COMPLETING THE RTF VECTOR FOR AN MVDR BEAMFORMER AS APPLIED TO A LOCAL MICROPHONE ARRAY AND AN EXTERNAL MICROPHONE

Randall Ali[†], *Toon van Waterschoot*[†*] *and Marc Moonen*[†]

[†]KU Leuven, Dept. of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium

[*]KU Leuven, Dept. of Electrical Engineering (ESAT-ETC), e-Media Research Lab, Andreas Vesaliusstraat 13, 3000 Leuven, Belgium

ABSTRACT

A minimum variance distortionless response (MVDR) beamformer can be an effective multi-microphone noise reduction strategy, provided that a vector of transfer functions from the desired speech signal at a reference microphone to the other microphones, i.e. a vector of the relative transfer functions (RTFs), is known. When using a local microphone array (LMA) and an external microphone (XM), this RTF vector has two distinct parts: an RTF vector for that of only the LMA and a single RTF component for the XM, with the reference microphone on the LMA. Whereas a priori assumptions can be made for the RTF vector for the LMA, the RTF for the XM must be estimated as the XM position is generally unknown. This paper investigates a procedure for estimating this unknown RTF by making use of the a priori RTF vector for the LMA, thereby completing the RTF vector for use of the MVDR beamformer. It is shown that such a procedure results in an Eigenvalue Decomposition (EVD) of a 2×2 matrix for a system of M microphones in the LMA and one XM. The resulting performance is evaluated within the context of a monaural MVDR beamformer.

Index Terms— Multi-Microphone Noise Reduction, Beamforming, MVDR, External Microphone, Relative Transfer Function.

1. INTRODUCTION

In hearing devices, such as hearing aids (HAs) and cochlear implants (CIs), the use of a multi-microphone noise reduction strategy is essential for preserving a desired speech signal and rejecting unwanted noise. Considerable attention has been devoted to this issue within the context of microphone arrays [1], but recently there has also been an interest in noise reduction strategies that include an external microphone (XM) [2–8]. In this paper, the minimum variance distortionless response (MVDR) beamformer [9] [10] as the multi-microphone noise reduction strategy is considered.

The MVDR beamformer can be effective provided that a vector of transfer functions from the desired speech signal at a reference microphone to the other microphones, i.e. a vector of the relative transfer functions (RTFs), is known. When using a local microphone array (LMA) and an external microphone (XM), this RTF vector has two distinct parts: an RTF vector corresponding to that of the LMA and a single RTF component for the XM, with the reference microphone on the LMA. A priori assumptions can be imposed on the RTF vector for the LMA due to the known, static relative microphone positions, whereas the position of the XM in relation to the LMA is typically unknown. Consequently the unknown RTF component for the XM must be estimated in order to complete the entire RTF vector for use in the MVDR beamformer.

Such an estimation can be done using the covariance subtraction or covariance whitening methods [11] as applied to correlation matrices involving both the LMA and XM signals. For instance, the procedure proposed in [6] uses the covariance whitening method to estimate the RTF component for the XM, which was consequently mixed with the a priori (anechoic) RTF vector for the LMA.

This paper investigates an alternative procedure whereby the a priori knowledge of the RTF vector for the LMA is explicitly used for estimating the RTF component for the XM. Such a procedure simply serves to augment an MVDR that has already been designed for use with the LMA, which could facilitate a practical implementation. Whether or not a pre-whitening operation is included, it is shown that this approach leads to an eigenvalue decomposition (EVD) of a 2×2 matrix for a system of M microphones in the LMA and one XM. The performance of the resulting MVDR beamformer using these estimates, as well as that from a previously developed method [8] in a monaural context is evaluated through simulations.

This paper is organised as follows. The data model is provided in Section 2. A review of the MVDR with a LMA and with an XM is given in Section 3. The RTF estimation methods are discussed in Section 4. Simulation results are presented in Section 5 and conclusions are drawn in Section 6.

2. DATA MODEL

A noise reduction system consisting of a LMA of M microphones plus one additional XM is considered. It is also assumed that there is only one desired speech signal in a noisy environment. In the short-time Fourier transform (STFT) domain, the received signal at one particular frequency, k, and one time frame, l, is represented as:

$$\mathbf{y}(k,l) = \underbrace{\mathbf{h}(k,l)\mathrm{s}_1(k,l)}_{\mathbf{x}(k,l)} + \mathbf{n}(k,l) \tag{1}$$

