Methods of extending a generalised sidelobe canceller with external microphones

Randall Ali, Giuliano Bernardi, Toon van Waterschoot and Marc Moonen

***Abstract*—While substantial noise reduction and speech enhancement can be achieved with multiple microphones organised in an array, in some cases, such as when the microphone spacings are quite close, it can also be quite limited. This degradation can however be resolved by the introduction of one or more external microphones (XMs) into the same physical space as the local microphone array (LMA). In this paper, three methods of extending an LMA-based generalised sidelobe canceller (GSC-LMA) with multiple XMs are proposed in such a manner that the relative transfer function pertaining to the LMA is treated as a priori knowledge. Two of these methods involve a procedure for completing an extended blocking matrix, while the third uses the speech estimate from the GSC-LMA directly with an orthogonalised version of the XM signals to obtain an improved speech estimate via a rank-1 generalised eigenvalue decomposition (GEVD). All three methods were evaluated with recorded data from an office room and it was found that the third method could offer the most improvement. It was also shown that in using this method, the speech estimate from the GSC-LMA was not compromised and would be available to the listener if so desired, along with the improved speech estimate that uses both the LMA and XMs.**

***Index Terms*—Multi-Microphone Noise Reduction, Speech Enhancement, External Microphone, GSC, beamforming**

I. Introduction

By exploiting their spatial variation, microphones organised in an array [1] have been successfully used for noise reduction and speech enhancement in several applications, including, but not limited to assistive hearing, mobile communication, and teleconferencing. In some cases, however, particularly for closely spaced microphone arrays, such as those on a hearing aid (HA), the spatial characteristics among the microphones may not be sufficiently distinct and hence the amount of noise reduction that can be achieved is limited. By introducing one or more external microphones (XMs) (such as on a mobile device or a wireless microphone clipped onto a desired speaker) into the same physical space as a 'local' microphone array (LMA), the spatial diversity among the microphones becomes greater, resulting in the potential for an increase in the amount of achievable noise reduction [2].

This fact has led to considerable research within the field of wireless acoustic sensor networks (WASNs) [3], where individual microphones and/or microphone arrays are randomly arranged in a physical space. For instance, several distributed speech enhancement algorithms have been developed [2] [4]–[6], which confirm the advantages of such WASNs.

For a WASN specifically consisting of an LMA (such as on an HA) and a single XM, early frequency modulation (FM) systems [7] [8] have been used to simply transmit an XM signal to a HA user, while disabling the LMA. It was assumed that the XM was always close to the desired speaker and hence a cleaner signal could be achieved. It was however noted in [7] that some subjects expressed concerns of persistent noise in very noisy environments as well as the problem of spatially localising the desired speaker.

Recently, though, a number of more sophisticated strategies have been proposed for this type of WASN. In [9]–[11], variants of the Multi-Channel Wiener Filter (MWF) [12] have been used for preservation of binaural cues for HA users. In [13], the use of the XM as a noise reference for speech enhancement was analysed while taking into account, issues associated with the wireless transmission of the audio signal. For single microphone HAs, the procedure in [14] used the XM to design a post-filter in order to resolve a front-back ambiguity. A different approach altogether used an XM (typically worn on the desired speaker) to estimate the sound direction of arrival (DoA) and then applied the appropriate binaural cues onto the "clean" XM signal [15] [16].

In this paper, the Minimum Variance Distortionless Response (MVDR) beamformer [17] [18] and its practical implementation, the Generalised Sidelobe Canceller (GSC) [19] will be considered for noise reduction. An extension of the previously discussed WASN to one that contains a single LMA collaborating with one or multiple XMs will also be considered. It will not be assumed that the XM(s) will always be close to the desired speaker, but rather that it (they) can be in any position within the physical space. The MVDR beamformer and the GSC can be effective provided that a vector of transfer functions relating the desired speech signal at a reference microphone to the desired speech signal at the other microphones, i.e. a vector of the relative transfer functions (RTFs), is known. In [2], [4]–[6], [9]–[11], the approach has been to estimate such an RTF vector for all of the microphones, i.e. for both the LMA and the XMs.

An alternative approach is however considered in this

R. Ali, G. Bernardi, T. van Waterschoot, and M. Moonen are with KU Leuven, Dept. of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium (email: {randall.ali, giuliano.bernardi, marc.moonen}@esat.kuleuven.be). T. van Waterschoot is also with KU Leuven, Dept. of Electrical Engineering (ESAT-ETC), e-Media Research Lab, Andreas Vesaliusstraat 13, 3000 Leuven, Belgium (email: toon.vanwaterschoot@esat.kuleuven.be).

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 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.

work, where available a priori knowledge of the RTF vector pertaining to only the LMA is explicitly used. For instance, in some hearing assistive devices it is not uncommon to assume a frontal location for the desired speaker [14], [20], which can subsequently be used to compute an a priori RTF vector for the LMA. It has been shown that designing an LMA-based noise reduction system for a hearing assistive device with such an a priori RTF vector can then lead to a practical and robust approach for the assumed desired speaker location [21], [22]. In this context, therefore, it is only the missing part of the RTF vector corresponding to the XMs that needs to be estimated. The advantage of this approach is that the XMs can be incorporated in a modular fashion or as "add-ons" for an improved performance to the LMA-based noise reduction system. With such modularity, this approach has a built-in contingency option of reverting to the original performance of the noise reduction system with the LMAs in cases where estimation becomes challenging. This is in contrast to a system where the entire RTF vector is estimated, since in such cases if the estimation is poor, there are no alternative options or decisions which can be taken to yield an acceptable performance.

Within an MVDR beamformer framework, this type of RTF estimation that uses the a priori information of the RTF vector of an LMA has already been considered in [23] for the case of one XM. In this paper, these procedures are generalised for multiple XMs and also extended to a practical GSC framework. In particular, three methods will be discussed and evaluated experimentally, two of which involve a process of completing a blocking matrix similar to that of [24]. The third method, which will be proven to offer the most improvement, uses the speech estimate from an LMA-based GSC (GSC-LMA) directly with an orthogonalised version of the XM signals to obtain an improved speech estimate via a rank-1 generalised eigenvalue decomposition (GEVD). This approach indeed does not compromise an existing GSC-LMA as both speech estimates, i.e. that from only using the existing GSC-LMA, and that from using the LMA in co-operation with the XMs are independently available.

The paper is organised as follows. In section II, the data model is presented. In section III, a review of processing schemes using only an LMA for an MVDR beamformer and a GSC is provided. In section IV, the extension of an LMA-based MVDR beamformer to include multiple XMs is introduced. In section V, the method of completing the blocking matrix for an extension of the GSC-LMA for two different RTF estimation procedures involving multiple XMs is discussed. In section VI, an alternative approach to extending the GSC-LMA is proposed, which involves an orthogonalisation of the XM signals and a rank-1 GEVD procedure. In section VII, the three methods are evaluated with recorded data taken in a typical office scenario. A summary and general conclusions are finally drawn in section VIII.

II. Data Model

A noise reduction system consisting of an LMA of M_{a} microphones plus M_{e} XMs 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:

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

where (dropping the dependency on k and l for brevity) $\mathbf{y} = [\mathbf{y_a}^T\ \mathbf{y_e}^T]^T$, $\mathbf{y_a} = [\mathrm{y}_{\mathrm{a},1}\ \mathrm{y}_{\mathrm{a},2}\ \dots \mathrm{y}_{\mathrm{a},M_{\mathrm{a}}}]^T$ are the LMA signals, $\mathbf{y_e} = [\mathrm{y}_{\mathrm{e},1}\ \mathrm{y}_{\mathrm{e},2}\ \dots \mathrm{y}_{\mathrm{e},M_{\mathrm{e}}}]^T$ are the XM signals, $\mathbf{x}$ is the speech contribution, represented by $\mathrm{s}_{\mathrm{a},1}$, the speech signal in the first microphone of the LMA, filtered with $\mathbf{h} = [\mathbf{h_a}^T\ \mathbf{h_e}^T]^T$, $\mathbf{h_a}$ is the RTF vector for the LMA (with the first microphone used as the reference, i.e. the first component of $\mathbf{h_a}$ equal to 1), and $\mathbf{h_e}$ is the RTF vector for the XM signals. Finally, $\mathbf{n} = [\mathbf{n_a}^T\ \mathbf{n_e}^T]^T$ represents the noise contribution. Variables with the subscript "$\mathbf{a}$" refer to the LMA and variables with the subscript "$\mathbf{e}$" refer to the XMs.

The $(M_{\mathrm{a}} + M_{\mathrm{e}}) \times (M_{\mathrm{a}} + M_{\mathrm{e}})$ speech-plus-noise, noise-only, and speech-only spatial correlation matrices are given respectively as:

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

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 the XM signals with no bandwidth constraints and with synchronous sampling.

The estimate of the speech component in the first microphone of the LMA, $\mathrm{z}_{\mathrm{a},1}$, is then obtained through the linear filtering of the microphone signals, such that:

$$\mathrm{z}_{\mathrm{a},1} = \mathbf{w}^H\mathbf{y} \tag{3}$$

where $\mathbf{w} = [\mathbf{w_a}^T\ \mathbf{w_e}^T]^T$ is a complex-valued filter.

III. Processing with a local microphone array

A. LMA-based MVDR

The MVDR beamformer as proposed in [17] [18] 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 $\widetilde{\mathbf{h}}_{\mathbf{a}} = [\widetilde{\mathrm{h}}_{\mathrm{a},1}\ \widetilde{\mathrm{h}}_{\mathrm{a},2}\ \dots \widetilde{\mathrm{h}}_{\mathrm{a},M_{\mathrm{a}}}]^T$ is the a priori RTF vector for the LMA that defines the constraint 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). The optimal noise reduction filter corresponding to (4) is then given by:

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

which is referred to as the MVDR-LMA. The speech estimate, $\widetilde{\mathrm{z}}_{\mathrm{a},1}$, is then obtained through the linear filtering of the microphone signals with the complex-valued filter $\mathbf{w_a}$:

$$\widetilde{\mathrm{z}}_{\mathrm{a},1} = \widetilde{\mathbf{w}}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}} \tag{6}$$

B. LMA-based GSC

In the practical implementation of the MVDR-LMA as proposed by Griffiths and Jim [19], the constrained minimisation problem of (4) is converted into an unconstrained one. The resulting beamformer, an LMA-based GSC (referred to before as the GSC-LMA), is displayed in Fig. 1. The top branch provides a speech reference by satisfying the constraint in (4) through the use of a fixed beamformer, $\mathbf{f_a}$. The output of the top branch is then given by:

$$\mathrm{y}_{\mathrm{f}} = \mathbf{f}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}} \tag{7}$$

The bottom branch provides the noise reference signals $\mathbf{u_a} = \mathbf{C}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}} = [\mathrm{u}_{\mathrm{a},1}\ \mathrm{u}_{\mathrm{a},2}\ldots\mathrm{u}_{\mathrm{a},M_{\mathrm{a}}-1}]^{T}$ through the $M_{\mathrm{a}} \times (M_{\mathrm{a}}-1)$ blocking matrix, $\mathbf{C_a}$, which is defined as being orthogonal to the corresponding RTFs such that $\mathbf{C}_{\mathbf{a}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a}} = \mathbf{0}$. Therefore, $\mathbf{C_a}$ can be defined as follows:

$$\mathbf{C}_{\mathbf{a}} = \begin{bmatrix} -\widetilde{\mathrm{h}}_{\mathrm{a},2}^{*} & -\widetilde{\mathrm{h}}_{\mathrm{a},3}^{*} & \ldots & -\widetilde{\mathrm{h}}_{\mathrm{a},M_{\mathrm{a}}}^{*} \\ & \mathbf{I}_{M_{\mathrm{a}}-1} & & \end{bmatrix} \tag{8}$$

where $\{.\}^{*}$ denotes the complex conjugate and $\mathbf{I}_{M_{\mathrm{a}}-1}$ is the $(M_{\mathrm{a}}-1)\times(M_{\mathrm{a}}-1)$ identity matrix (in general $\mathbf{I}_{\vartheta}$ will denote the $\vartheta \times \vartheta$ identity matrix).

The adaptive noise cancelling (ANC) filter, $\mathbf{v_a}$, is then updated such as to reduce the residual noise in the speech reference at each time frame, l[1], by solving the following unconstrained optimisation problem:

$$\min_{\mathbf{v_a}(l)} \quad \mathbb{E}\{|\mathbf{f}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}}(l) - \mathbf{v}_{\mathbf{a}}^{H}(l)\mathbf{C}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}}(l)|^{2}\} \tag{9}$$

In order to avoid speech cancellation due to speech leakage into the noise reference, $\mathbf{v_a}$ is usually updated in frames where only noise is present. The optimal solution for $\mathbf{v_a}$ is given as:

$$\hat{\mathbf{v}}_{\mathbf{a}}(l) = (\mathbf{C}_{\mathbf{a}}^{H}\mathbf{R}_{\mathbf{n_a n_a}}(l)\mathbf{C_a})^{-1}\mathbf{C}_{\mathbf{a}}^{H}\mathbf{R}_{\mathbf{n_a n_a}}(l)\mathbf{f_a} \tag{10}$$

from which the filter output representing a speech estimate follows as:

$$\widetilde{\mathrm{e}}_{\mathrm{a},1}(l) = \mathbf{f}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}}(l) - \hat{\mathbf{v}}_{\mathbf{a}}^{H}(l)\mathbf{C}_{\mathbf{a}}^{H}\mathbf{y}_{\mathbf{a}}(l) \tag{11}$$

In practice, the solution to (9) is often implemented with a Normalised Least Mean Squares (NLMS) approach [25].

[1]The dependency on l will be re-introduced to highlight the importance of the time dependence on some quantities. These quantities are still per frequency and the dependency on k will continue to be omitted for brevity.

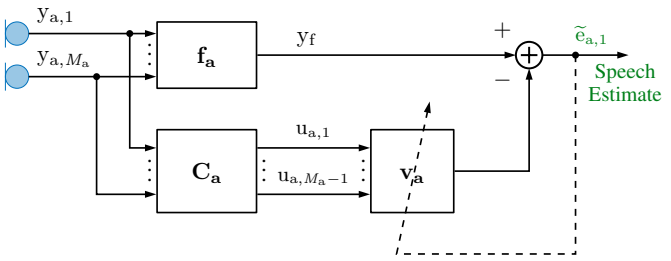

Fig. 1: LMA-based Generalised Sidelobe Canceller, GSC-LMA.

