2×2 block is diagonalised from the first term in (21) and $\mathbf{P_{11}}, \mathbf{P_{12}}$, and $\mathbf{P_{21}}$ are the residual matrices. The solution to (23) is now the best approximation of the first term in (23) by means of a sum of a scaled identity matrix and a rank-1 2×2 matrix, which is in contrast to the rank-1 $(M_\text{a}+1)\times(M_\text{a}+1)$ matrix required to solve (11).

From (22), an estimate for d_e then follows as the last element from:

$$\left[\widetilde{\mathbf{d}}_\mathbf{a}^T \ \ \hat{\text{d}}_\text{e}^\text{pwt}\right]^T = \frac{1}{\zeta}\,\mathbf{\Omega}^{-H}\mathbf{J}\,\underline{\underline{\mathbf{U}}}\,\underline{\underline{\mathbf{e}}}_\mathbf{1} \tag{24}$$

where $\zeta = \mathbf{e_1}^T\mathbf{\Omega}^{-H}\mathbf{J}\,\underline{\underline{\mathbf{U}}}\,\underline{\underline{\mathbf{e}}}_\mathbf{1}$, and the 2-element selection vector, $\underline{\underline{\mathbf{e}}}_\mathbf{1} = [1,\ 0]^T$. Substitution of (24) for $\widetilde{\mathbf{d}}$ in (23) eventually

$$\min_{\Phi_\text{r}(l),\Phi_\text{s}(l),\,\text{d}_\text{e}} ||\left[\begin{array}{c|c} \mathbf{P_{11}} & \mathbf{P_{12}} \\ \hline \mathbf{P_{21}} & \underline{\underline{\mathbf{\Lambda}}}(l) \end{array}\right] - \Phi_\text{r}(l)\left[\begin{array}{c|c} \mathbf{I}_{M_\text{a}-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_2 \end{array}\right] - \mathbf{G}^H\mathbf{\Omega}^H\Phi_\text{s}(l)\widetilde{\mathbf{d}}\widetilde{\mathbf{d}}^H\mathbf{\Omega}\mathbf{G}||_F^2 \tag{23}$$

$$\min_{\Phi_\text{r}(l),\Phi_\text{s}(l)} ||\left[\begin{array}{c|c} \mathbf{P_{11}} & \mathbf{P_{12}} \\ \hline \mathbf{P_{21}} & \underline{\underline{\mathbf{\Lambda}}}(l) \end{array}\right] - \Phi_\text{r}(l)\left[\begin{array}{c|c} \mathbf{I}_{M_\text{a}-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_2 \end{array}\right] - \Phi_\text{s}(l)\left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \underline{\underline{\mathbf{\Lambda}}}_\mathbf{x} \end{array}\right] ||_F^2 \tag{25}$$

results in (25), where $\underline{\underline{\mathbf{\Lambda}}}_{\mathbf{x}} = \text{diag}\{\frac{1}{|\zeta|^2}, 0\}$. Similar to (13), it is once again the diagonal elements which contribute to the solution of (25). Estimates for $\Phi_{\text{r}}(l)$ and $\Phi_{\text{s}}(l)$ then follow similarly to those in section III:

$$\boxed{\hat{\Phi}_{\text{r}}^{\text{pwt}}(l) = \frac{1}{M_{\text{a}}}\left(\text{trace}\{\mathbf{P}(l)\} - \underline{\underline{\lambda}}_1(l)\right)} \tag{26}$$

$$\boxed{\hat{\Phi}_{\text{s}}^{\text{pwt}}(l) = (\underline{\underline{\lambda}}_1(l) - \hat{\Phi}_{\text{r}}^{\text{pwt}}(l))\,|\zeta|^2} \tag{27}$$

where $\mathbf{P}(l) = \mathbf{G}^H\mathbf{\Omega}^H\hat{\mathbf{\Phi}}_{\mathbf{y}}(l)\mathbf{\Omega}\mathbf{G}$, i.e., the first term of (25).