where (dropping the dependency on k and l for brevity) $\mathbf{y} = [\mathbf{y}_\mathbf{a}^T \; \mathrm{y}_\mathrm{e}]^T$, $\mathbf{y}_\mathbf{a} = [\mathrm{y}_1 \; \mathrm{y}_2 \; \ldots \mathrm{y}_\mathrm{M}]^T$ are the LMA signals, y_e is the XM signal, $\mathbf{x}$ is the speech contribution, represented by s_1, the speech signal in the first microphone of the LMA, filtered with $\mathbf{h} = [\mathbf{h}_\mathbf{a}^T \; \mathrm{h}_\mathrm{e}]^T$, $\mathbf{h}_\mathbf{a}$ is the RTF vector for the LMA (with the first microphone used as the reference, i.e. the first component of $\mathbf{h}_\mathbf{a}$ equal to 1), h_e is the RTF component for the XM. $\mathbf{n} = [\mathbf{n}_\mathbf{a}^T \; \mathrm{n}_\mathrm{e}]^T$ represents the noise contribution, which consists of correlated and uncorrelated noise. Variables with the subscript "$\mathbf{a}$" refer to the

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 Impulsfonds IMP/14/037, 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 scientific responsibility is assumed by its authors.

LMA and those with the subscript "e" refer to the XM.

The $(M+1)\times(M+1)$ speech-plus-noise, noise-only, and speech-only spatial correlation matrix are given respectively as:

$$\mathbf{R_{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\}; \quad \mathbf{R_{nn}} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\}; \quad \mathbf{R_{xx}} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} \tag{2}$$

where $\mathbb{E}\{.\}$ is the expectation operator and ^{H} is the Hermitian transpose. It is assumed that the speech signal is uncorrelated with the noise signal, and hence $\mathbf{R_{yy}} = \mathbf{R_{xx}} + \mathbf{R_{nn}}$. The speech-plus-noise and the noise-only spatial correlation matrix can also be calculated solely for the LMA signals respectively as $\mathbf{R_{y_a y_a}} = \mathbb{E}\{\mathbf{y_a}\mathbf{y_a}^H\}$ and $\mathbf{R_{n_a n_a}} = \mathbb{E}\{\mathbf{n_a}\mathbf{n_a}^H\}$. It is assumed that all signal correlations can be estimated as if all signals were available in a centralised processor, i.e., a perfect communication link is assumed between the LMA and XM with no bandwidth constraints and synchronous sampling.

The estimate of the speech component in the first microphone of the LMA, z_1, is then obtained through the linear filtering of the microphone signals, such that:

$$\mathrm{z}_1 = \mathbf{w}^H\mathbf{y} \tag{3}$$

where $\mathbf{w} = [\mathbf{w_a}^T \;\, \mathrm{w_e}]^T$ is the complex-valued filter to be designed.

3. MVDR BEAMFORMING

3.1. MVDR with an a priori RTF vector (MVDR-LM)

The MVDR as proposed in [9] [10] minimises the total noise power (minimum variance), while preserving the received signal in a particular direction (distortionless response). Considering only the LMA, the problem can be formulated as follows:

$$\begin{aligned} \min_{\mathbf{w_a}} \quad & \mathbf{w_a}^H \mathbf{R_{n_a n_a}} \mathbf{w_a} \\ \text{s.t.} \quad & \mathbf{w_a}^H \widetilde{\mathbf{h}}_\mathbf{a} = 1 \end{aligned} \tag{4}$$

where $\widetilde{\mathbf{h}}_\mathbf{a} = [\widetilde{\mathrm{h}}_{\mathrm{a},1} \;\, \widetilde{\mathrm{h}}_{\mathrm{a},2} \;\, \ldots \widetilde{\mathrm{h}}_{\mathrm{a,M}}]^T$ is the a priori RTF vector for the LMA that defines the direction for which the speech is to be preserved. $\widetilde{\mathbf{h}}_\mathbf{a}$ can be based on a priori assumptions regarding microphone characteristics, position, speaker location and room acoustics (e.g. no reverberation). For instance, it is not uncommon in hearing devices to assume knowledge of the speaker location [12–14]. The optimal noise reduction filter for (4) is then given by:

$$\mathbf{w_a} = \frac{\mathbf{R}_{\mathbf{n_a n_a}}^{-1}\widetilde{\mathbf{h}}_\mathbf{a}}{\widetilde{\mathbf{h}}_\mathbf{a}^H \mathbf{R}_{\mathbf{n_a n_a}}^{-1}\widetilde{\mathbf{h}}_\mathbf{a}} \tag{5}$$

which will be referred to as the MVDR-LM.