IV. MVDR Beamformer with a Local Microphone Array and Multiple External Microphones

The MVDR-LMA can be simply extended to include the XM signals into what is referred to here as the MVDR-LMA-XM:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^{H}\mathbf{R}_{\mathbf{nn}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{H}\mathbf{h} = 1 \end{aligned} \tag{12}$$

where $\mathbf{h}$ is the RTF vector that consists of M_{a} components corresponding to the LMA, $\mathbf{h_a}$, and M_{e} components corresponding to the XM signals, $\mathbf{h_e}$.

As the RTF vector is in general not known, its definition proves to be a challenging aspect in designing the MVDR beamformer. The MVDR-LMA as defined in section III-A imposes a priori assumptions on the RTF vector for the LMA. In the case of including one or several XMs, however, no such a priori assumptions can be made on the relative positions of the XMs in relation to the LMA as they are subject to change (consider using an XM on a mobile phone for instance). Consequently, there are two potential approaches that can be taken in order to define $\mathbf{h}$ - (i) only the missing section of the RTF vector corresponding to that of the XM signals is estimated, while the a priori assumed RTF vector for the LMA signals is preserved or (ii) the entire RTF vector is estimated for the LMA signals and the XM signals.

As discussed in section I, the first of these approaches is considered as it intends to preserve the reliability of an existing LMA-based system, while treating the XMs as "add-ons". Additionally, it will only be necessary to compute M_{e} estimates for the missing RTF section (as opposed to $M_{\mathrm{a}}+M_{\mathrm{e}}$ in an entire RTF vector estimation). Such an RTF vector will therefore be defined as follows:

$$\widetilde{\mathbf{h}} = [\widetilde{\mathbf{h}}_{\mathbf{a}}^{T} \mid \hat{\mathbf{h}}_{\mathbf{e}}^{T}]^{T} \tag{13}$$

where various methods for computing $\hat{\mathbf{h}}_{\mathbf{e}}$ (in the case of $M_{\mathrm{e}} = 1$) have been presented in [23]. It should also be noted that although $\widetilde{\mathbf{h}}$ partially contains an estimated RTF vector, this is done with respect to the a priori assumptions set by $\widetilde{\mathbf{h}}_{\mathbf{a}}$, and hence the notation for $\widetilde{\mathbf{h}}$ is kept to be that of an a priori RTF vector (i.e. $\{\widetilde{\ .\ }\}$).

Replacing $\mathbf{h}$ in (12) with the definition from (13), the

MVDR-LMA-XM is then given by (similar to (5)):

$$\widetilde{\mathbf{w}} = \frac{\mathbf{R}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}}}{\widetilde{\mathbf{h}}^H \mathbf{R}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}}} \tag{14}$$

with the speech estimate, $\widetilde{z}_1 = \widetilde{\mathbf{w}}^H \mathbf{y}$.

In the following sections, three methods are discussed for the implementation of the MVDR-LMA-XM in a GSC framework, referred to here as the GSC-LMA-XM.

V. Completing the Blocking Matrix

One approach for implementing a GSC structure with a LMA and multiple XMs is to use the estimate, $\hat{\mathbf{h}}_{\mathbf{e}}$ to complete the additional columns of the blocking matrix. In [24], this was demonstrated (for $M_{\mathrm{e}} = 1$) using a computation of $\hat{\mathbf{h}}_{\mathbf{e}}$ based on a cross-correlation method. In this section, this approach of completing the blocking matrix will be extended for $M_{\mathrm{e}} \geq 1$ with a further discussion of relevant implementation details. Two block schemes for a GSC will also be presented: (i) using the cross-correlation method to compute $\hat{\mathbf{h}}_{\mathbf{e}}$, and (ii) using the EVD method to compute $\hat{\mathbf{h}}_{\mathbf{e}}$ adopted from [23] (where $M_{\mathrm{e}} = 1$).

The cost function of (9) is firstly extended to include the XMs:

$$\min_{\mathbf{g}} \quad \mathbb{E}\{|\mathbf{f}^H(l)\mathbf{y}(l) - \mathbf{g}^H(l)\mathbf{C}^H(l)\mathbf{y}(l)|^2\} \tag{15}$$

where $\mathbf{f}$ is an $(M_{\mathrm{a}}+M_{\mathrm{e}})\times 1$ fixed beamformer acting on both the LMA and XM signals, $\mathbf{g} = [\mathbf{g}_{\mathbf{a}}^T\ \mathbf{g}_{\mathbf{e}}^T]^T$ is the ANC filter to be designed, and the extended $(M_{\mathrm{a}} + M_{\mathrm{e}}) \times (M_{\mathrm{a}} + M_{\mathrm{e}} - 1)$ blocking matrix is now given as:

$$\mathbf{C}(l) = \left[\begin{array}{c|c} \mathbf{C}_{\mathbf{a}} & \begin{matrix} -\hat{\mathbf{h}}_{\mathbf{e}}^H(l) \\ \mathbf{0}_{(M_{\mathrm{a}}-1)\times M_{\mathrm{e}}} \end{matrix} \\ \hline \mathbf{0}_{M_{\mathrm{e}}\times(M_{\mathrm{a}}-1)} & \mathbf{I}_{M_{\mathrm{e}}} \end{array}\right] \tag{16}$$

where $\mathbf{C}_{\mathbf{a}}$ is defined from (8), the zero blocks are indicated with their dimensions.

The role of the fixed beamformer within the context of a GSC is to satisfy the distortionless constraint, which can be accomplished regardless of the XMs. Consequently, the fixed beamformer, $\mathbf{f}$, can be readily simplified by setting $\mathbf{f} = [\mathbf{f}_{\mathbf{a}}^T\ \mathbf{0}_{(M_{\mathrm{e}}\times 1)}^T]^T$, i.e. using the fixed beamformer from (7) for the LMA signals and an $(M_{\mathrm{e}} \times 1)$ vector of zeros for the XM signals, hence $\mathbf{f}^H\widetilde{\mathbf{h}} = \mathbf{f}_{\mathbf{a}}^H\widetilde{\mathbf{h}}_{\mathbf{a}}$. As a result, only $\hat{\mathbf{h}}_{\mathbf{e}}(l)$ will be required to complete the blocking matrix, $\mathbf{C}(l)$, which requires an update for each time frame. The optimal solution for $\mathbf{g}(l)$ is also computed in noise-only periods in a similar manner to $\mathbf{v}_{\mathbf{a}}(l)$ for the GSC-LMA, and is given by:

$$\hat{\mathbf{g}}(l) = (\mathbf{C}^H(l)\mathbf{R}_{\mathbf{nn}}(l)\mathbf{C}(l))^{-1}\mathbf{C}^H(l)\mathbf{R}_{\mathbf{nn}}(l)\mathbf{f} \tag{17}$$

On substitution of (16) into (15), and with $\mathbf{f} = [\mathbf{f}_{\mathbf{a}}^T\ \mathbf{0}_{(M_{\mathrm{e}}\times 1)}^T]^T$, the new speech estimate then follows as:

$$\widetilde{\mathrm{e}}_1(l) = \underbrace{\mathbf{f}_{\mathbf{a}}^H\mathbf{y}_{\mathbf{a}}(l) - \hat{\mathbf{g}}_{\mathbf{a}}^H(l)\underbrace{\mathbf{C}_{\mathbf{a}}^H\mathbf{y}_{\mathbf{a}}(l)}_{\mathbf{u}_{\mathbf{a}}(l)}}_{\text{LMA contribution, } \widetilde{\varepsilon}_{\mathrm{a}}(l)} \underbrace{- \hat{\mathbf{g}}_{\mathbf{e}}^H(l)\underbrace{\hat{\mathbf{C}}_{\mathbf{e}}^H(l)\begin{bmatrix}\mathrm{y}_1(l)\\ \mathbf{y}_{\mathbf{e}}(l)\end{bmatrix}}_{\mathbf{u}_{\mathrm{e}}(l)}}_{\text{XM contribution, } \widetilde{\varepsilon}_{\mathrm{e}}(l)} \tag{18}$$

where $\hat{\mathbf{C}}_{\mathbf{e}}(l)$ is defined as:

$$\hat{\mathbf{C}}_{\mathbf{e}}(l) = \begin{bmatrix} -\hat{\mathbf{h}}_{\mathbf{e}}^H(l) \\ \mathbf{I}_{M_{\mathrm{e}}} \end{bmatrix} \tag{19}$$

and $\mathbf{u}_{\mathbf{a}}(l)$ and $\mathbf{u}_{\mathrm{e}}(l)$ are the noise reference signals corresponding to the LMA and the XM signals respectively. It is apparent that there are two sets of updates that are required - (i) an update for $\hat{\mathbf{h}}_{\mathbf{e}}(l)$, which will subsequently be used to complete the blocking matrix $\mathbf{C}(l)$, by defining $\hat{\mathbf{C}}_{\mathbf{e}}(l)$, and (ii) an update for the ANC filter, $\hat{\mathbf{g}}(l)$.

It is also evident that the speech estimate in (18) consists of two distinct components, $\widetilde{\varepsilon}_{\mathrm{a}}$, as a result of the contribution from the LMA signals, and $\widetilde{\varepsilon}_{\mathrm{e}}$, from the contribution from the XM signals. It is clear that when $\hat{\mathbf{g}}_{\mathbf{e}} = \mathbf{0}$, the contribution from the XM signals is disabled and the error or speech estimate will be identical to that of the GSC-LMA in (11), i.e. $\hat{\mathbf{g}}_{\mathbf{a}} = \hat{\mathbf{v}}_{\mathbf{a}}$, and hence $\widetilde{\varepsilon}_{\mathrm{a}} = \widetilde{\mathrm{e}}_{\mathrm{a},1}$. However, in general, $\widetilde{\varepsilon}_{\mathrm{a}} \neq \widetilde{\mathrm{e}}_{\mathrm{a},1}$ as two different errors are minimised from (9) and (15).

Whereas in practice an NLMS approach could be used for updating $\hat{\mathbf{v}}_{\mathbf{a}}$ in the GSC-LMA, care should be taken for the approach used for updating $\hat{\mathbf{g}}$. This is because the power of the noise references from the XMs could be quite different as opposed to the case of the LMA, where it would be expected that the power of noise references from the LMA would be similar. Consequently, it is suggested that a diagonal step size normalised by the respective noise references be used in an NLMS context, or that a recursive least squares (RLS) [25] algorithm be used for updating $\hat{\mathbf{g}}$. A further analysis of adaptive techniques and their respective trade-offs is outside the scope of this paper.

A. Cross Correlation RTF estimate

In [24], using the cross-correlation method to compute $\hat{\mathbf{h}}_{\mathbf{e}}$ (for $M_{\mathrm{e}} = 1$), a GSC method as previously described was presented. The signal, $\widetilde{\varepsilon}_{\mathrm{a}}$ from (18), is used as a speech reference in order to carry out a cross-correlation with the XM signal for computing the RTF estimate. As opposed to $\widetilde{\varepsilon}_{\mathrm{a}}$, an alternative speech reference may be the output from the fixed beamformer, i.e. $\mathrm{y}_{\mathrm{f}} = \mathbf{f}_{\mathbf{a}}^H\mathbf{y}_{\mathbf{a}}$. Although this signal would be more noisy than $\widetilde{\varepsilon}_{\mathrm{a}}$, it will still be preferred to $\widetilde{\varepsilon}_{\mathrm{a}}$ due to its stability, i.e., it would be fixed and not time-varying due to adaptation. It should be noted however, that in using such a speech reference, this estimator takes into consideration the a priori information of the LMA. Hence, for $M_{\mathrm{e}} \geq 1$, the update of the i^{th} component of $\hat{\mathbf{h}}_{\mathbf{e,cc}}$ follows as:

$$\hat{\mathrm{h}}_{\mathrm{e},i,\mathrm{cc}}(l) = \begin{cases} \frac{r_{\mathrm{ea},i}(l)}{r_{\mathrm{aa},i}(l)} & \text{speech frames} \\ \hat{\mathrm{h}}_{\mathrm{e},i,\mathrm{cc}}(l-1) & \text{otherwise} \end{cases} \tag{20}$$

where

$$r_{\mathrm{ea},i}(l) = \alpha_{\mathrm{e},i}\, r_{\mathrm{ea},i}(l-1) + (1-\alpha_{\mathrm{e},i})\,\mathrm{y}_{\mathrm{e},i}(l)\,\mathrm{y}_{\mathrm{f}}^*(l) \tag{21}$$

$$r_{\mathrm{aa},i}(l) = \alpha_{\mathrm{e},i}\, r_{\mathrm{aa},i}(l-1) + (1-\alpha_{\mathrm{e},i})\,|\mathrm{y}_{\mathrm{f}}(l)|^2 \tag{22}$$

are computed in frames where speech is present and $\alpha_{e,i} \in [0,1]$ is a forgetting factor for the i^{th} XM component. Although this estimator is of low complexity, it is a biased estimator due to the presence of noise in y_{f}.

This GSC that uses the cross-correlation RTF estimate for the XM signals is referred to here as the GSC-LMA-XM-CC and can be encapsulated by the block diagram as shown in Fig. 2, similar to [24], except that $\mathrm{y_f}$ is used as opposed to $\widetilde{\varepsilon}_{\mathrm{a}}$ for updating $\hat{\mathrm{h}}_{\mathrm{e},i,\mathrm{cc}}$. It is reiterated here that the top branch remains unchanged from the GSC-LMA, and hence only changes are made to the lower branch. The cross-correlation RTF estimation procedure is used to complete the blocking matrix, $\mathbf{C}$ (i.e. define $\hat{\mathbf{C}}_{\mathbf{e}}$) and generate the extended set of noise references, $\mathbf{u} = [\mathbf{u_a}^T\ \mathbf{u_e}^T]^T$. The block diagram also intuitively depicts the two separate components of (18), with the speech estimate denoted as $\widetilde{\mathrm{e}}_{1,\mathrm{cc}}$. A further advantage of such a block scheme is that it does not compromise the initial structure of the GSC-LMA and can be interpreted as an "add-on" since it can easily be seen that if $\mathbf{g_e} = \mathbf{0}$, the GSC-LMA-XM-CC is reduced to the GSC-LMA.