Alternative estimates for $\Phi_{\text{r}}(l)$ and $\Phi_{\text{s}}(l)$ may also be considered by approximating (25) with its lower 2×2 blocks only, i.e. by solving the following problem:

$$\min_{\Phi_{\text{r}}(l),\Phi_{\text{s}}(l)} ||\underline{\underline{\mathbf{\Lambda}}}(l) - \Phi_{\text{r}}(l)\mathbf{I}_2 - \Phi_{\text{s}}(l)\underline{\underline{\mathbf{\Lambda}}}_{\mathbf{x}}||_F^2 \tag{28}$$

Estimates for $\Phi_{\text{r}}(l)$ and $\Phi_{\text{s}}(l)$ would then follow as:

$$\boxed{\hat{\Phi}_{\text{r}}^{\text{pwt,22}}(l) = \text{trace}\{\underline{\underline{\mathbf{\Lambda}}}(l)\} - \underline{\underline{\lambda}}_1(l) = \underline{\underline{\lambda}}_2(l)} \tag{29}$$

$$\boxed{\hat{\Phi}_{\text{s}}^{\text{pwt,22}}(l) = (\underline{\underline{\lambda}}_1(l) - \hat{\Phi}_{\text{r}}^{\text{pwt,22}}(l))\,|\zeta|^2} \tag{30}$$

where the estimate for $\Phi_{\text{r}}(l)$ is not anymore an average of the diagonal elements of $\mathbf{P}(l)$. The advantage here is that it is not necessary to compute $\mathbf{P}(l)$, but then it is only an approximation to the original problem of (21).

In terms of complexity of this approach, an EVD is performed on a 2×2 matrix as opposed to a $(M_{\text{a}}+1)\times(M_{\text{a}}+1)$ matrix, but the pre-whitening-transformation operation, $\mathbf{\Omega}^H$ still remains to be computed. However, as $\mathbf{L}$, $\mathbf{\Upsilon_1}$, and $\hat{\mathbf{\Gamma}}$ are all known and are data-independent, $\mathbf{\Omega}^H$ can be pre-computed and multiplied with the microphone signal vector as a pre-processing stage. It is then the last two elements of this pre-processed vector which can be used to construct the 2×2 matrix on the left hand side of (22).

V. Simulations

The simulated acoustic scenario to be evaluated consisted of a linear LMA with five omnidirectional microphones separated by $8\,\text{cm}$, along with one XM, and an end-fire positioned speech source $2\,\text{m}$ from the LMA in a room of dimensions $5.1\,\text{m} \times 6.3\,\text{m} \times 2.5\,\text{m}$ with a reverberation time of $600\,\text{ms}$. The room impulse responses were obtained using the randomised image method [15] and implemented from [16]. The speech source consisted of five sentences from the hearing in noise test (HINT) database [17]. All simulations were performed using the Weighted Overlap and Add (WOLA) method [18], with a Discrete Fourier Transform (DFT) size of 512, 50% overlap, and sampling frequency of 16 kHz. $\hat{\mathbf{\Phi}}_{\mathbf{y}}$ was computed using recursive averaging with a time constant of $100\,\text{ms}$.

A far-field approximation was used to define $\widetilde{\mathbf{d}}_{\mathbf{a}}$, such that $\widetilde{\mathbf{d}}_{\mathbf{a}} = [1\; e^{-j\omega\tau_2(\theta)} \ldots e^{-j\omega\tau_{M_a}(\theta)}]^T$, where $\tau_m(\theta)$ is the relative time delay between the m^{th} microphone and the first microphone, and θ is the a priori assumed location of the source with respect to the LMA, with 0^o defined as the end-fire direction. Using this definition of $\widetilde{\mathbf{d}}_{\mathbf{a}}$, $\mathbf{C}_{\mathbf{a}}$, and $\mathbf{f}_{\mathbf{a}}$ were defined accordingly from (18).

TABLE I: MWF quantities used for the evaluated algorithms.

Algorithm	Signals used	RTF vector	$\Phi_{\text{r}}(l)$	$\Phi_{\text{s}}(l)$
XM	XM	-	-	-
LMA	LMA	$\widetilde{\mathbf{d}}_{\mathbf{a}}$	$\hat{\Phi}_{\text{r}}^{\text{pw}(1)}$	$\hat{\Phi}_{\text{s}}^{\text{pw}(1)}$
PW	LMA+XM	$\hat{\mathbf{d}}^{\text{pw}}$	$\hat{\Phi}_{\text{r}}^{\text{pw}}$	$\hat{\Phi}_{\text{s}}^{\text{pw}}$
PW-PR	LMA+XM	$[\widetilde{\mathbf{d}}_{\mathbf{a}}^T \;\; \hat{\text{d}}_{\text{e}}^{\text{pwt}}]^T$	$\hat{\Phi}_{\text{r}}^{\text{pw}}$	$\hat{\Phi}_{\text{s}}^{\text{pw}}$
PWT	LMA+XM	$[\widetilde{\mathbf{d}}_{\mathbf{a}}^T \;\; \hat{\text{d}}_{\text{e}}^{\text{pwt}}]^T$	$\hat{\Phi}_{\text{r}}^{\text{pwt}}$	$\hat{\Phi}_{\text{s}}^{\text{pwt}}$
PWT-22	LMA+XM	$[\widetilde{\mathbf{d}}_{\mathbf{a}}^T \;\; \hat{\text{d}}_{\text{e}}^{\text{pwt}}]^T$	$\hat{\Phi}_{\text{r}}^{\text{pwt,22}}$	$\hat{\Phi}_{\text{s}}^{\text{pwt,22}}$