3.2. MVDR with an XM (MVDR-XM)

The MVDR-LM can be simply extended to incorporate the XM into what is referred to as the MVDR-XM:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^H \mathbf{R_{nn}} \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^H \widetilde{\mathbf{h}} = 1 \end{aligned} \tag{6}$$

where $\widetilde{\mathbf{h}} = [\widetilde{\mathbf{h}}_\mathbf{a}^T \;\, \hat{\mathrm{h}}_\mathrm{e}]^T$ consisting of $\widetilde{\mathbf{h}}_\mathbf{a}$, the a priori RTF vector for the LMA and $\hat{\mathrm{h}}_\mathrm{e}$ the RTF component for the XM to be estimated. Similarly to (4)-(5), the solution to (6) is:

$$\mathbf{w} = \frac{\mathbf{R}_\mathbf{nn}^{-1}\widetilde{\mathbf{h}}}{\widetilde{\mathbf{h}}^H \mathbf{R}_\mathbf{nn}^{-1}\widetilde{\mathbf{h}}} \tag{7}$$

With such a definition for $\widetilde{\mathbf{h}}$, only a single estimate for the RTF component for the XM, $\hat{\mathrm{h}}_\mathrm{e}$ is required (as opposed to estimating the entire RTF vector). In the following section, a previously developed method and the proposed method (with and without pre-whitening) for computing $\hat{\mathrm{h}}_\mathrm{e}$ will be discussed.

4. RTF ESTIMATION

4.1. Cross-Correlation Method

As previously proposed in [8], $\hat{\mathrm{h}}_\mathrm{e}$ can be found from a cross-correlation between an estimate of the speech signal in the first microphone of the LMA and the speech contribution in the XM. Using the estimate of the speech signal from the MVDR-LM, i.e. $\widetilde{\mathrm{z}}_{\mathrm{a},1} = \mathbf{w_a}^H\mathbf{y_a}$, a mean square error (MSE) problem can be formulated with the XM:

$$\min_{\hat{\mathrm{h}}_\mathrm{e}} \quad \mathbb{E}\{|\hat{\mathrm{h}}_\mathrm{e}\,\widetilde{\mathrm{z}}_{\mathrm{a},1} - \mathrm{y_e}|^2\} \tag{8}$$

The estimate for the RTF component for the XM is then (where * is the complex conjugate):

$$\hat{\mathrm{h}}_{\mathrm{e,xc}} = \frac{\mathbb{E}\{\mathrm{y_e}\,\widetilde{\mathrm{z}}_{\mathrm{a},1}^*\}}{\mathbb{E}\{\widetilde{\mathrm{z}}_{\mathrm{a},1}\,\widetilde{\mathrm{z}}_{\mathrm{a},1}^*\}} \tag{9}$$

4.2. EVD with a priori knowledge

In order to estimate an entire RTF vector, a method is proposed in [15] whereby, for a given $\mathbf{R_{yy}}$ and a given $\mathbf{R_{nn}}$, an improved speech-only correlation matrix, $\mathbf{R_{x,r1}}$ is computed, along with an improved noise-only correlation matrix, $\mathbf{R_{n,r1}}$ such that $\{\mathbf{R_{x,r1}}, \mathbf{R_{n,r1}}\}$ minimises the cost function:

$$J = \;\alpha||\mathbf{R_{yy}} - (\mathbf{R_{x,r1}} + \mathbf{R_{n,r1}})||_F^2 + (1-\alpha)||\mathbf{R_{nn}} - \mathbf{R_{n,r1}}||_F^2 \tag{10}$$

where $||.||_F$ is the Frobenius norm and $\alpha \in [0 \;\, 1]$ is a weighting parameter. In other words, $\mathbf{R_{x,r1}} + \mathbf{R_{n,r1}}$ should give an accurate approximation to $\mathbf{R_{yy}}$ and $\mathbf{R_{n,r1}}$ an accurate approximation to $\mathbf{R_{nn}}$, with α placing more weight on the respective approximation. Furthermore, a priori knowledge can be exploited here, such that $\mathbf{R_{x,r1}}$ should be low rank. Using a rank-1 model for $\mathbf{R_{x,r1}}$, it is shown in [15], that $\mathbf{R_{x,r1}}$ should minimise the following cost function:

$$J = \;\alpha(1-\alpha)||(\mathbf{R_{yy}} - \mathbf{R_{nn}}) - \mathbf{R_{x,r1}}||_F^2 \tag{11}$$

$\mathbf{R_{x,r1}}$ can then be found from an Eigenvalue Decomposition (EVD) of the matrix $(\mathbf{R_{yy}} - \mathbf{R_{nn}})$, where the entire RTF vector can be computed from the principal eigenvector.