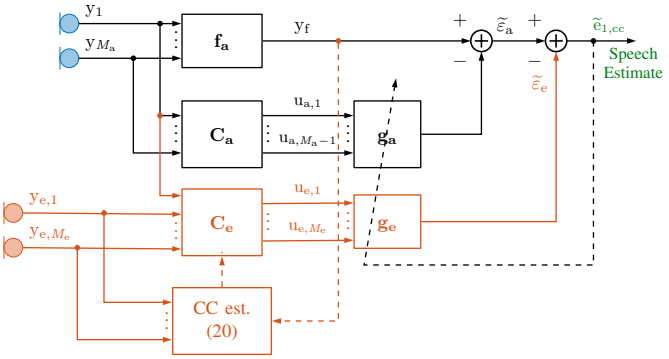

Fig. 2: GSC-LMA extended with XMs and using the cross corrleation method of estimating the missing RTF component for the XMs, GSC-LMA-XM-CC.

B. EVD RTF estimate

As a natural extension from [26] (i.e. from using an LMA to an LMA with XMs), a rank-1 model, $\mathbf{R_{x,r1}}$, for the speech-only correlation matrix, $\mathbf{R_{xx}}$, can be found from an eigenvalue decomposition (EVD) of the matrix $(\mathbf{R_{yy}} - \mathbf{R_{nn}})$, where the associated RTF vector is computed from the principal eigenvector. However, as shown in [23], 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 $\mathbf{R_{xx}}$, which can then be expressed as:

$$\mathbf{R_{x,r1}} = \hat{\sigma}^2_{\mathrm{x_{a,1}}}\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H = \hat{\sigma}^2_{\mathrm{x_{a,1}}}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^H\ \hat{\mathbf{h}}_{\mathbf{e}}^H\end{bmatrix} \tag{23}$$

where $\hat{\sigma}^2_{\mathrm{x_{a,1}}}$ is the estimated speech power in the first microphone of the LMA. Hence, computing $\hat{\mathbf{h}}_{\mathbf{e}}$ reduces to the following estimation problem:

$$\min_{\hat{\sigma}^2_{\mathrm{x_{a,1}}},\hat{\mathbf{h}}_{\mathbf{e}}} \quad ||(\mathbf{R_{yy}} - \mathbf{R_{nn}}) - \hat{\sigma}^2_{\mathrm{x_{a,1}}}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^H\ \hat{\mathbf{h}}_{\mathbf{e}}^H\end{bmatrix}||_F^2 \tag{24}$$

where $||.||_F$ is the Frobenius norm. In [23], for $M_{\mathrm{e}} = 1$, it has been demonstrated that by introducing a transform, (24) is further simplified and $\hat{\mathbf{h}}_{\mathbf{e}}$ can be computed from a 2×2 correlation matrix. In the following, this procedure is generalised for the case of $M_{\mathrm{e}} \geq 1$.

Proceeding to solve (24), an $M_{\mathrm{a}} \times (M_{\mathrm{a}}-1)$ blocking matrix $\mathbf{B_a}$ and a specific $M_{\mathrm{a}} \times 1$ fixed beamformer, $\mathbf{b_a}$ are defined such that:

$$\mathbf{B}_{\mathbf{a}}^H\widetilde{\mathbf{h}}_{\mathbf{a}} = \mathbf{0}; \qquad \mathbf{b_a} = \frac{\widetilde{\mathbf{h}}_{\mathbf{a}}}{||\widetilde{\mathbf{h}}_{\mathbf{a}}||} \tag{25}$$

where $\mathbf{B}_{\mathbf{a}}^H\mathbf{B_a} = \mathbf{I}_{(M_{\mathrm{a}}-1)}$. It should be noted that $\mathbf{B_a}$ can be computed from a QR decomposition of $\mathbf{C_a}$. Using $\mathbf{B_a}$ and $\mathbf{b_a}$, an $(M_{\mathrm{a}} + M_{\mathrm{e}}) \times (M_{\mathrm{a}} + M_{\mathrm{e}})$ unitary transformation matrix, $\mathbf{T}$, can be subsequently defined:

$$\mathbf{T} = \left[\begin{array}{c|c}\mathbf{T_a} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{I}_{M_{\mathrm{e}}}\end{array}\right] \tag{26}$$

where $\mathbf{T_a} = [\mathbf{B_a}\ \mathbf{b_a}]$, $\mathbf{T}_{\mathbf{a}}^H\mathbf{T_a} = \mathbf{I}_{M_{\mathrm{a}}}$, and hence $\mathbf{T}^H\mathbf{T} = \mathbf{I}_{(M_{\mathrm{a}}+M_{\mathrm{e}})}$. As the Frobenius norm is invariant under a unitary transformation [27], (24) can be rewritten as:

$$\min_{\hat{\sigma}^2_{\mathrm{x_{a,1}}},\hat{\mathbf{h}}_{\mathbf{e}}} \quad ||\mathbf{T}^H((\mathbf{R_{yy}} - \mathbf{R_{nn}}) - \hat{\sigma}^2_{\mathrm{x_{a,1}}}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix}\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^H\ \hat{\mathbf{h}}_{\mathbf{e}}^H\end{bmatrix})\mathbf{T}||_F^2 \tag{27}$$

By using (25) and (26), it can be seen that a transformed version of the RTF vector can be expressed as follows:

$$\mathbf{T}^H\begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix} = \begin{bmatrix}\mathbf{B}_{\mathbf{a}}^H\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \mathbf{b}_{\mathbf{a}}^H\widetilde{\mathbf{h}}_{\mathbf{a}}\\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ ||\widetilde{\mathbf{h}}_{\mathbf{a}}||\\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix} \tag{28}$$

and hence the expansion of (27) becomes:

$$\min_{\hat{\sigma}^2_{\mathrm{x_{a,1}}},\hat{\mathbf{h}}_{\mathbf{e}}} \quad ||\left[\begin{array}{c|c}\mathbf{K_{a_-}} & \mathbf{K}_{\mathbf{c}}^H\\ \hline \mathbf{K_c} & \mathbf{K_{e_+}}\end{array}\right] - \left[\begin{array}{c|c}\mathbf{0} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{K_{x,r1}}\end{array}\right]||_F^2 \tag{29}$$

where $\mathbf{K_{a_-}}$ is an $(M_{\mathrm{a}}-1) \times (M_{\mathrm{a}}-1)$ matrix, $\mathbf{K_c}$ an $(M_{\mathrm{e}}+1) \times (M_{\mathrm{a}}-1)$ matrix and $\mathbf{K_{e_+}}$ and $\mathbf{K_{x,r1}}$ are $(M_{\mathrm{e}}+1) \times (M_{\mathrm{e}}+1)$ matrices realised as:

$$\begin{aligned}\mathbf{K_{e_+}} &= \left[\begin{array}{c|c}\mathbf{b}_{\mathbf{a}}^H & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{I}_{M_{\mathrm{e}}}\end{array}\right](\mathbf{R_{yy}} - \mathbf{R_{nn}})\left[\begin{array}{c|c}\mathbf{b_a} & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{I}_{M_{\mathrm{e}}}\end{array}\right]\\ &= \mathbb{E}\left\{\begin{bmatrix}\mathbf{b}_{\mathbf{a}}^H\mathbf{y_a}\\ \mathbf{y_e}\end{bmatrix}\begin{bmatrix}\mathbf{y}_{\mathbf{a}}^H\mathbf{b_a}\ \mathbf{y}_{\mathbf{e}}^H\end{bmatrix}\right\} - \mathbb{E}\left\{\begin{bmatrix}\mathbf{b}_{\mathbf{a}}^H\mathbf{n_a}\\ \mathbf{n_e}\end{bmatrix}\begin{bmatrix}\mathbf{n}_{\mathbf{a}}^H\mathbf{b_a}\ \mathbf{n}_{\mathbf{e}}^H\end{bmatrix}\right\}\end{aligned} \tag{30}$$

$$\mathbf{K_{x,r1}} = \hat{\sigma}^2_{\mathrm{x_{a,1}}}\begin{bmatrix}||\widetilde{\mathbf{h}}_{\mathbf{a}}||\\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix}\begin{bmatrix}||\widetilde{\mathbf{h}}_{\mathbf{a}}||\ \hat{\mathbf{h}}_{\mathbf{e}}^H\end{bmatrix} \tag{31}$$

From (29), it can be seen that the additional a priori knowledge of a known $\widetilde{\mathbf{h}}_{\mathbf{a}}$ reduces the estimation problem further to:

$$\min_{\hat{\sigma}^2_{\mathrm{x_{a,1}}},\hat{\mathbf{h}}_{\mathbf{e}}} \quad ||\mathbf{K_{e_+}} - \mathbf{K_{x,r1}}||_F^2 \tag{32}$$

which is that of a rank-1 approximation of the $(M_{\mathrm{e}}+1) \times (M_{\mathrm{e}}+1)$ matrix, $\mathbf{K_{e_+}}$. Computing $\hat{\mathbf{h}}_{\mathbf{e}}$ follows by initially extracting the principal eigenvector, $\mathbf{k_{max}} = [\mathrm{k_a}\ \mathbf{k}_{\mathbf{e}}^T]^T$, corresponding to the largest eigenvalue of $\mathbf{K_{e_+}}$. Applying the appropriate scaling and normalisation of the elements in $\mathbf{k_{max}}$, $\hat{\mathbf{h}}_{\mathbf{e}}$ is then given by:

$$\hat{\mathbf{h}}_{\mathbf{e,evd}} = \frac{||\widetilde{\mathbf{h}}_{\mathbf{a}}||\ \mathbf{k_e}}{\mathrm{k_a}} \tag{33}$$

This EVD-based RTF estimation method can easily be realised in a GSC scheme similar to that of the cross-correlation method as illustrated in Fig. 3, which will be referred to as the GSC-LMA-XM-EVD. In this case, however, a specific fixed beamformer of $\mathbf{f_a} = \mathbf{b_a}$ is required. The output from the fixed beamformer, $\mathrm{y_f}$, and XM signals, $\mathbf{y_e}$, are then used to generate the correlation matrix $\mathbf{K}_{\mathbf{e}_+}$ from (30). The first term of (30) is updated when speech is active, and the second term updated in noise-only periods. $\hat{\mathbf{h}}_{\mathbf{e,evd}}$ is computed accordingly and used to generate the extra noise reference, which completes the missing part of the blocking matrix, $\hat{\mathbf{C}}_\mathbf{e}$. It is also noted that although another blocking matrix is defined in (25), this is only used for the derivation in computing $\hat{\mathbf{h}}_{\mathbf{e,evd}}$. Consequently, the GSC-LMA-XM-EVD scheme as depicted in Fig. 3 still uses $\mathbf{C_a}$ and $\hat{\mathbf{C}}_\mathbf{e}$ as the blocking matrices, and the procedure of completing the blocking matrix follows as previously described, with the speech estimate, $\widetilde{\mathrm{e}}_{1,\mathrm{evd}}$.

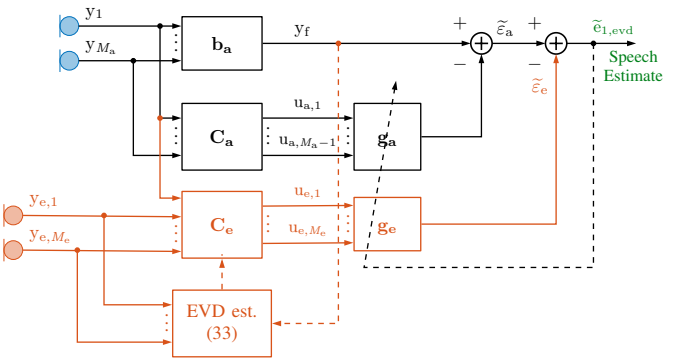

Fig. 3: GSC-LMA extended with XMs and using the EVD method of estimating the missing RTF component for the XMs, GSC-LMA-XM-EVD.

VI. Rank-1 GEVD Method

In [23], a method of computing $\hat{\mathbf{h}}_\mathbf{e}$ (for $M_\mathrm{e} = 1$) using covariance whitening, or equivalently, a GEVD has been presented. In this section, some modifications will be made to this method, as well as an extension for the general case of $M_\mathrm{e} \geq 1$, which will lead to an alternative scheme compared to the previous section. This new scheme will still make use of the GSC-LMA, and the inclusion of the XM will once again be used as an "add-on" to the noise reduction system. As the mathematical derivations involved may detract from the conceptual aspect of this method, an overview of the resulting scheme and its utility is firstly presented in this section, followed by the relevant mathematical details.

A. Overview of the method

Fig. 4 reveals the resulting scheme, which will be referred to as the GSC-LMA-XM-GEVD. Firstly, the (M_a+ M_e) signals will undergo the transformation from (39), which is simply the application of the fixed beamformer, $\mathbf{f_a}$, and the blocking matrix, $\mathbf{C_a}$ on the LMA signals as is done in the GSC-LMA, along with the unmodified XM signals.

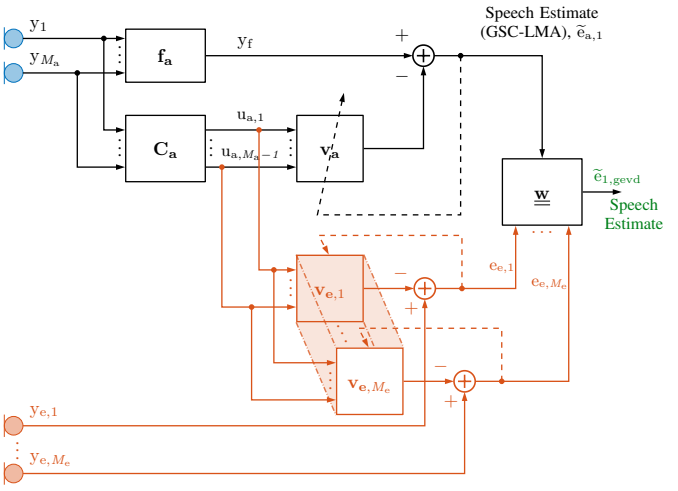

Fig. 4: GSC-LMA extended with XMs involving a rank-1 GEVD-based RTF estimation procedure, GSC-LMA-XM-GEVD.