[1] $\hat{\Phi}_{\text{r}}^{\text{pw}}$ and $\hat{\Phi}_{\text{s}}^{\text{pw}}$ for the LMA algorithm were modified accordingly using only the LMA signals (i.e. as per [2] and [12]).

Table I summarises the list of algorithms evaluated and the estimates used for the direct path RTF vector, $\Phi_{\text{s}}(l)$, and $\Phi_{\text{r}}(l)$ for the MWF filter. PW is the pre-whitened procedure from Section III, PW-PR is the PW but with the a priori RTF vector for the LMA and estimate of d_{e}, PWT uses the pre-whitening-transformation procedure involving the 2×2 matrices from Section IV, and PWT-22 is the approximation to PWT considering only the diagonalised matrices from (25). Processing with the LMA only and the unprocessed XM signal were included as benchmarks against which processing with both LMA signals and the XM signal together could be compared. Figure 2 illustrates the two scenarios evaluated: (a) a scenario with the XM close to the speech source, and (b) a scenario with the XM further away from the speech source.

Figure 3 displays the results of these scenarios, with the figures on the left-hand column (i.e, (a), (b), and (c)) corresponding to the scenario where the XM was closer to the speech source and the figures on the right-hand column (i.e, (d), (e), and (f)) corresponding to the scenario where the XM was further away from the speech source. The difference (Δ) in the metrics, STOI [19], Cepstral Distance (CD) [20], and unweigthed segmental SNR (SNRseg) (i.e fwSNRseg from [20] with a neutralised weighting) from the reference signal were used for evaluation. This reference signal was the direct component of the speech signal in the first microphone of the LMA. Higher values of Δ-STOI, and Δ-SNRseg indicate a benefit, whereas lower values for Δ-CD indicate a benefit.

On observation of the left-hand column of Fig. 3, it can be seen that the PW-PR, PWT and PWT-22 algorithms perform better than using the LMA algorithm, and all exhibit a similar performance. This suggests that the PWT and PWT-22 methods can indeed be appropriate for estimating the MWF quantities. The difference in performance between the PW and PW-PR is due to the fact that $\hat{\mathbf{d}}^{\text{pw}}$ for the PW from (12) would have contained both direct and early components, and hence resulted in an estimate different from the anechoic reference. Finally, while it may seem that the XM outperforms all other algorithms, it should be noted that the spatial cue would be different from that of an estimate of the source in the reference microphone of the LMA, which may not be desirable in some applications.

In the right-hand column of Fig. 3, it can be observed that the XM yields a poor performance, which also indicates that listening to an XM signal alone could yield unpredictable quality as its location is subject to change. It is also seen once again that the PW-PR, PWT and PWT-22 algorithms all exhibit

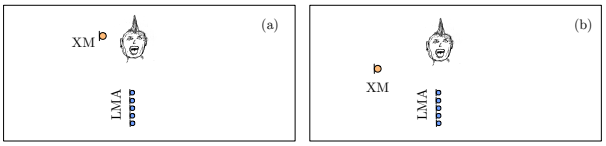

Fig. 2: Sketches of the simulated scenarios (a) XM at an angle of $15°$ and $1.7\,\mathrm{m}$ away from the LMA, (b) XM at an angle of $50°$ and $1.3\,\mathrm{m}$ away from the LMA. The LMA was positioned at $(1.9\,\mathrm{m},\ 3.6\,\mathrm{m},\ 1.4\,\mathrm{m})$.