However, for the case where the RTF vector for the LMA is known, such additional a priori knowledge can also be included on top of the rank-1 approximation for $\mathbf{R_{x,r1}}$. Consequently, $\mathbf{R_{x,r1}}$ can be expressed as:

$$\mathbf{R_{x,r1}} = \hat{\Phi}_{\mathrm{x,r1}}\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H = \hat{\Phi}_{\mathrm{x,r1}} \begin{bmatrix} \widetilde{\mathbf{h}}_\mathbf{a} \\ \hat{\mathrm{h}}_\mathrm{e} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{h}}_\mathbf{a}^H & \hat{\mathrm{h}}_\mathrm{e}^* \end{bmatrix} \tag{12}$$

where now, only $\hat{\Phi}_{\mathrm{x,r1}}$, the estimated speech power in the first microphone and $\hat{\mathrm{h}}_\mathrm{e}$ need to minimise the cost function of (11), i.e. the estimation problem is reduced to:

$$\min_{\hat{\Phi}_{\mathrm{x,r1}},\, \hat{\mathrm{h}}_{\mathrm{e}}} \quad ||(\mathbf{R_{yy}} - \mathbf{R_{nn}}) - \hat{\Phi}_{\mathrm{x,r1}} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathrm{h}}_{\mathrm{e}} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}}^{H} & \hat{\mathrm{h}}_{\mathrm{e}}^{*} \end{bmatrix} ||_F^2 \tag{13}$$

Proceeding to solve (13), an $M \times (M-1)$ unitary blocking matrix $\mathbf{B_a}$ and an $M \times 1$ vector $\mathbf{b_a}$ are defined such that:

$$\mathbf{B}_{\mathbf{a}}^{H} \widetilde{\mathbf{h}}_{\mathbf{a}} = \mathbf{0}; \qquad \mathbf{b_a} = \frac{\widetilde{\mathbf{h}}_{\mathbf{a}}}{||\widetilde{\mathbf{h}}_{\mathbf{a}}||} \tag{14}$$

where $\mathbf{B}_{\mathbf{a}}^{H}\mathbf{B_a} = \mathbf{I}_{(M-1)}$ and in general $\mathbf{I}_{\vartheta}$ is a $\vartheta \times \vartheta$ identity matrix. Using $\mathbf{B_a}$ and $\mathbf{b_a}$, an $(M+1) \times (M+1)$ unitary transformation matrix, $\mathbf{T}$, can be subsequently defined:

$$\mathbf{T} = \left[\begin{array}{c|c} \mathbf{T_a} & \mathbf{0} \\ \hline \mathbf{0} & 1 \end{array}\right] \tag{15}$$

where $\mathbf{T_a} = [\mathbf{B_a} \;\; \mathbf{b_a}]$, $\mathbf{T}_{\mathbf{a}}^{H}\mathbf{T_a} = \mathbf{I}_M$, and hence $\mathbf{T}^H\mathbf{T} = \mathbf{I}_{(M+1)}$. As the Frobenius norm is invariant under a unitary transformation [16], (13) can be rewritten as:

$$\min_{\hat{\Phi}_{\mathrm{x,r1}},\, \hat{\mathrm{h}}_{\mathrm{e}}} \quad ||\mathbf{T}^H((\mathbf{R_{yy}} - \mathbf{R_{nn}}) - \hat{\Phi}_{\mathrm{x,r1}} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathrm{h}}_{\mathrm{e}} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}}^{H} & \hat{\mathrm{h}}_{\mathrm{e}}^{*} \end{bmatrix})\,\mathbf{T}||_F^2 \tag{16}$$

which upon expansion results in:

$$\min_{\hat{\Phi}_{\mathrm{x,r1}},\, \hat{\mathrm{h}}_{\mathrm{e}}} \quad || \left[\begin{array}{c|c} \mathbf{K_{11}} & \mathbf{K_{12}} \\ \hline \mathbf{K_{21}} & \mathbf{K_{22}} \end{array}\right] - \left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{K_{x,r1}} \end{array}\right] ||_F^2 \tag{17}$$

where $\mathbf{K_{11}}$ is an $(M-1) \times (M-1)$ matrix, $\mathbf{K_{12}}$ an $(M-1) \times 2$ matrix, $\mathbf{K_{21}}$ a $2 \times (M-1)$ matrix and $\mathbf{K_{22}}$ and $\mathbf{K_{x,r1}}$ are 2×2 matrices realised as:

$$\mathbf{K_{22}} = \left[\begin{array}{ccc|c} & \mathbf{b}_{\mathbf{a}}^{H} & & 0 \\ \hline 0 & \dots & 0 & 1 \end{array}\right] (\mathbf{R_{yy}} - \mathbf{R_{nn}}) \left[\begin{array}{c|c} & 0 \\ \mathbf{b_a} & \vdots \\ & 0 \\ \hline 0 & 1 \end{array}\right] \tag{18}$$

$$\mathbf{K_{x,r1}} = \hat{\Phi}_{\mathrm{x,r1}} \begin{bmatrix} ||\widetilde{\mathbf{h}}_{\mathbf{a}}|| \\ \hat{\mathrm{h}}_{\mathrm{e}} \end{bmatrix} \begin{bmatrix} ||\widetilde{\mathbf{h}}_{\mathbf{a}}|| & \hat{\mathrm{h}}_{\mathrm{e}}^{*} \end{bmatrix} \tag{19}$$