This is then followed by the orthogonalisation of the noise components of $\mathrm{y_f}$ and $\mathbf{y_e}$ onto the noise components of $\mathbf{u_a}$. Such an orthogonalisation can be performed in noise-only periods by using adaptive filters. The resulting $(M_\mathrm{e}+1)$ signals after this orthogonalisation procedure are then denoted as:

$$\underline{\underline{\mathbf{y}}}(l) = \left[\widetilde{\mathrm{e}}_{\mathrm{a},1}(l) \;\; \mathrm{e}_{\mathrm{e},1}(l) \;\; \ldots \;\; \mathrm{e}_{\mathrm{e},M_\mathrm{e}}(l)\right]^T \tag{34}$$

consisting of the speech output from a GSC-LMA, $\widetilde{\mathrm{e}}_{\mathrm{a},1}(l)$, and the vector of XM signals who have had their noise components orthogonalised onto the noise components of $\mathbf{u_a}$, $\left[\mathrm{e}_{\mathrm{e},1}(l) \;\; \ldots \;\; \mathrm{e}_{\mathrm{e},M_\mathrm{e}}(l)\right]^T$. Since the orthogonalisation of the noise components of $\mathrm{y_f}$ onto the noise components of $\mathbf{u_a}$ is equivalent to considering the optimisation equation of (9) from the GSC-LMA, the speech output from a GSC-LMA, $\widetilde{\mathrm{e}}_{\mathrm{a},1}$ corresponds to the first element in $\underline{\underline{\mathbf{y}}}$.

For the orthogonalisation involving the XM signals, a separate $(M_\mathrm{a} - 1) \times 1$ adaptive filter, $\mathbf{v}_{\mathbf{e},i}$ will have to be introduced for each of the i^{th} XMs, such that it minimises the same equation as in (9), but with $\mathrm{y}_{\mathrm{e},i}$ as the desired signal. Therefore, the optimal filter for $\mathbf{v}_{\mathbf{e},i}$ can be computed in noise-only frames as:

$$\hat{\mathbf{v}}_{\mathbf{e},i}(l) = (\mathbf{C}_\mathbf{a}^H \mathbf{R}_{\mathbf{n_a n_a}}(l)\mathbf{C_a})^{-1}\mathbf{C}_\mathbf{a}^H \mathbf{R}_{\mathbf{n_a}\mathrm{n}_{\mathbf{e},i}}(l) \tag{35}$$

where $\mathbf{R}_{\mathbf{n_a}\mathrm{n}_{\mathbf{e},i}} = \mathbb{E}\{\mathbf{n_a}\mathrm{n}^*_{\mathrm{e},i}\}$ The resulting error from this orthogonalisation step for the i^{th} XM is then:

$$\mathrm{e}_{\mathrm{e},i}(l) = \mathrm{y}_{\mathrm{e},i}(l) - \mathbf{v}^H_{\mathbf{e},i}(l)\mathbf{C}_\mathbf{a}^H \mathbf{y_a}(l) \tag{36}$$

Finally, the filter, $\underline{\underline{\mathbf{w}}}$, which involves a GEVD procedure (derived in the following section), can be used to filter the signals $\underline{\underline{\mathbf{y}}}$ in the corresponding time frame to yield the corresponding speech estimate, $\widetilde{\mathrm{e}}_{1,\mathrm{gevd}}$.

From Fig. 4, it can easily be observed that the XMs are truly incorporated in a modular fashion or as "add-ons" to an existing GSC-LMA. One advantage of this implementation over the previously described approach of completing the blocking

matrix is that the speech estimate, $\widetilde{\mathrm{e}}_{\mathrm{a},1}$ is still available and could be reverted to in cases where using the XM signals may yield undesirable behaviour.

Furthermore, as will be demonstrated in (62), $\underline{\underline{\mathbf{w}}}$ is a low dimensional MVDR beamformer, which uses a rank-1 GEVD-based RTF estimate. Consequently, the GSC-LMA-XM-GEVD represents a combination of a GSC-LMA and an MVDR with a rank-1 GEVD-based RTF estimate, which encompasses a variety of filtering schemes. In particular, for $M_{\mathrm{e}} = 0$, the GSC-LMA-XM-GEVD is equivalent to the GSC-LMA, and for $M_{\mathrm{a}} = 1, M_{\mathrm{e}} \neq 0$, the GSC-LMA-XM-GEVD is a MVDR with a rank-1 GEVD-based RTF estimate, which uses the single local microphone as the reference signal.

In the sections that follow, a derivation is firstly given for computing $\hat{\mathbf{h}}_{\mathbf{e}}$, followed by how this is incorporated into the MVDR-LMA-XM from section IV in order to yield the GSC-LMA-XM-GEVD depicted in Fig. 4.

B. GEVD-based RTF estimation

In order to extend the estimation problem of (24) to a GEVD, a spatial pre-whitening (or orthogonalisation) operation is firstly defined from the noise-only correlation matrix using the Cholesky decomposition:

$$\mathbf{R}_{\mathbf{nn}} = \mathbf{R}_{\mathbf{nn}}^{1/2}\,\mathbf{R}_{\mathbf{nn}}^{H/2} \tag{37}$$

where $\mathbf{R}_{\mathbf{nn}}^{1/2}$ is a lower triangular matrix, and $\mathbf{R}_{\mathbf{nn}}^{H/2}$ is its hermitian transpose. Spatial pre-whitening is then performed by pre-multiplying the signal vector of interest by $\mathbf{R}_{\mathbf{nn}}^{-1/2}$. For an autocorrelation matrix, spatial pre-whitening is performed by pre-multiplying it by $\mathbf{R}_{\mathbf{nn}}^{-1/2}$ and post-multiplying it by $\mathbf{R}_{\mathbf{nn}}^{-H/2}$. Therefore, the pre-whitened version of (24) becomes:

$$\min_{\hat{\sigma}^2_{\mathrm{x}_{\mathrm{a},1}},\,\hat{\mathbf{h}}_{\mathbf{e}}} ||\mathbf{R}_{\mathbf{nn}}^{-1/2}((\mathbf{R}_{\mathbf{yy}} - \mathbf{R}_{\mathbf{nn}}) - \hat{\sigma}^2_{\mathrm{x}_{\mathrm{a},1}} \begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix} \begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^H & \hat{\mathbf{h}}_{\mathbf{e}}^H\end{bmatrix})\mathbf{R}_{\mathbf{nn}}^{-H/2}||_F^2 \tag{38}$$

In [23], an appropriate transformation matrix has been defined and then (38) is solved (for $M_{\mathrm{e}} = 1$) by performing an EVD on a 2×2 pre-whitened correlation matrix.

An alternative approach can however be taken to solve (38) that will yield a practical scheme and include the general case of $M_{\mathrm{e}} \geq 1$. Firstly, using the blocking matrix, $\mathbf{C}_{\mathbf{a}}$ from (8), and the fixed beamformer, $\mathbf{f}_{\mathbf{a}}$, the $(M_{\mathrm{a}} + M_{\mathrm{e}}) \times (M_{\mathrm{a}} + M_{\mathrm{e}})$ transformation matrix, $\boldsymbol{\Upsilon}_{\mathbf{1}}$, can be defined:

$$\boldsymbol{\Upsilon}_{\mathbf{1}} = \left[\begin{array}{c|c} [\mathbf{C}_{\mathbf{a}} \;\; \mathbf{f}_{\mathbf{a}}] & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_{M_{\mathrm{e}}} \end{array}\right] \tag{39}$$

from which the transformed speech-plus-noise signals and the transformed noise-only signals are defined respectively as:

$$\boldsymbol{\Upsilon}_{\mathbf{1}}^H\mathbf{y} = \begin{bmatrix}\mathbf{C}_{\mathbf{a}}^H\mathbf{y}_{\mathbf{a}} \\ \mathbf{f}_{\mathbf{a}}^H\mathbf{y}_{\mathbf{a}} \\ \mathbf{y}_{\mathbf{e}}\end{bmatrix};\;\; \boldsymbol{\Upsilon}_{\mathbf{1}}^H\mathbf{n} = \begin{bmatrix}\mathbf{C}_{\mathbf{a}}^H\mathbf{n}_{\mathbf{a}} \\ \mathbf{f}_{\mathbf{a}}^H\mathbf{n}_{\mathbf{a}} \\ \mathbf{n}_{\mathbf{e}}\end{bmatrix} \tag{40}$$

consisting of the blocking matrix signals from the LMA, the fixed beamformer output signal, and the XM signals.

Another spatial pre-whitening operation can be subsequently defined from the noise-only correlation matrix of the transformed noise-only signals, similar to that of (37):

$$\mathbb{E}\{(\boldsymbol{\Upsilon}_{\mathbf{1}}^H\mathbf{n})(\boldsymbol{\Upsilon}_{\mathbf{1}}^H\mathbf{n})^H\} = \mathbf{L}\mathbf{L}^H \tag{41}$$

where $\mathbf{L}$ is an $(M_{\mathrm{a}}+M_{\mathrm{e}})\times(M_{\mathrm{a}}+M_{\mathrm{e}})$ lower triangular matrix:

$$\mathbf{L} = \left[\begin{array}{c|c} \mathbf{L}_{\mathbf{a}_-} & \mathbf{0} \\ \hline \mathbf{L}_{\mathbf{c}} & \mathbf{L}_{\mathbf{e}_+} \end{array}\right] \tag{42}$$

whose block dimensions are such that $\mathbf{L}_{\mathbf{a}_-}$ is an $(M_{\mathrm{a}}-1) \times (M_{\mathrm{a}}-1)$ lower triangular matrix, $\mathbf{L}_{\mathbf{c}}$ an $(M_{\mathrm{e}}+1) \times (M_{\mathrm{a}}-1)$, and $\mathbf{L}_{\mathbf{e}_+}$ is an $(M_{\mathrm{e}}+1) \times (M_{\mathrm{e}}+1)$ lower triangular matrix. By computing the block inverse of $\mathbf{L}$ in (42), $\mathbf{L}^{-1}$ is then:

$$\begin{aligned}\mathbf{L}^{-1} &= \left[\begin{array}{c|c} \mathbf{L}_{\mathbf{a}_-}^{-1} & \mathbf{0} \\ \hline -\mathbf{L}_{\mathbf{e}_+}^{-1}\mathbf{L}_{\mathbf{c}}\,\mathbf{L}_{\mathbf{a}_-}^{-1} & \mathbf{L}_{\mathbf{e}_+}^{-1} \end{array}\right] \\ &= \underbrace{\left[\begin{array}{c|c} \mathbf{L}_{\mathbf{a}_-}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{L}_{\mathbf{e}_+}^{-1} \end{array}\right]}_{\boldsymbol{\Upsilon}_{\mathbf{3}}^H} \underbrace{\left[\begin{array}{c|c} \mathbf{I}_{M_{\mathrm{a}}-1} & \mathbf{0} \\ \hline -\mathbf{L}_{\mathbf{c}}\,\mathbf{L}_{\mathbf{a}_-}^{-1} & \mathbf{I}_{M_{\mathrm{e}}+1} \end{array}\right]}_{\boldsymbol{\Upsilon}_{\mathbf{2}}^H}\end{aligned} \tag{43}$$

which is split into two separate orthogonalisations, $\boldsymbol{\Upsilon}_{\mathbf{2}}^H$, which will orthogonalise the noise components from the last $(M_{\mathrm{e}} + 1)$ signals onto the first $(M_{\mathrm{a}} - 1)$ signals, and $\boldsymbol{\Upsilon}_{\mathbf{3}}^H$, which will complete the entire orthogonalisation operation. A second transformation of the transformed speech-plus-noise and the noise-only signals from (40) can then be defined respectively such that:

$$\underline{\mathbf{y}} = \boldsymbol{\Upsilon}_{\mathbf{2}}^H\boldsymbol{\Upsilon}_{\mathbf{1}}^H\mathbf{y};\;\; \underline{\mathbf{n}} = \boldsymbol{\Upsilon}_{\mathbf{2}}^H\boldsymbol{\Upsilon}_{\mathbf{1}}^H\mathbf{n} \tag{44}$$

which can be interpreted as the same set of transformed signals from (40), but with the noise components from the fixed beamformer output as well as the XM signals orthogonalised onto the noise components of the blocking matrix signals from the LMA (i.e. orthogonalised onto $\mathbf{u}_{\mathbf{a}}$).

With the relevant transformations defined, the estimation problem from (38) can be subsequently re-formulated as:

$$\begin{aligned}\min_{\hat{\sigma}^2_{\mathrm{x}_{\mathrm{a},1}},\,\hat{\mathbf{h}}_{\mathbf{e}}} ||\mathbf{R}_{\mathbf{nn}}^{-1/2}\boldsymbol{\Upsilon}_{\mathbf{1}}^{-H}\boldsymbol{\Upsilon}_{\mathbf{2}}^{-H}\boldsymbol{\Upsilon}_{\mathbf{2}}^H\boldsymbol{\Upsilon}_{\mathbf{1}}^H((\mathbf{R}_{\mathbf{yy}} - \mathbf{R}_{\mathbf{nn}}) \\ - \hat{\sigma}^2_{\mathrm{x}_{\mathrm{a},1}} \begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix} \begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^H & \hat{\mathbf{h}}_{\mathbf{e}}^H\end{bmatrix})\boldsymbol{\Upsilon}_{\mathbf{1}}\boldsymbol{\Upsilon}_{\mathbf{2}}\boldsymbol{\Upsilon}_{\mathbf{2}}^{-1}\boldsymbol{\Upsilon}_{\mathbf{1}}^{-1}\mathbf{R}_{\mathbf{nn}}^{-H/2}||_F^2\end{aligned} \tag{45}$$

which can be simplified by realising:

$$\begin{aligned}\mathbf{R}_{\mathbf{nn}}^{-1/2}\boldsymbol{\Upsilon}_{\mathbf{1}}^{-H}\boldsymbol{\Upsilon}_{\mathbf{2}}^{-H} &= (\boldsymbol{\Upsilon}_{\mathbf{1}}^H\mathbf{R}_{\mathbf{nn}}^{1/2})^{-1}\boldsymbol{\Upsilon}_{\mathbf{2}}^{-H} \\ &= (\mathbf{L}\boldsymbol{\Theta})^{-1}\boldsymbol{\Upsilon}_{\mathbf{2}}^{-H} \\ &= \boldsymbol{\Theta}^H\underbrace{\mathbf{L}^{-1}\boldsymbol{\Upsilon}_{\mathbf{2}}^{-H}}_{\boldsymbol{\Upsilon}_{\mathbf{3}}^H}\end{aligned} \tag{46}$$

where $\boldsymbol{\Theta}$ is some unitary matrix. Since the Frobenius norm is invariant under a unitary transformation, (45) can be re-written as:

$$\begin{aligned}\min_{\hat{\sigma}^2_{\mathrm{x}_{\mathrm{a},1}},\,\hat{\mathbf{h}}_{\mathbf{e}}} ||\boldsymbol{\Upsilon}_{\mathbf{3}}^H(\underline{\mathbf{R}}_{\mathbf{yy}} - \underline{\mathbf{R}}_{\mathbf{nn}})\boldsymbol{\Upsilon}_{\mathbf{3}} \\ - \boldsymbol{\Upsilon}_{\mathbf{3}}^H\boldsymbol{\Upsilon}_{\mathbf{2}}^H\boldsymbol{\Upsilon}_{\mathbf{1}}^H(\hat{\sigma}^2_{\mathrm{x}_{\mathrm{a},1}} \begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathbf{h}}_{\mathbf{e}}\end{bmatrix} \begin{bmatrix}\widetilde{\mathbf{h}}_{\mathbf{a}}^H & \hat{\mathbf{h}}_{\mathbf{e}}^H\end{bmatrix})\boldsymbol{\Upsilon}_{\mathbf{1}}\boldsymbol{\Upsilon}_{\mathbf{2}}\boldsymbol{\Upsilon}_{\mathbf{3}}||_F^2\end{aligned} \tag{47}$$

where:

$$\underline{\underline{\mathbf{R}}}_{\mathbf{yy}} = \boldsymbol{\Upsilon}_{\mathbf{2}}^H \boldsymbol{\Upsilon}_{\mathbf{1}}^H \mathbf{R}_{\mathbf{yy}} \boldsymbol{\Upsilon}_{\mathbf{1}} \boldsymbol{\Upsilon}_{\mathbf{2}} = \mathbb{E}\{\underline{\mathbf{y}}\,\underline{\mathbf{y}}^H\} \tag{48}$$

$$\underline{\underline{\mathbf{R}}}_{\mathbf{nn}} = \boldsymbol{\Upsilon}_{\mathbf{2}}^H \boldsymbol{\Upsilon}_{\mathbf{1}}^H \mathbf{R}_{\mathbf{nn}} \boldsymbol{\Upsilon}_{\mathbf{1}} \boldsymbol{\Upsilon}_{\mathbf{2}} = \left[\begin{array}{c|c} \mathbf{L}_{\mathbf{a}_-} \mathbf{L}_{\mathbf{a}_-}^H & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{L}_{\mathbf{e}_+} \mathbf{L}_{\mathbf{e}_+}^H \end{array}\right] \tag{49}$$

and also since $\mathbf{C}_{\mathbf{a}}^H \widetilde{\mathbf{h}}_{\mathbf{a}} = \mathbf{0}$:

$$\boldsymbol{\Upsilon}_{\mathbf{2}}^H \boldsymbol{\Upsilon}_{\mathbf{1}}^H \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathbf{h}}_{\mathbf{e}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathrm{d_a} \\ \hat{\mathbf{h}}_{\mathbf{e}} \end{bmatrix} \tag{50}$$

which consists of a vector with $(M_\mathrm{a} - 1)$ zeros, together with $\mathrm{d_a} = \mathbf{f}_{\mathbf{a}}^H \widetilde{\mathbf{h}}_{\mathbf{a}}$ and the RTF vector for the XM signals to be estimated. Upon substitution of (50) into (47), it can be observed that the dimensionality of (47) can be reduced to a lower order of $(M_\mathrm{e} + 1)$, similarly to what was done in the EVD method of section V-B. The lower order $(M_\mathrm{e}+1)$ speech-plus-noise and noise-only signals can be defined as follows, denoted with the double underbar:

$$\underline{\underline{\mathbf{y}}} = \begin{bmatrix} \mathbf{0}_{(M_\mathrm{e}+1)\times(M_\mathrm{a}-1)} & \mathbf{I}_{M_\mathrm{e}+1} \end{bmatrix} \underline{\mathbf{y}} \tag{51}$$

$$\underline{\underline{\mathbf{n}}} = \begin{bmatrix} \mathbf{0}_{(M_\mathrm{e}+1)\times(M_\mathrm{a}-1)} & \mathbf{I}_{M_\mathrm{e}+1} \end{bmatrix} \underline{\mathbf{n}} \tag{52}$$

which consists of the single fixed beamformer output signal and the M_e XM signals, both of which have had their noise components orthogonalised onto the blocking matrix signals. Therefore, (47) can be reduced to:

$$\min_{\hat{\sigma}^2_{\mathrm{x_{a,1}}},\, \hat{\mathbf{h}}_{\mathbf{e}}} ||\mathbf{L}_{\mathbf{e}_+}^{-1}((\underline{\underline{\mathbf{R}}}_{\mathbf{yy}} - \underline{\underline{\mathbf{R}}}_{\mathbf{nn}}) - (\hat{\sigma}^2_{\mathrm{x_{a,1}}} \begin{bmatrix} \mathrm{d_a} \\ \hat{\mathbf{h}}_{\mathbf{e}} \end{bmatrix} \begin{bmatrix} \mathrm{d_a^*} & \hat{\mathbf{h}}_{\mathbf{e}}^H \end{bmatrix}))\mathbf{L}_{\mathbf{e}_+}^{-H}||_F^2 \tag{53}$$

where the $(M_\mathrm{e} + 1) \times (M_\mathrm{e} + 1)$ matrices $\underline{\underline{\mathbf{R}}}_{\mathbf{yy}}$ and $\underline{\underline{\mathbf{R}}}_{\mathbf{nn}}$ are:

$$\underline{\underline{\mathbf{R}}}_{\mathbf{yy}} = \mathbb{E}\{\underline{\underline{\mathbf{y}}}\,\underline{\underline{\mathbf{y}}}^H\}; \quad \underline{\underline{\mathbf{R}}}_{\mathbf{nn}} = \mathbb{E}\{\underline{\underline{\mathbf{n}}}\,\underline{\underline{\mathbf{n}}}^H\} = \mathbf{L}_{\mathbf{e}_+} \mathbf{L}_{\mathbf{e}_+}^H \tag{54}$$

and hence $\mathbf{L}_{\mathbf{e}_+}^{-1} \underline{\underline{\mathbf{R}}}_{\mathbf{nn}} \mathbf{L}_{\mathbf{e}_+}^{-H} = \mathbf{I}_{M_\mathrm{e}+1}$.

The solution then follows from a GEVD of the matrix pencil $\{\underline{\underline{\mathbf{R}}}_{\mathbf{yy}}, \underline{\underline{\mathbf{R}}}_{\mathbf{nn}}\}$, where a joint diagonalisation is done:

$$\underline{\underline{\mathbf{R}}}_{\mathbf{yy}} = \mathbf{Q} \boldsymbol{\Sigma}_{\mathbf{y}} \mathbf{Q}^H; \;\; \underline{\underline{\mathbf{R}}}_{\mathbf{nn}} = \mathbf{Q} \boldsymbol{\Sigma}_{\mathbf{n}} \mathbf{Q}^H \tag{55}$$

where $\mathbf{Q}$ is a full-rank, $(M_\mathrm{e} + 1) \times (M_\mathrm{e} + 1)$ invertible matrix, $\boldsymbol{\Sigma}_{\mathbf{y}}$ and $\boldsymbol{\Sigma}_{\mathbf{n}}$ are real valued $(M_\mathrm{e} + 1) \times (M_\mathrm{e} + 1)$ diagonal matrices arranged in descending order according to the magnitude of the generalised eigenvalues, i.e. $\boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \boldsymbol{\Sigma}_{\mathbf{y}}$. Using the principal eigenvalue and corresponding eigenvector from the matrix, $(\underline{\underline{\mathbf{R}}}_{\mathbf{yy}} - \underline{\underline{\mathbf{R}}}_{\mathbf{nn}})$, it then evident that:

$$\hat{\sigma}^2_{\mathrm{x_{a,1}}} \begin{bmatrix} \mathrm{d_a} \\ \hat{\mathbf{h}}_{\mathbf{e}} \end{bmatrix} \begin{bmatrix} \mathrm{d_a^*} & \hat{\mathbf{h}}_{\mathbf{e}}^H \end{bmatrix} = \mathbf{Q} \mathbf{e_1} \mathbf{e_1}^T (\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{n}}) \mathbf{e_1} \mathbf{e_1}^T \mathbf{Q}^H \tag{56}$$

where the $(M_\mathrm{e}+1) \times 1$ vector, $\mathbf{e_1} = [1 \;\; 0 \ldots 0]^T$. The lower dimensional RTF vector then follows as:

$$\begin{bmatrix} \mathrm{d_a} \\ \hat{\mathbf{h}}_{\mathbf{e}} \end{bmatrix} = \frac{\mathbf{Q} \mathbf{e_1}\, \mathrm{d_a}}{\mathbf{e_1}^T \mathbf{Q} \mathbf{e_1}} \tag{57}$$

This GEVD-based $\hat{\mathbf{h}}_{\mathbf{e}}$ from (57) could then, in fact, be used as a third option for completing the blocking matrix as was done in section V. However, as will be demonstrated in the following section, a substitution of (57) into the MVDR-LMA-XM beamformer will reveal a convenient sequence of operations that leads to the alternative practical implementation depicted in Fig. 4 that does not affect the output of the GSC-LMA.

C. MVDR-LMA-XM beamformer

Using the definitions from (41), and the result of (57), the numerator of (14) can firstly be written as:

$$\mathbf{R}_{\mathbf{nn}}^{-1} \widetilde{\mathbf{h}} = \boldsymbol{\Upsilon}_{\mathbf{1}} \mathbf{L}^{-H} \mathbf{L}^{-1} \boldsymbol{\Upsilon}_{\mathbf{1}}^H \begin{bmatrix} \widetilde{\mathbf{h}}_{\mathbf{a}} \\ \hat{\mathbf{h}}_{\mathbf{e}} \end{bmatrix} \tag{58}$$

Using (42), (43), (54), and (57) eventually results in:

$$\mathbf{R}_{\mathbf{nn}}^{-1} \widetilde{\mathbf{h}} = \boldsymbol{\Upsilon}_{\mathbf{1}} \begin{bmatrix} -\mathbf{L}_{\mathbf{a}_-}^{-H} \mathbf{L}_{\mathbf{c}}^H \\ \mathbf{I}_{M_\mathrm{e}+1} \end{bmatrix} \mathbf{Q}^{-H} \boldsymbol{\Sigma}_{\mathbf{n}}^{-1} \frac{\mathbf{e_1}\, \mathrm{d_a}}{\mathbf{e_1}^T \mathbf{Q} \mathbf{e_1}} \tag{59}$$

Finally, making the relevant substitutions in the denominator of (14), the MVDR-LMA-XM beamformer becomes:

$$\widetilde{\mathbf{w}} = \underbrace{\boldsymbol{\Upsilon}_{\mathbf{1}}}_{\text{Trans}} \underbrace{\begin{bmatrix} -\mathbf{L}_{\mathbf{a}_-}^{-H} \mathbf{L}_{\mathbf{c}}^H \\ \mathbf{I}_{M_\mathrm{e}+1} \end{bmatrix}}_{\text{Orthogonalisation}} \underbrace{\mathbf{Q}^{-H} \mathbf{e_1}}_{\text{GEVD}} \underbrace{\frac{\mathbf{e_1}^T \mathbf{Q}^H \mathbf{e_1}}{\mathrm{d_a}}}_{\text{scaling}} \tag{60}$$

which reveals the distinct operations that are required in order to implement the scheme of Fig. 4. Namely, it consists of the transformation operation from (39), the orthogonalisation operation onto the noise components of the blocking matrix signals from the LMA, the GEVD operation and the appropriate scaling.

Consequently, the resulting speech estimate, i.e. $\mathbf{w}^H \mathbf{y}$, is then computed as:

$$\widetilde{\mathrm{e}}_{1,\mathrm{gevd}} = \underbrace{\frac{\mathbf{e_1}^T \mathbf{Q} \mathbf{e_1}}{\mathrm{d_a}} \;\; \mathbf{e_1}^T \mathbf{Q}^{-1}}_{\underline{\underline{\mathbf{w}}}^H} \underbrace{[-\mathbf{L_c} \mathbf{L}_{\mathbf{a}_-}^{-1} \mid \mathbf{I}_{M_\mathrm{e}+1}] \begin{bmatrix} \mathbf{C}_{\mathbf{a}}^H \mathbf{y_a} \\ \mathbf{f}_{\mathbf{a}}^H \mathbf{y_a} \\ \mathbf{y_e} \end{bmatrix}}_{\underline{\underline{\mathbf{y}}}} \tag{61}$$

which can be realised as a lower dimensional [$(M_\mathrm{e} + 1)$ tap] beamformer, $\underline{\underline{\mathbf{w}}}$ that acts directly on the transformed, orthogonalised signals, $\underline{\underline{\mathbf{y}}}$ (see (34)). It is also noted that $\underline{\underline{\mathbf{w}}}$ can be equivalently formulated in the familiar MVDR beamformer structure as:

$$\underline{\underline{\mathbf{w}}} = \frac{\underline{\underline{\mathbf{R}}}_{\mathbf{nn}}^{-1} \underline{\underline{\widetilde{\mathbf{h}}}}}{\underline{\underline{\widetilde{\mathbf{h}}}}^H \underline{\underline{\mathbf{R}}}_{\mathbf{nn}}^{-1} \underline{\underline{\widetilde{\mathbf{h}}}}} \tag{62}$$

where the lower dimensional RTF vector, $\underline{\underline{\widetilde{\mathbf{h}}}} = \begin{bmatrix} \mathrm{d_a} & \hat{\mathbf{h}}_{\mathbf{e}} \end{bmatrix}^T$.

VII. Evaluation and Discussion

The various algorithms were evaluated on audio recordings made in an office room of dimensions $5.4\,\mathrm{m} \times 3.5\,\mathrm{m} \times 2.5\,\mathrm{m}$ with an estimated broadband reverberation time of $0.3\,\mathrm{s}$. The scenario under which audio recordings were made is depicted in Fig. 5. At the centre of the scenario, a test subject wore a single dummy behind-the-ear (BTE) hearing aid (HA) equipped with two microphones spaced approximately $1.3\,\mathrm{cm}$ apart, which served as the LMA. The test subject was instructed to always face towards the direction of $0°$, i.e. in the direction of the speech source, which was placed $1\,\mathrm{m}$ away. Four XMs were distributed as shown, with XM3

being worn by the test subject, clipped onto the chest. AKG-CK-97-O microphones were used for XM1 and XM3, and AKG-CK-32 microphones were used for XM2 and XM4. A Genelec 8020C loudspeaker was used to generate the speech signal (SS1), which was that of a male speaker from [28]. Several loudspeakers to serve as noise sources (NS) were also distributed as shown, also $1\,\mathrm{m}$ away from the test subject. NS1 and NS2 were JBL Control 1 Pro loudspeakers, and NS3 and NS4 were Harman-Kardon HK206 loudspeakers. The noise signals considered were two uncorrelated excerpts of multitalker babble noise (`bb1` and `bb2`) from [29], speech-shaped noise (`sn1`) constructed from a 12-coefficient linear predictive coder (LPC) using the male speaker from [28], and noise from the interior of a busy coffee shop (`cc1`) [30]. With the test subject facing toward $0°$, each of these noises played through each of the NS positions was recorded separately. The speech signal from SS1 was also separately recorded. All recorded signals had a duration of $60\,\mathrm{s}$. With this database of speech and noises, subsets of several scenarios stemming from that of Fig. 5 could then be constructed and analysed accordingly.