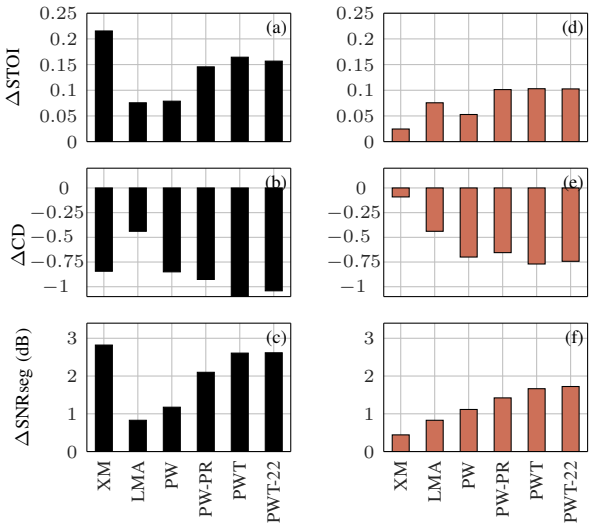

Fig. 3: Performance of the algorithms from table I: (a)-(c) when the XM is closer to the speech source (Fig.2 (a)); (d)-(f) when the XM is further from the speech source (Fig.2 (b)).

a similar performance and are preferable to the LMA algorithm or the XM alone. It is noted, however, that the absolute values of the metrics have decreased in comparison to when the XM was closer to the speech source. Nevertheless, this scenario also confirms that the PWT and PWT-22 methods would be appropriate for estimating the MWF quantities. Audio samples from these simulations may be heard at [21].

VI. Conclusions

A method has been proposed to estimate the relevant quantities in an MWF for speech dereverberation using a microphone system consisting of an LMA and an XM. With a priori knowledge of the RTF vector pertaining to only the LMA and when the reverberant component of the signals received by the LMA is uncorrelated with that of the XM, it was shown that by using a pre-whitening-transformation operation that these MWF quantities could be estimated from a 2×2 matrix. Simulations have also confirmed that using such an estimate results in a similar performance to what would obtained by using a higher-dimensional correlation matrix, and that using an LMA with an XM is generally advantageous for speech dereverberation in comparison to using an LMA alone or an XM alone.

References

[1] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. Wiley, Aug. 2018.

[2] I. Kodrasi and S. Doclo, "Analysis of Eigenvalue Decomposition-Based Late Reverberation Power Spectral Density Estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1102–1114, 2018.

[3] A. Kuklasiński, S. Doclo, and J. Jensen, "Maximum likelihood psd estimation for speech enhancement in reverberant and noisy conditions," in *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '16)*, March 2016, pp. 599–603.

[4] O. Schwartz, S. Gannot, and E. A. Habets, "Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments," *Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '16)*, vol. 2016-May, pp. 151–155, 2016.

[5] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and Comparison of Late Reverberation Power Spectral Density Estimators," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1052–1067, 2018.

[6] J. Szurley, A. Bertrand, B. van Dijk, and M. Moonen, "Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 5, pp. 952–966, 2016.

[7] D. Yee, H. Kamkar-Parsi, R. Martin, and H. Puder, "A Noise Reduction Post-Filter for Binaurally-linked Single-Microphone Hearing Aids Utilizing a Nearby External Microphone," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 5–18, 2017.

[8] N. Gößling and S. Doclo, "Relative Transfer Function Estimation Exploiting Spatially Separated Microphones in a Diffuse Noise Field," in *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC '18)*, Tokyo, Japan, Sept 2018.

[9] R.Ali, , G. Bernardi, T. van Waterschoot, and M. Moonen, "Methods of extending a generalized sidelobe canceller with external microphones," *IEEE/ACM Trans. Audio Speech Lang. Process.*, to appear, 2019.

[10] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st European Signal Process. Conf. (EUSIPCO '13)*, Sept 2013.

[11] E. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint." *J. Acoust. Soc. Amer.*, vol. 124, no. November, pp. 2911–2917, 2008.

[12] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction," in *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC '18)*, Tokyo, Japan, Sept 2018.

[13] I. Markovsky, *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer, 2012.

[14] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '15)*, Brisbane, Australia, April 2015, pp. 544–548.

[15] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 4, pp. 774–786, April 2015.

[16] N. Antonello. (2016) Room impulse response generator with the randomized image method. [Online]. Available: https://github.com/nantonel/RIM.jl/tree/master/src/MATLAB

[17] M. Nilsson, S. D. Soli, and J. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise." *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[18] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/Synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 99–102, 1980.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[20] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, p. 3387, 2009.

[21] R.Ali. (2018). [Online]. Available: ftp://ftp.esat.kuleuven.be/pub/SISTA/rali/Reports/EUSIPCO2019/AudioDRVB