From (17), it can be seen that the additional a priori knowledge of a known $\widetilde{\mathbf{h}}_{\mathbf{a}}$ reduces the estimation problem further to:

$$\min_{\hat{\Phi}_{\mathrm{x,r1}},\, \hat{\mathrm{h}}_{\mathrm{e}}} \quad ||\mathbf{K_{22}} - \mathbf{K_{x,r1}}||_F^2 \tag{20}$$

which is that of a rank-1 approximation of the 2×2 matrix, $\mathbf{K_{22}}$. The solution then follows by initially performing an EVD on $\mathbf{K_{22}}$ and extracting the principal eigenvector, $\mathbf{k_{max}} = [\mathrm{k_a} \;\; \mathrm{k_e}]^T$, corresponding to the largest eigenvalue. $\hat{\mathrm{h}}_{\mathrm{e}}$ is consequently calculated by the appropriate scaling and normalisation of the elements in $\mathbf{k_{max}}$ upon comparison with (19) and hence given by:

$$\hat{\mathrm{h}}_{\mathrm{e,evd}} = \frac{||\widetilde{\mathbf{h}}_{\mathbf{a}}||\; \mathrm{k_e}}{\mathrm{k_a}} \tag{21}$$

4.3. Covariance whitening with a priori knowledge

A natural extension to the EVD method previously described is that of covariance whitening (CW) [11], which involves a spatial pre-whitening operation followed by an EVD (subsequently referred to as EVD-CW). The spatial pre-whitening operation is defined from the noise-only correlation matrix using the Cholesky decomposition:

$$\mathbf{R_{nn}} = \mathbf{R}_{\mathbf{nn}}^{1/2}\; \mathbf{R}_{\mathbf{nn}}^{H/2} \tag{22}$$

where $\mathbf{R}_{\mathbf{nn}}^{1/2}$ is a lower triangular matrix, and $\mathbf{R}_{\mathbf{nn}}^{H/2}$ is its conjugate transpose. Spatial pre-whitening is then performed by multiplying the signal vector of interest by $\mathbf{R}_{\mathbf{nn}}^{-1/2}$. Therefore, the pre-whitened version of (13) becomes:

$$\min_{\hat{\Phi}_{\mathrm{x,r1}},\, \hat{\mathrm{h}}_{\mathrm{e}}} \; ||\mathbf{R}_{\mathbf{nn}}^{-1/2}((\mathbf{R_{yy}} - \mathbf{R_{nn}}) - \hat{\Phi}_{\mathrm{x,r1}} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathrm{h}}_{\mathrm{e}} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}}^{H} & \hat{\mathrm{h}}_{\mathrm{e}}^{*} \end{bmatrix})\mathbf{R}_{\mathbf{nn}}^{-H/2}||_F^2 \tag{23}$$

Representing the pre-whitened version of $\widetilde{\mathbf{h}}$ as:

$$\begin{bmatrix} \underline{\widetilde{\mathbf{h}}}_{\mathbf{a}} \\ \underline{\hat{\mathrm{h}}}_{\mathrm{e}} \end{bmatrix} = \mathbf{R}_{\mathbf{nn}}^{-1/2} \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathrm{h}}_{\mathrm{e}} \end{bmatrix} \tag{24}$$

pre-whitened versions of the unitary blocking matrix, $\underline{\mathbf{B}}_{\mathbf{a}}$, vector, $\underline{\mathbf{b}}_{\mathbf{a}}$, and transformation matrix, $\underline{\mathbf{T}}$ can all be defined such that:

$$\underline{\mathbf{B}}_{\mathbf{a}}^{H} \underline{\widetilde{\mathbf{h}}}_{\mathbf{a}} = \mathbf{0}; \qquad \underline{\mathbf{b}}_{\mathbf{a}} = \frac{\underline{\widetilde{\mathbf{h}}}_{\mathbf{a}}}{||\underline{\widetilde{\mathbf{h}}}_{\mathbf{a}}||}; \qquad \underline{\mathbf{T}} = \left[\begin{array}{c|c} \underline{\mathbf{T}}_{\mathbf{a}} & \mathbf{0} \\ \hline \mathbf{0} & 1 \end{array}\right] \tag{25}$$