For the processing of the algorithms in the various acoustic scenarios, the Weighted Overlap and Add (WOLA) method [31], with a Discrete Fourier Transform (DFT) size of 256, $50\,\%$ overlap, a square-root Hanning window, and sampling frequency of $16\,\mathrm{kHz}$ was used. Separate experiments were done using either a perfect voice activity detector (VAD) or an imperfect VAD (using the minimum statistics method from [32]) to indicate the time frames where speech was present. All RTF estimates were performed in time frames where periods where the speech was present (as indicated by the respective VAD) and all the relevant correlation matrices were computed using a forgetting factor corresponding to an averaging time of $1\,\mathrm{s}$. For all experiments, the correlation matrices were initialised such that all elements were set to zero. It should be noted that these VADs were applied on the time domain noisy signal of the microphone, LM1. Hence if it was detected that a certain time frame contained speech, all frequency bins were subsequently treated as such. As not all frequency bins will truly contain speech, the use of the word "perfect" in this context is somewhat of a misnomer. It is rather used to distinguish two instances of estimating the correlation matrices, i.e. one which is better ("perfect") and another that does not perform as well ("imperfect"). Further to this point, the "imperfect" VAD from [32] is just one option of a realistic VAD, and one may choose from other methods such as the speech presence probability [33]. The results that ensue are only meant to give an idea of the potential range in performance one can expect from using the various algorithms depending on how well the relevant correlation matrices are estimated.

The performance of the algorithms was also evaluated using two different procedures for defining the a priori RTF vector, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, for the LMA. In the first procedure, a white noise signal of $60\,\mathrm{s}$ was played through SS1, which was subsequently used to compute a rank-1 correlation matrix per frequency. With the frontal microphone of the LMA, LM1, as the reference microphone, an EVD on the correlation matrix was performed

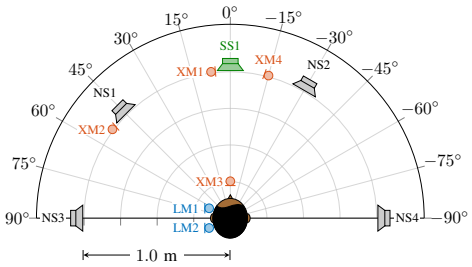

Fig. 5: Acoustic scenario illustrating the spatial distribution of the speech source (SS1), the noise sources (NS1 - NS4), the LMA (LM1, LM2), and the XMs (XM1 - XM4).

and the resulting principal eigenvector was used to define $\widetilde{\mathbf{h}}_{\mathbf{a}}$. In the following, this definition of $\widetilde{\mathbf{h}}_{\mathbf{a}}$ will be referred to as $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$. In the second procedure, only the direct path of the RTF vector was used for defining $\widetilde{\mathbf{h}}_{\mathbf{a}}$ and will be referred to as $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$. Hence $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}} = [1\; e^{-j\omega\tau_2(\theta)}]^T$, where $\tau_2(\theta)$ is the relative time delay between the two microphones of the LMA, and θ is the a priori assumed location of the source with respect to the LMA, which was 0^o in the experiments that follow.

The metrics used to evaluate the following experiments were the change in Speech Intelligibility-weighted SNR improvement [34] (Δ SI-SNR) from the input SI-SNR at LM1, and the change in short-time objective intelligibility (Δ STOI) [35] from the unprocessed speech only signal in LM1. The Δ SI-SNR was calculated as:

$$\Delta\text{SI-SNR} = \sum_i I_i \left(\text{SNR}_{\text{i,out}} - \text{SNR}_{\text{i,in}}\right) \tag{63}$$

where the band importance function I_i expresses the importance of the i-th one-third octave band with centre frequency, f_i^c for intelligibility, $\text{SNR}_{i,\text{in}}$ is the input SNR (dB), and $\text{SNR}_{i,\text{out}}$ is the output SNR (dB) in the i-th one-third octave band. The centre frequencies, f_i^c and the values for I_i are defined in [36]. The input SNR was computed accordingly using the unprocessed speech only and unprocessed noise only components in the discrete time domain at LM1, and the output SNR from the individually processed speech-only and processed noise-only components in the discrete time domain resulting from the particular algorithm. For the Δ STOI, higher values indicate an improved speech intelligibility.

In addition to the GSC algorithms, the metrics were also computed on the particular XMs used in the various acoustic scenarios. As the XM signals can be considered as speech estimates, they were also treated as separate algorithms. It should be reiterated that throughout this paper, the question being addressed was that of how XMs could be incorporated into an existing LMA-based noise reduction system, i.e. when an a priori RTF vector, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, is available. As such, in this section, the interest is only in that of a relative comparison among the GSC-LMA, the XMs themselves, and the proposed algorithms that extend the GSC-LMA with XMs. In this context, a comparison with a system where the entire RTF vector is estimated is not provided and left for future work.

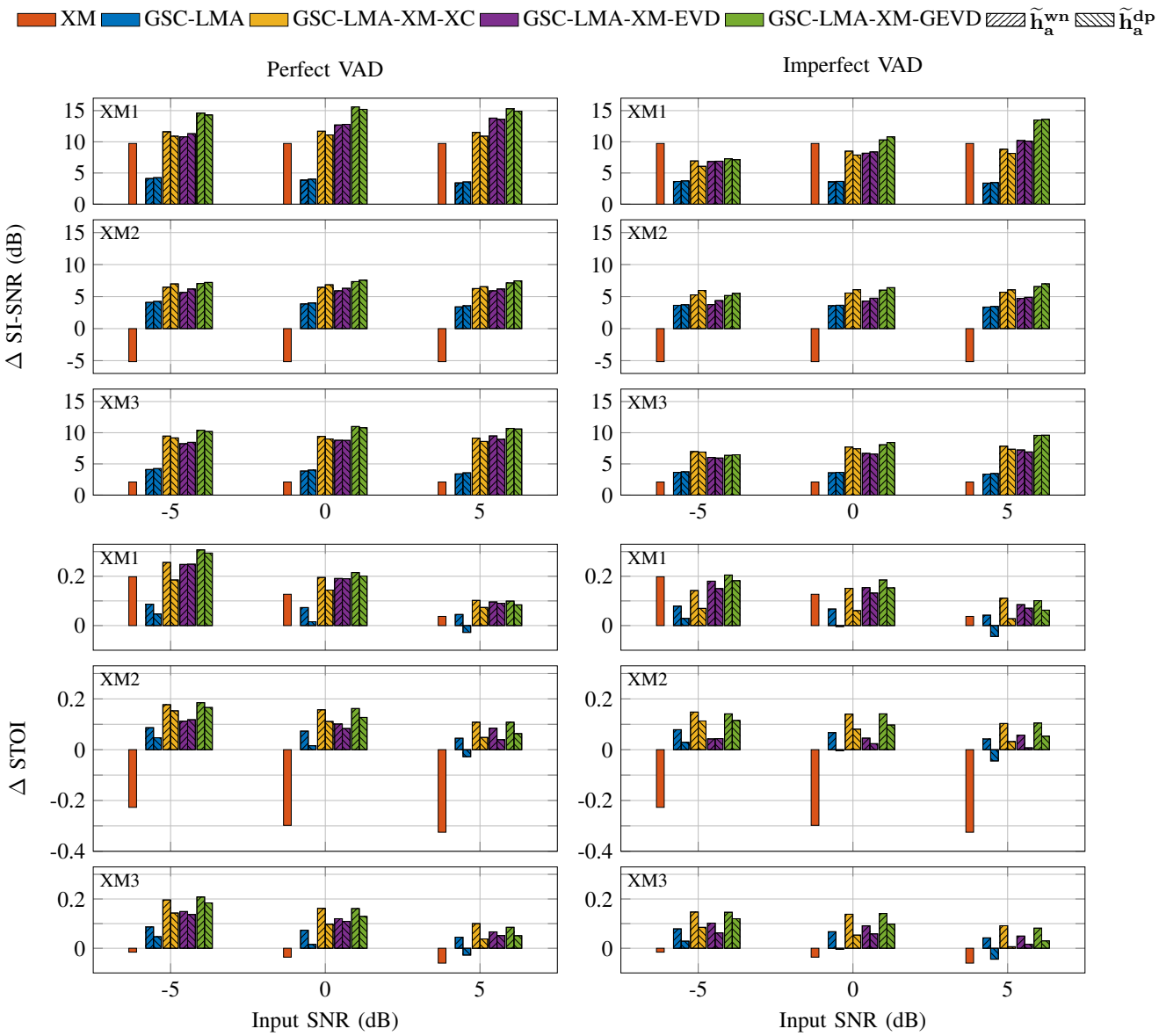

Fig. 6: Objective metrics for the various GSC algorithms, as well as for the XMs as a function of input SNR at LM1 using a perfect VAD (left) and an imperfect VAD (right). The scenario considered one male speech source (SS1), one mutli-talker babble noise source (NS1), the LMA, and one XM. Each of the three sub-plots within the particular metric uses a different XM as indicated, which corresponds to the XMs from Fig. 5. Each group of 2 bars for a particular algorithm represents the processing as performed with either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA.

A. Single XM, Single SS, Single NS

In the first experiment, only SS1 and NS1 were considered (i.e. a single correlated noise) along with the LMA and a single XM, hence 3 microphones in total. The single XM was chosen to be either XM1, XM2 or XM3 in order to observe the impact of the SNR of the XM signal on the performance of the algorithms. The input signal created was a combination of the speech signal (i.e. the male speaker from [28]) and the `bb1` noise. As these noises were recorded separately from the speech, each noise signal was also scaled such as to vary the unweighted input SNR at LM1 from $-5\,\mathrm{dB}$ to $5\,\mathrm{dB}$. This unweighted SNR (as opposed to an input SI-SNR) was computed simply from the ratio of the variance of the unprocessed speech signal to the variance of the unprocessed noise signal.

The algorithms evaluated were the GSC-LMA from section III-B, GSC-LMA-XM-CC from section V-A, GSC-LMA-XM-EVD from section V-B, and GSC-LMA-XM-GEVD from section VI-C, as well as the individual XM signals. For each of the GSC methods, the optimal filters were used for the computation of the ANC filters, i.e. (10) for $\hat{\mathbf{v}}_{\mathbf{a}}$ (in the GSC-LMA and GSC-LMA-XM-GEVD), (17) for $\hat{\mathbf{g}}$ (in the GSC-LMA-XM-CC and GSC-LMA-XM-EVD) and (35) for $\hat{\mathbf{v}}_{\mathbf{e},i}$ (in the GSC-LMA-XM-GEVD). For the

GSC-LMA-XM-GEVD, $\underline{\underline{\mathbf{w}}}$ was computed using (61) with $\underline{\underline{\mathbf{R}}}_{\mathbf{yy}}$ and $\underline{\underline{\mathbf{R}}}_{\mathbf{nn}}$ computed per frame using the same forgetting factor of $1\,\mathrm{s}$. Since the GSC-LMA-XM-EVD is the only algorithm that requires a strict definition of its fixed beamformer, i.e. $\mathbf{b_a}$ (see (25)), the fixed beamformers of all other algorithms were also set to $\mathbf{b_a}$, i.e. $\mathbf{f_a} = \mathbf{b_a}$ so as to provide an unambiguous comparison between the methods. Finally, despite the input signals having a duration of $60\,\mathrm{s}$, only the latter $30\,\mathrm{s}$ of the processed signals were used for evaluation to avoid any transient behaviour that may have resulted from the convergence of the covariance estimates in computing the filters.

Figure 6 compiles the results from this experiment when a perfect VAD and an imperfect VAD was used. Each of the three sub-plots within the particular metric uses a different XM as indicated. Each group of 2 bars for a particular algorithm represents the processing as performed with either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA. The x-axis indicates the unweighted input SNR of LM1 at which all of the algorithms were evaluated.

Focusing on the left-hand plots of Fig. 6 for the perfect VAD, in terms of Δ SI-SNR, it is clear that all of the algorithms which use the XM have an improved performance over that of the GSC-LMA as well as the XM for all input SNRs. The GSC-LMA-XM-GEVD also seems to offer a better performance than the GSC-LMA-XM-CC or GSC-LMA-XM-EVD algorithms. In terms of Δ-STOI, it is once again evident that the algorithms which use the XM have an improved performance over that of the GSC-LMA as well as the XM for all input SNRs. Among these algorithms which use the XM, however, the GSC-LMA-XM-CC and GSC-LMA-XM-GEVD have similar intelligibility improvements and are both mostly better than the GSC-LMA-XM-EVD. It is not surprising, however that the GSC-LMA-XM-EVD exhibits a poorer performance, as it is known that estimating RTFs from a subtraction of correlation matrices is prone to error at low input SNRs [4] [26]. It is also noted at the highest input SNR that the Δ-STOI values are generally smaller as the input signal would have already been fairly intelligible. In general, it can also be observed that the best improvements in terms of both metrics is obtained when XM1 is used, which is close to the source and hence subject to a high SNR. Additionally, improvements are also evident even in cases where the SNR of the XM is quite low as in the case of XM2. In such a case the XMs would have the potential to act as a noise references and hence improve performance. Finally, using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ did not demonstrate any considerable differences in performance, but using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ was in most cases slightly preferred to $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$.

Focusing on the right-hand plots of Figure 6 where an imperfect VAD was now used, as expected, the absolute values of the metrics have decreased. However, for both metrics the algorithms using an XM still demonstrate an improvement over the GSC-LMA as well as the XM except for lower input SNRs when XM1 is used. At an input SNR of $-5\,\mathrm{dB}$ for instance, the Δ SI-SNR is better than all the other algorithms, however the Δ-STOI indicates that the intelligibility is still on the order of using the GSC-LMA-XM-GEVD. Among the algorithms that use the XM, in terms of Δ SI-SNR, the GSC-LMA-XM-GEVD performs better than the GSC-LMA-XM-CC or GSC-LMA-XM-EVD particularly at higher input SNRs and when the SNR of the XM is higher, for instance for the case of XM1 or XM3. Apart from such conditions, the differences between the GSC-LMA-XM-CC and GSC-LMA-XM-GEVD are less pronounced and better than the GSC-LMA-XM-EVD. In terms of Δ STOI, the GSC-LMA-XM-CC and GSC-LMA-XM-GEVD are again similar and better than the GSC-LMA-XM-EVD. The difference in performance from using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ is now more evident, more so in terms of the intelligibility, where $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ would be preferred to $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$.

From these results of Fig. 6, the concluding point would be that the proposed algorithms that use the XM can indeed be beneficial as opposed to using only a GSC-LMA or an XM alone despite imperfect estimation of the relevant correlation matrices. Furthermore, there is the most benefit to be gained from GSC-LMA-XM-GEVD algorithm as it performs either equally well or better than the GSC-LMA-XM-CC or GSC-LMA-XM-EVD algorithms depending on the acoustic conditions.

The processed audio files from all of the algorithms, the reference, and the XM signals can be listened to for a personal subjective evaluation at [37].

B. Multiple XM combinations

In this section, the full scenario of Fig. 5 was now considered with the single speech source and the four noise sources active. The noise signals, `bb1`, `bb2`, `cc1`, and `sn1`, were used respectively for NS1, NS2, NS3, and NS4. The sum of these noises were scaled such that the unweighted input SNR at LM1 was $0\,\mathrm{dB}$. Using the optimal filters, the GSC-LMA, GSC-LMA-XM-CC, GSC-LMA-XM-EVD, and GSC-LMA-XM-GEVD algorithms were once again evaluated, however using all the possible permutations from the set of XM signals available. Hence, for the four XMs that were available, there were 15 possible XM combinations to choose from for use with the LMA.

Figure 7 displays the results of this experiment when using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA, as well as when using a perfect VAD. The x-axis indicates which set of XMs was used and is grouped by the number of XMs used in the respective algorithm. So for instance, the first four points indicate that only one XM was used, while the second set of six points indicates that two XMs were used. The numbers indicated in these groupings correspond to the XM positions as depicted in Fig. 5. The corresponding Δ SI-SNR and Δ STOI metrics for the individual XMs used are also displayed in Table I.

TABLE I: Corresponding metrics for the individual XMs used for the algorithms corresponding to Fig. 7 and Fig. 8.

	XM1	XM2	XM3	XM4
Δ SI-SNR	7.1	-1.8	1.6	3.2
Δ STOI	0.12	-0.2	-0.02	0.05

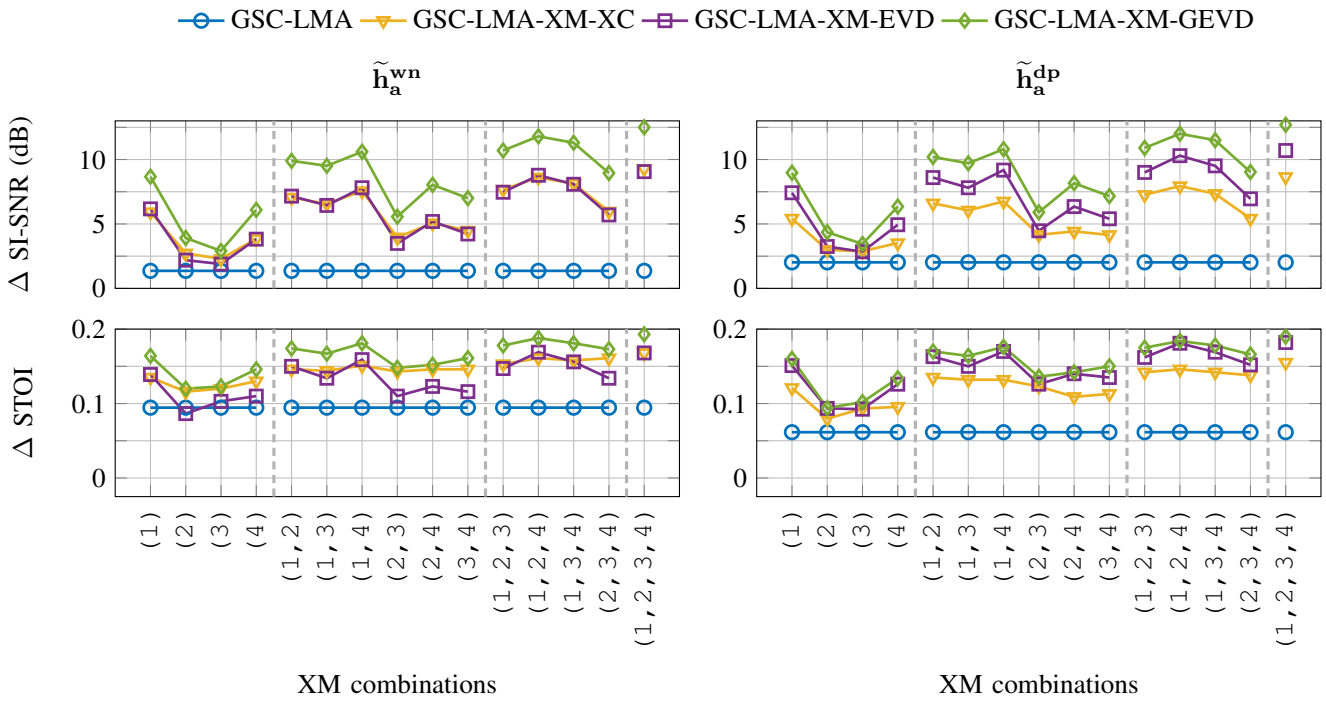

Fig. 7: Objective metrics from the acoustic scenario of Fig. 5 with one speech source and four noise sources, as a function of various combinations of XMs when using a perfect VAD.

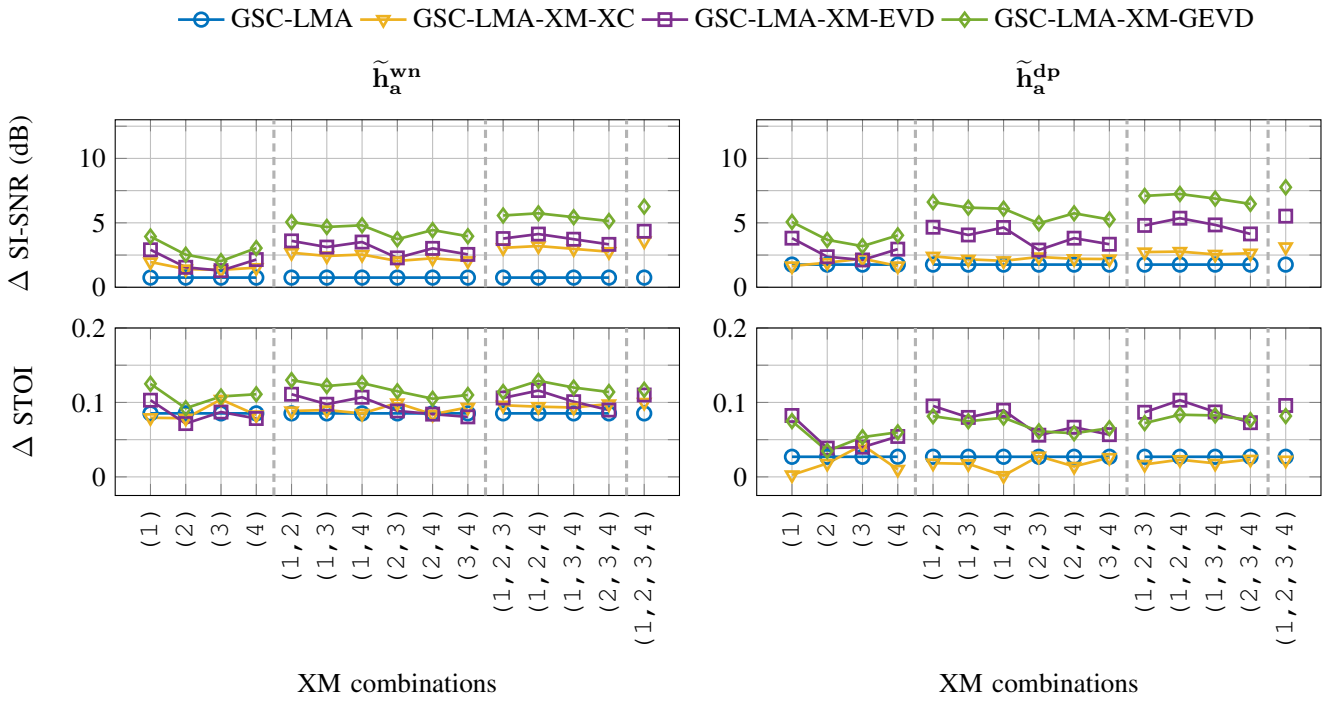

Fig. 8: Objective metrics from the acoustic scenario of Fig. 5 with one speech source and four noise sources, as a function of various combinations of XMs when using an imperfect VAD.

Focusing on the left-hand plot of Fig. 7 when $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ was used, it can be observed that the GSC-LMA-XM-GEVD outperforms the GSC-LMA, GSC-LMA-XM-CC, GSC-LMA-XM-EVD as well as any of the XMs (upon comparison with the metrics in Table I). In general, a greater improvement can be achieved with the addition of more XMs. The relative position of the XM to the speech source is also observed to have an influence, where XMs closer to the speech source are more beneficial. This can be seen by the increase in the metrics whenever XM1 is included

as one of the XMs in the respective algorithm. Such results are indeed consistent with previous findings such as in [2]. Focusing now on the right-hand plot of Fig. 7 when $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ was used, the performance of the GSC-LMA-XM-GEVD is quite similar to that when using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ and the trend is maintianed in most cases where it outperforms the GSC-LMA, GSC-LMA-XM-CC, GSC-LMA-XM-EVD as well as any of the XMs. In terms of Δ STOI, however, the GSC-LMA-XM-GEVD and GSC-LMA-XM-EVD demonstrate a similar performance. In general, the GSC-LMA-XM-CC and GSC-LMA-XM-EVD algorithms appear to be more sensitive than the GSC-LMA-XM-GEVD to the choice of either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA.

Figure 8 now displays the results of the experiment when using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ as the a priori RTF vector for the LMA, but when using an imperfect VAD. Focusing on the left-hand plot of Fig. 8 when $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ was used, the reduction in the absolute performance can immediately be seen due to the misclassification of frames where speech was present. Nevertheless, the GSC-LMA-XM-GEVD algorithm maintains its trend of outperforming the GSC-LMA, GSC-LMA-XM-CC, GSC-LMA-XM-EVD. In terms of Δ SI-SNR, however, XM1 now offers a better performance (in cases when it is used), but the intelligibility improvement as indicated by its Δ STOI is still on the order of that of the GSC-LMA-XM-GEVD algorithm. On the right-hand plot of Fig. 8 when $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$ was used, it can be observed that the Δ SI-SNR maintains a similar trend for all algorithms. In terms of intelligibility improvement however, all algorithms have a reduced performance, with GSC-LMA-XM-GEVD and GSC-LMA-XM-EVD demonstrating a similar improvement (similar to the right-hand plot of Fig. 7). Furthermore, in cases where XM1 is used, this XM may offer some intelligibility improvement over the other algorithms. However, the GSC-LMA-XM-EVD and GSC-LMA-XM-GEVD still offer intelligibility improvements over the other XMs, for instance in cases when XM2 and XM3 or XM2, XM3, and XM4 are used.

The processed audio files from all of the algorithms, the reference, and the XM signals for this experiment can be listened to for a personal subjective evaluation at [37].

C. Switching XMs

In this experiment, the impact of instantaneously switching between different XMs over time was investigated. Such a scenario can occur for instance if an XM on a mobile device is being used and then one moves the mobile device from one location to the other. As demonstrated in the previous two sections, the GSC-LMA-XM-GEVD has the potential for the best performance among all the algorithms that use the XMs. Hence, a comparison will only be considered only for this algorithm will along with the GSC-LMA (using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$), as well as the XMs in this section.

The scenario as depicted in Fig. 9 was considered. The speech and noise signals used were identical to that of section VII-B, however only two XMs were used along with the LMA.

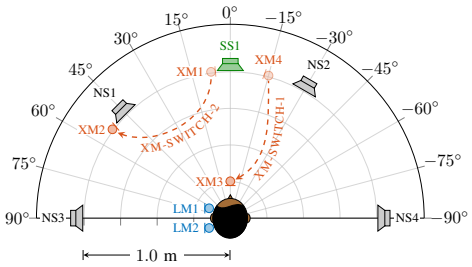

Fig. 9: Acoustic scenario analagous to that of Fig 5, except with two XMs. After 20 s, XM4 was switched to XM3 as indicated by the arrow XM-SWITCH-1, and after 40 s, XM1 was switched to XM1 as indicated by the arrow XM-SWITCH-2.

Initially both XM1 and XM4 were used. After 20 s, XM4 was switched to XM3 as indicated by the arrow XM-SWITCH-1, and after 40 s, XM1 was switched to XM2 as indicated by the arrow XM-SWITCH-2.

For both the GSC-LMA and the GSC-LMA-XM-GEVD, an NLMS procedure was used to compute the adaptive filters $\mathbf{v}_{\mathbf{a}}$ and $\mathbf{v}_{\mathbf{e},i}$. As a comparison was not done with the GSC-LMA-XM-EVD, the fixed beamformer was now set such that $\mathbf{f}_{\mathbf{a}} = \widetilde{\mathbf{h}}_{\mathbf{a}}/||\widetilde{\mathbf{h}}_{\mathbf{a}}||^2$, which is generally more of a common choice as it is a matched filter. The difference between this definition and $\mathbf{b}_{\mathbf{a}}$ is simply a scaling and does not affect the relative comparison of the various algorithms. With the input SNR at the LM1 scaled to 0 dB, the metrics of Δ SI-SNR and Δ STOI were computed over time in 3 s frames with a 50 % overlap.