where $\underline{\mathbf{T}}_{\mathbf{a}} = [\underline{\mathbf{B}}_{\mathbf{a}} \;\; \underline{\mathbf{b}}_{\mathbf{a}}]$, $\underline{\mathbf{T}}_{\mathbf{a}}^{H}\underline{\mathbf{T}}_{\mathbf{a}} = \mathbf{I}_M$, and hence $\underline{\mathbf{T}}^H\underline{\mathbf{T}} = \mathbf{I}_{(M+1)}$. The transformed version of (23) is then:

$$\min_{\hat{\Phi}_{\mathrm{x,r1}},\, \hat{\mathrm{h}}_{\mathrm{e}}} \quad ||\underline{\mathbf{T}}^H((\underline{\mathbf{R}}_{\mathbf{yy}} - \underline{\mathbf{R}}_{\mathbf{nn}}) - \hat{\Phi}_{\mathrm{x,r1}} \begin{bmatrix} \underline{\widetilde{\mathbf{h}}}_{\mathbf{a}} \\ \underline{\hat{\mathrm{h}}}_{\mathrm{e}} \end{bmatrix} \begin{bmatrix} \underline{\widetilde{\mathbf{h}}}_{\mathbf{a}}^{H} & \underline{\hat{\mathrm{h}}}_{\mathrm{e}}^{*} \end{bmatrix})\,\underline{\mathbf{T}}||_F^2 \tag{26}$$

where $\underline{\mathbf{R}}_{\mathbf{yy}} = \mathbf{R}_{\mathbf{nn}}^{-1/2}\mathbf{R_{yy}}\mathbf{R}_{\mathbf{nn}}^{-H/2}$ and $\underline{\mathbf{R}}_{\mathbf{nn}} = \mathbf{I}_{(M+1)}$, whose form is identical to that of (16), except that the pre-whitened quantities are used. Consequently, the estimation problem is reduced to:

$$\min_{\hat{\Phi}_{\mathrm{x,r1}},\, \hat{\mathrm{h}}_{\mathrm{e}}} \quad ||\underline{\mathbf{K}}_{\mathbf{22}} - \underline{\mathbf{K}}_{\mathbf{x,r1}}||_F^2 \tag{27}$$

where $\underline{\mathbf{K}}_{\mathbf{22}}$ and $\underline{\mathbf{K}}_{\mathbf{x,r1}}$ are 2×2 matrices realised as in (18) and (19) respectively, but replaced with the respective pre-whitened quantities. Once again, the solution follows from the rank-1 approximation of a 2×2 matrix: $\underline{\mathbf{K}}_{\mathbf{22}}$. Performing an EVD on $\underline{\mathbf{K}}_{\mathbf{22}}$ and extracting the principal eigenvector, $\underline{\mathbf{k}}_{\mathbf{max}} = [\underline{\mathrm{k}}_{\mathrm{a}} \;\; \underline{\mathrm{k}}_{\mathrm{e}}]^T$, corresponding to the largest eigenvalue, $\underline{\hat{\mathrm{h}}}_{\mathrm{e}}$ is initially calculated:

$$\underline{\hat{\mathrm{h}}}_{\mathrm{e}} = \frac{||\underline{\widetilde{\mathbf{h}}}_{\mathbf{a}}||\; \underline{\mathrm{k}}_{\mathrm{e}}}{\underline{\mathrm{k}}_{\mathrm{a}}} \tag{28}$$

following which the pre-whitening operation is undone to achieve the RTF estimate (where the $(M+1) \times 1$ selection vector, $\mathbf{e}_{\mathrm{e}} = [0 \dots 0 \; 1]^T$):

$$\hat{\mathrm{h}}_{\mathrm{e,evd-cw}} = \mathbf{e}_{\mathrm{e}}^{T}\, \mathbf{R}_{\mathbf{nn}}^{1/2} \begin{bmatrix} \underline{\widetilde{\mathbf{h}}}_{\mathbf{a}} \\ \underline{\hat{\mathrm{h}}}_{\mathrm{e}} \end{bmatrix} \tag{29}$$

5. SIMULATIONS

A LMA with two omnidirectional microphones separated by 1 cm, with an end-fire positioned speech source 1 m from the array, and an XM in a room of dimensions 6.9 m x 4.3 m x 2.6 m was considered. All simulations were performed using the Weighted Overlap and Add (WOLA) method [17], with a Discrete Fourier Transform (DFT) size of 512, 50% overlap, and sampling frequency of 16 kHz.