Figure 10 displays the results of this experiment when using a perfect VAD (left) and an imperfect VAD (right). The uppermost plot displays the reference speech signal at LM1 with the respective VAD superimposed, while the bottom two plots display the Δ SI-SNR and Δ STOI metrics. Focusing on the left-hand plot of Fig. 10, i.e., with the perfect VAD, it can firstly be observed that at the points of switching, both of the XMs transition from having a higher SI-SNR and intelligibility improvement to a lower SI-SNR and intelligibility improvement. This transition obviously does not affect the GSC-LMA, whose SNR and intelligibility improvement remain relatively constant over time. It is also clear that the GSC-LMA-XM-GEVD results in the best performance regardless of which XMs are used for processing. Additionally, it can also be seen that using XM1 results in the most improvement (in line with the previous results) since at 40 s, when XM1 switches to XM3, the absolute values of the metrics for the GSC-LMA-XM-GEVD are reduced. In terms of the different a priori RTF vectors for the LMA, there was no significant difference between using either $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$ or $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$.

Focusing now on the right-hand plot of Fig. 10, i.e., with the imperfect VAD, the missclassification of the time periods for which speech was present can be observed from the uppermost plot. For the GSC-LMA and GSC-LMA-XM-GEVD that use $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$, it can be seen that the GSC-LMA-XM-GEVD suffers a reduction in performance in both SI-SNR and STOI improvement, but still performs better than the GSC-LMA in most cases. During the first 40 s, the GSC-LMA-XM-GEVD also

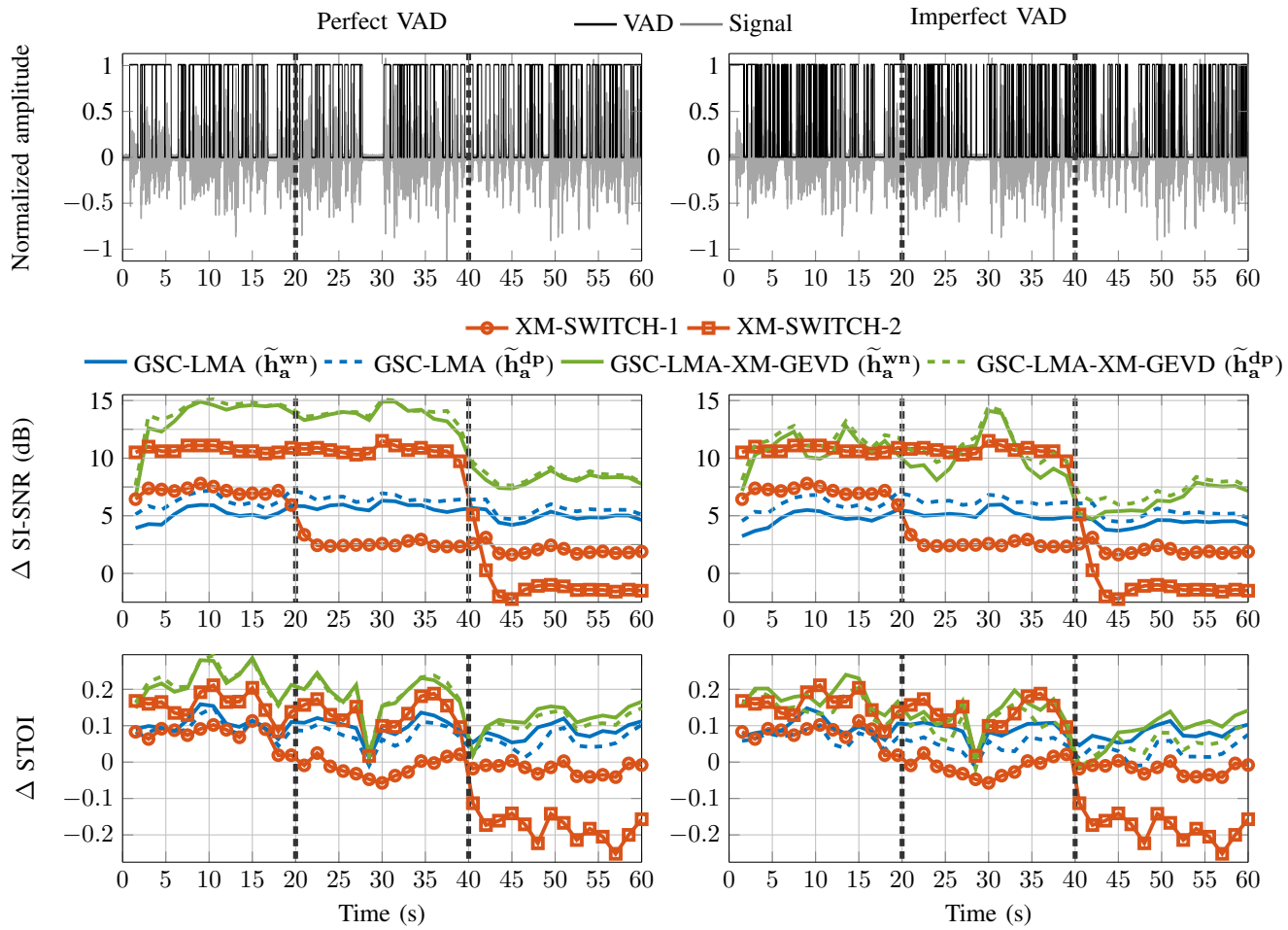

Fig. 10: Performance of the algorithms when switching between different XMs over time using a perfect VAD (left) and an imperfect VAD (right). The uppermost plots display the clean speech reference signal in LM1 with the respective VAD superimposed. The middle plots show the Δ SI-SNR metric and the final row of plots show the Δ STOI metric. Both Δ SI-SNR and Δ STOI metrics were computed with 3 s frames with a 50 % overlap. The markers for the XMs correspond to the middle of these time frames. Markers were not used for the other plots for clarity. The vertical dotted line at 20 s indicate the switching of XMs from XM-SWITCH-1 and the vertical dotted line at 40 s indicate the switching of XMs from XM-SWITCH-2.

exhibits a performance on the order of XM1, being slightly better or slightly worse at times. However, when any of the other XMs are used, the GSC-LMA-XM-GEVD maintains a better performance. When using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{dp}}$, the relative trends among the algorithms and the XMs are maintained in comparison with using $\widetilde{\mathbf{h}}_{\mathbf{a}}^{\mathbf{wn}}$, except that the absolute intelligibility has been reduced.

The processed audio files from all of the algorithms, the reference, and the XM signals for this experiment can be listened to for a personal subjective evaluation at [37].

VIII. Conclusions

Three methods of extending a local-microphone-array-based Generalised Sidelobe Canceller (GSC-LMA) with external microphones (XMs) have been presented. These methods have considered the relative transfer functions (RTF) for the GSC-LMA as a priori knowledge, upon which the complementary RTFs for the XMs could be estimated. Such an approach has intended to preserve the reliability of an existing GSC-LMA, while including the XMs as "add-ons" that could improve the performance.

Two of the methods presented, the GSC-LMA-XM-CC and the GSC-LMA-XM-EVD, involved a procedure for completing an extended blocking matrix using either a cross-correlation or an eigenvalue decomposition (EVD) respectively. The third method, GSC-LMA-XM-GEVD, proposed an alternative approach, where the speech estimate from an GSC-LMA is directly used along with an orthogonalised version of the XM signals to obtain an improved speech estimate via a generalised eigenvalue decomposition (GEVD). When a perfect VAD was used to estimate the relevant correlation matrices, it was found that all of these methods offered an improvement over the GSC-LMA, with the GSC-LMA-XM-GEVD having the best performance. In cases of imperfect estimation of the relevant correlation matrices, the performance of these algorithms was inevitably reduced, but the GSC-LMA-XM-GEVD continued to maintain its performance above the GSC-LMA and the XMs, unless the XMs

were very close to the speech source. As it is not expected that the XM will always be close to the speech source, using the GSC-LMA-XM-GEVD would then be the preferable option for a consistent performance when the XMs are subject to movement.

One final point to re-iterate is that the speech estimate from the GSC-LMA has not been compromised in any way (as illustrated in Fig 4) and if so desired, this signal is still available as an option to the listener. In extreme cases of poor estimation with the XMs, the GSC-LMA-XM-GEVD offers a contingency option of simply using the LMA-based solution (which was what would be used in the absence of XMs), and hence the XMs can be truly be treated in a modular fashion or as "add-ons".

REFERENCES

[1] M. Brandstein and D. 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 J. Adv. Signal Process.*, vol. 2009, 2009.

[3] A. Bertrand, S. Doclo, S. Gannot, N. Ono, and T. van Waterschoot, "Special issue on wireless acoustic sensor networks and ad hoc microphone arrays," *Signal Processing*, vol. 107, pp. 1–3, 2015.

[4] A. Hassani, "Distributed signal processing algorithms for multi-task wireless acoustic sensor networks," Ph.D. dissertation, KU Leuven, Oct. 2017.

[5] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 2, pp. 343–356, 2013.

[6] A. I. Koutrouvelis, T. W. Sherson, R. Heusdens, and R. C. Hendriks, "A Low-Cost Robust Distributed Linearly Constrained Beamformer for Wireless Acoustic Sensor Networks with Arbitrary Topology," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 8, pp. 1434–1448, 2018.

[7] A. Boothroyd, "Hearing aid accessories for adults: the remote FM microphone." *Ear and Hearing*, vol. 25, no. 1, pp. 22–33, 2004.

[8] E. C. Schafer, K. Sanders, D. Bryant, K. Keeney, and N. Baldus, "Effects of Voice Priority in FM Systems for Children with Hearing Aids," *J. Educ. Audiol.*, vol. 19, pp. 12–24, 2013.

[9] 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.

[10] 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.

[11] N. Gößling and S. Doclo, "RTF-based binaural MVDR beamformer exploiting an external microphone in a diffuse noise field," in *Proc. ITG Conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 1–5.

[12] 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.

[13] 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.

[14] 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.

[15] M. Farmani, M. S. Pedersen, Z.-H. Tan, and J. Jensen, "Informed Sound Source Localization Using Relative Transfer Functions for Hearing Aid Applications," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 3, pp. 611–623, 2017.

[16] G. Courtois, "Spatial hearing rendering in wireless microphone systems for binaural hearing aids," Ph.D. dissertation, École polytechnique fédérale de Lausanne (EPFL), Lausanne, 2016.

[17] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[18] 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.

[19] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[20] N. Yousefian and P. Loizou, "A Dual-Microphone Speech Enhancement Algorithm Based on the Coherence Function," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 2, pp. 599–609, 2011.

[21] 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.

[22] 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.

[23] R.Ali, T. van Waterschoot, and M. Moonen, "Completing the RTF vector for an MVDR beamformer as applied to a local microphone array and an external microphone," in *Proc. 2018 Int. Workshop Acoustic Signal Enhancement (IWAENC '18)*, Tokyo, Japan, Sept 2018, pp. 1–4.

[24] ——, "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.

[25] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 2013.

[26] 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.

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

[28] Bang and Olufsen, "Music for Archimedes," CD B&O 101, 1992.

[29] Auditec, "Auditory Tests (Revised), Compact Disc, Auditec, St. Louis," St. Louis, 1997.

[30] BBC Sound Effects data for Research and Education Space, 16,000 sound effects and field recordings, 2018. [Online]. Available: http://bbcsfx.acropolis.org.uk/assets/07070195.wav

[31] 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.

[32] M. Brookes *et al.* (1997) Voicebox: Speech processing toolbox for matlab. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox

[33] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. 2011 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '11)*, Oct 2011, pp. 145–148.

[34] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance." *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, 1993.

[35] 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.

[36] ANSI-S3.5-1997, "American National Standard Methods for calculation of the speech intelligibility index," *J. Acoust. Soc. Amer.*, June 1997.

[37] (2018). [Online]. Available: http://homes.esat.kuleuven.be/~ali/meth_extend_gsc_xms.html

Randall Ali received a B.Sc. degree in electrical and computer engineering from the University of the West Indies in 2007 and an M.S. degree in acoustics from the Pennsylvania State University in 2013. He is currently working toward the Ph.D. degree in electrical engineering at KU Leuven, where his research is focused on speech enhancement strategies for hearing assistive devices.

Giuliano Bernardi (S'12) was born in Asolo, Italy in 1987. He received the M.Sc. in Engineering Acoustics from Denmark Technical University (DTU), Denmark, in 2011, the MEng in Bioengineering from University of Padua, Padua, Italy, in 2012, and the Ph.D. in engineering science from KU Leuven, Belgium, in 2018. Currently, he is a Postdoctoral Researcher at KU Leuven focusing on acoustic feedback control, echo cancellation, and noise reduction strategies for hearing-aid applications and voice communication systems. His research interests include audio signal processing, speech perception and psychoacoustics, real-time audio signal processing, and speech enhancement.

Toon van Waterschoot (S'04, M'12) received MSc (2001) and PhD (2009) degrees in Electrical Engineering, both from KU Leuven, Belgium, where he is currently an Associate Professor and Consolidator Grantee of the European Research Council (ERC). He has previously also held teaching and research positions at Delft University of Technology in The Netherlands and the University of Lugano in Switzerland. His research interests are in signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction.

He has been serving as an Associate Editor for the Journal of the Audio Engineering Society and for the EURASIP Journal on Audio, Music, and Speech Processing, and as a Guest Editor for Elsevier Signal Processing. He is a Director of the European Association for Signal Processing (EURASIP), a Member of the IEEE Audio and Acoustic Signal Processing Technical Committee, a Member of the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and a Founding Member of the EAA Technical Committee in Audio Signal Processing. He was the General Chair of the 60th AES International Conference in Leuven, Belgium (2016), and has been serving on the Organizing Committee of the European Conference on Computational Optimization (EUCCO 2016) and the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017). He is a member of EURASIP, IEEE, ASA, and AES.

Marc Moonen (M'94, SM'06, F'07) is a Full Professor at the Electrical Engineering Department of KU Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing. He is a Fellow of the IEEE (2007) and a Fellow of EURASIP (2018). He received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He received journal best paper awards from the IEEE Transactions on Signal Processing (with Geert Leus and with Daniele Giacobello) and from Elsevier Signal Processing (with Simon Doclo). He was chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and President of EURASIP (European Association for Signal Processing, 2007-2008 and 2011-2012). He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003-2005), Area Editor for Feature Articles in IEEE Signal Processing Magazine (2012-2014), and has been a member of the editorial board of Signal Processing, IEEE Transactions on Circuits and Systems II, IEEE Signal Processing Magazine, Integration-the VLSI Journal, EURASIP Journal on Wireless Communications and Networking and EURASIP Journal on Advances in Signal Processing.