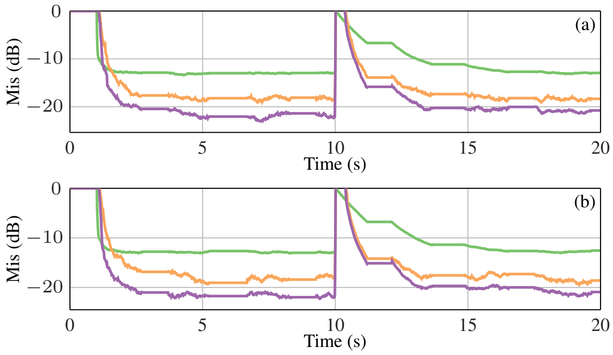

Fig. 1: (Colour online) Misalignment plots for (a) Real and (b) Imaginary parts for $\hat{\mathrm{h}}_{\mathrm{e,xc}}$ (——), $\hat{\mathrm{h}}_{\mathrm{e,evd}}$ (——), and $\hat{\mathrm{h}}_{\mathrm{e,evd-cw}}$ (——).

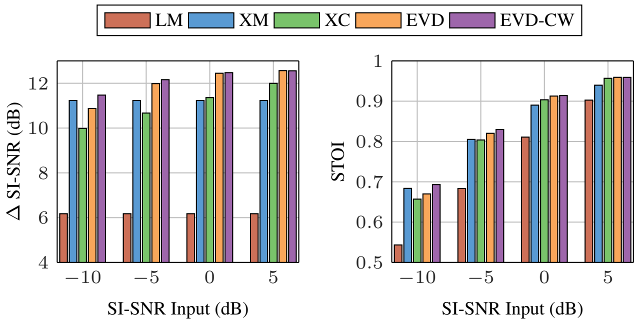

Fig. 2: Performance of the MVDR-LM, XM, and the MVDR-XM with the various RTF estimates in an anechoic scenario as a function of the SI-SNR at the first microphone of the LMA.

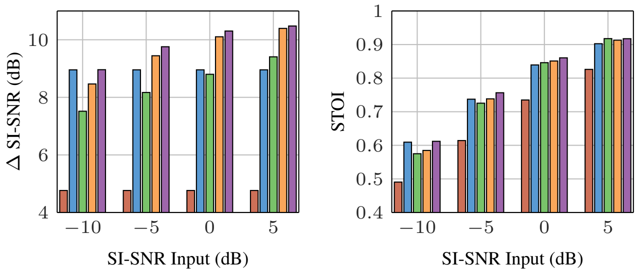

Fig. 3: Performance of the MVDR-LM, XM, and the MVDR-XM with the various RTF estimates with reverberation (T_{60} = 250ms) as a function of the SI-SNR at the first microphone of the LMA.

A perfect voice activity detector (VAD) was also used to retrieve the signals in the speech-plus-noise and noise-only frames. All RTF estimates were performed in periods where the speech source was active. The room impulse responses were obtained using the randomised image method [18] and implemented from [19].

In order to initially evaluate the relative performance of the RTF estimation methods discussed, an anechoic condition was considered, where white noise was used as the speech source signal, with an on-off behaviour dictated by the VAD. The noise field was a white diffuse noise field generated according to the method in [20]. The XM was initially placed 35 cm away from the speech source and instantaneously moved closer to only 9 cm away from the speech source after 10 s. The relevant correlation matrices were estimated with an exponential forgetting factor [21], corresponding to an averaging time of 1 s. The misalignment between the true RTF for the XM, $\mathrm{h_e}$, and the respective estimate, $\hat{\mathrm{h}}_{\mathrm{e}}$, was then calculated in each time frame up to the K^{th} frequency bin corresponding to 7125 Hz (for the real and imaginary parts accordingly) as:

$$\text{Mis (dB)} = 10\log_{10}\frac{\sum_{k=1}^{K}|\mathrm{h_e}(k)-\hat{\mathrm{h}}_{\mathrm{e}}(k)|^2}{\sum_{k=1}^{K}|\mathrm{h_e}(k)|^2} \tag{30}$$

Figure 1 displays the convergence of this misalignment for the three methods. All methods are able to adapt to changes in the position of the XM. It can also clearly be seen that the EVD-CW method performs better than the EVD method without pre-whitening, which in turn performs better than the cross-correlation method.

In a more realistic scenario, seven sentences separated by silence from the hearing in noise test (HINT) database [22] were used for the speech source signal. A diffuse noise field was generated from [20] using multitalker babble noise from Audiotec [23]. A scenario was considered for the XM, where it was placed just 26 cm away from the speech source and an averaging time of 3s was used in the estimation of the correlation matrices. The input signal-to-noise ratio (SNR) at the first microphone of the array was varied and the performance of the MVDR-XM using all the RTF estimation procedures was evaluated in terms of a change in speech-intelligibility-weighted signal-to-noise ratio (Δ SI-SNR) [24] in relation to the SI-SNR at the first microphone of the LMA and the short-time objective intelligibility (STOI) measure [25].

Figure 2 and 3 display the results of the MVDR-XM for the three RTF estimation methods, along with the MVDR-LM and the XM signal itself for an anechoic scenario and a scenario with a reverberation time of 0.25s respectively. Firstly, it can be seen that using an MVDR-XM with any of the estimation procedures offers an improvement over using an MVDR-LM. With respect to the Δ SI-SNR, in both anechoic and reverberant scenarios, the trend of an increasing performance from the cross-correlation method, to the EVD method without pre-whitening and then the EVD-CW method is observed, which corroborates with the result of Figure 1. A similar trend is observed for the STOI metric, although the differences are not as pronounced. For this particular position of the XM, it is also interesting to note that at lower input SI-SNRs, the performance of the EVD-CW method is better than or at least equivalent to the performance gained by simply switching to the use of the XM. However, it should be noted that switching to the XM will result in a loss of the spatial cues for the speech source. This suggests that future work should observe the effect of the XM position on the performance of the algorithms. Audio samples for an SI-SNR input of 0 dB can be heard at [26].

6. CONCLUSIONS

A procedure for estimating the unknown RTF component for an XM using the a priori information of the RTF vector for an LMA has been developed, thereby completing the entire RTF vector for an MVDR beamformer as applied to a LMA and an XM. It has been demonstrated that this procedure reduces to an EVD of a 2 × 2 matrix for a system of M microphones in the LMA and one XM. Simulation results have also indicated that the method with a pre-whitening operation would exhibit an improved performance over that without the pre-whitening operation and a cross-correlation method previously developed, within the context of a monaural MVDR beamformer.

7. REFERENCES

[1] M. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing, Techniques and Applications*. New York: Springer, 2001.

[2] A. Bertrand and M. Moonen, “Robust distributed noise reduction in hearing aids with external acoustic sensor nodes,” *Eurasip Journal on Advances in Signal Processing*, vol. 2009, 2009.

[3] N. Cvijanovic, O. Sadiq, and S. Srinivasan, “Speech enhancement using a remote wireless microphone,” *IEEE Trans. on Consumer Electronics*, vol. 59, no. 1, pp. 167–174, February 2013.

[4] 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.

[5] 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.

[6] N. Gößling, D. Marquardt, and S. Doclo, “Comparison of RTF Estimation Methods between a Head-Mounted Binaural Hearing Device and an External Microphone,” in *Proc. International Workshop on Challenges in Hearing Assistive Technology (CHAT)*, Stockholm, Sweden, August 2017, pp. 101–106.

[7] R. Ali, T. van Waterschoot, and M. Moonen, “A noise reduction strategy for hearing devices using an external microphone,” 2017, ESAT-STADIUS Technical Report TR 17-37, KU Leuven, Belgium.

[8] R.Ali, T. van Waterschoot, and M. Moonen, “Generalised sidelobe canceller for noise reduction in hearing devices using an external microphone,” in *Proc. 2018 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP ’18)*, Calgary, AB, Canada, April 2018.

[9] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[10] E. Habets, J. Benesty, S. Gannot, and I. Cohen, *Speech Processing in Modern Communication: Challenges and Perspectives*. Berlin Heidelberg: Springer, 2010, ch. 9, pp. 225–254.

[11] 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.

[12] J. Greenberg and P. Zurek, “Evaluation of an adaptive beamforming method for hearing aids,” *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1662–1676, 1992.

[13] J. M. Kates and M. R. Weiss, “A comparison of hearing-aid array-processing techniques,” *J. Acoust. Soc. Amer.*, vol. 99, no. 5, pp. 3138–3148, 1996.

[14] A. Spriet, L. Van Deun, K. Eftaxiadis, J. Laneau, M. Moonen, B. van Dijk, A. van Wieringen, and J. Wouters, “Speech understanding in background noise with the two-microphone adaptive beamformer BEAM in the Nucleus Freedom Cochlear Implant System.” *Ear and hearing*, vol. 28, no. 1, pp. 62–72, 2007.

[15] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, “Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 4, pp. 785–799, 2014.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] S. Haykin, *Adaptive Filter Theory Fifth Edition*. Prentice Hall, 2013.

[22] 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.

[23] Auditec, “Auditory Tests (Revised), Compact Disc, Auditec, St. Louis,” St. Louis, 1997.

[24] A. Spriet, M. Moonen, and J. Wouters, “Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction,” *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[25] 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.

[26] R.Ali. (2018). [Online]. Available: ftp://ftp.esat.kuleuven.be/stadius/rali/Reports/IWAENC%202018/Audio%20Data