RESEARCH

An integrated MVDR beamformer for speech enhancement using a local microphone array and external microphones

Randall Ali*, Toon van Waterschoot and Marc Moonen

Abstract

An integrated version of the minimum variance distortionless response (MVDR) beamformer for speech enhancement using a microphone array has been recently developed, which merges the benefits of imposing constraints defined from both a relative transfer function (RTF) vector based on a priori knowledge and an RTF vector based on a data-dependent estimate. In this paper, the integrated MVDR beamformer is extended for use with a microphone configuration where a microphone array, local to a speech processing device, has access to the signals from multiple external microphones (XMs) randomly located in the acoustic environment. The integrated MVDR beamformer is reformulated as a quadratically constrained quadratic program (QCQP) with two constraints, one of which is related to the maximum tolerable speech distortion for the imposition of the a priori RTF vector and the other related to the maximum tolerable speech distortion for the imposition of the data-dependent RTF vector. An analysis of how these maximum tolerable speech distortions affect the behaviour of the QCQP is presented, followed by the discussion of a general tuning framework. The integrated MVDR beamformer is then evaluated with audio recordings from behind-the-ear hearing aid microphones and three XMs for a single desired speech source in a noisy environment. In comparison to relying solely on an a priori RTF vector or a data-dependent RTF vector, the results demonstrate that the integrated MVDR beamformer can be tuned to yield different enhanced speech signals, which may be more suitable for improving speech intelligibility despite changes in the desired speech source position and imperfectly estimated spatial correlation matrices.

Keywords: Speech Enhancement; Beamforming; Minimum Variance Distortionless Response (MVDR) beamformer; External Microphones

1 Introduction

Speech processing devices such as a hearing aid, a cochlear implant, or a mobile telephone are commonly equipped with an array of microphones to capture the acoustic environment. The received microphone signals are often a mixture of a desired speech signal plus some undesired noise (any combination of interfering speakers, background noises, and reverberation). As the quality and intelligibility of the desired speech signal is susceptible to considerable degradation in the presence of such noise, the task of suppressing this noise and extracting the desired speech signal, known as speech enhancement, is of critical importance and has been the subject of extensive research [1–3].

*Correspondence: randall.ali@esat.kuleuven.be
KU Leuven, Dept. of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
Full list of author information is available at the end of the article

While successful speech enhancement strategies have been developed with microphone arrays, in some applications, due to physical space constraints, the spatial variation between the observed microphone signals may not be sufficient to yield an acceptable degree of speech enhancement. Consequently, the potential using more ad hoc microphone configurations consisting of randomly placed microphones to increase the spatial sampling of the acoustic environment has developed interest [4–12]. In this paper, a specific ad hoc microphone configuration is considered, where a microphone array located on some speech processing device, hereafter referred to as a local microphone array (LMA), is linked with multiple remote or external microphones (XMs) in a centralised processing framework, i.e. all microphone signals are sent to a fusion centre for processing. The terminology of a *local* microphone array is introduced since the microphone array is considered

to be confined or fixed within some area of the acoustic environment relative to the XMs which are subject to movement.

When there is a single desired speech source, speech enhancement can be accomplished by using the minimum variance distortionless response (MVDR) beamformer [13, 14]. One of the important quantities required for computing the MVDR beamformer is a vector of acoustic transfer functions from the desired speech source to all of the microphones. More commonly, however, a vector of relative transfer functions (RTFs) is used instead, which is a normalised version of the acoustic transfer function vector with respect to some reference microphone [15]. In practice, for an LMA, this RTF vector may be measured a priori or based on assumptions regarding microphone characteristics, position, speaker location and room acoustics (e.g. no reverberation). For instance, in assistive hearing devices, it is sometimes assumed that the desired speech source location is known and this knowledge can be subsequently used to define an a priori RTF vector [16–19]. Alternatively, it may be estimated in an online fashion from the observed microphone data [20, 21] so that it is a fully data-dependent estimate.

The situation under consideration throughout this paper is one in which there is an available a priori RTF vector pertaining only to the LMA that may or may not be sufficiently accurate with respect to the true RTF vector. In cases where the a priori RTF vector is not sufficiently accurate, then incorporating the use of a data-dependent RTF vector can be viewed as an opportunity for an improved performance provided that the data-dependent RTF vector is a better estimate of the true RTF vector. On the other hand, when acoustic conditions are adverse enough to significantly affect the accuracy of the data-dependent RTF vector, then relying on the a priori RTF vector can be viewed as a fall back or contingency strategy.

It would therefore be seemingly advantageous to use both an a priori and a data-dependent RTF vector in practice. Such an approach has recently been investigated for an LMA only and resulted in an integrated version of the MVDR beamformer [22]. As opposed to imposing either the a priori RTF vector or the data-dependent RTF vector as a hard constraint, they were both softened into an unconstrained optimisation problem. It was demonstrated that the resulting integrated MVDR beamformer is a convex combination of an MVDR beamformer that uses the a priori RTF vector, an MVDR beamformer that uses the data-dependent RTF vector, a linearly constrained minimum variance (LCMV) beamformer that uses both the a priori and data-dependent RTF vector, and an all-zero vector, each with real-valued weightings, revealing the versatile nature of such an integrated beamformer.

This paper therefore re-examines the integrated MVDR beamformer for the ad hoc microphone configuration consisting of an LMA located on some speech processing device linked with multiple XMs. Specifically, the integrated MVDR beamformer is reformulated from an alternative perspective, namely that of a quadratically constrained quadratic program (QCQP). This QCQP will consist of two constraints, one of which is related to the maximum tolerable speech distortion for the imposition of the a priori RTF vector and the other related to the maximum tolerable speech distortion for the imposition of the data-dependent RTF vector. With respect to the procedures for obtaining the RTF vectors, it is straightforward to obtain a data-dependent RTF vector, however, the notion of an a priori RTF vector when XMs are used with an LMA is a bit more ambiguous. In particular, since only partial a priori knowledge is usually available for the part of the RTF vector pertaining to the LMA, the other part pertaining to the XMs will have to be a data-dependent estimate and hence a procedure based on partial a priori knowledge [9] would be necessary. As a result, an integrated MVDR beamformer for a microphone configuration with an LMA and XMs will merge an a priori RTF vector that is based on partial a priori knowledge and a fully data-dependent one.

With the a priori and the data-dependent RTF vector for the LMA and XMs estimated, it will become evident that the optimal filter from the integrated MVDR beamformer, formulated as a QCQP, is identical to that which was derived from [22], where the Lagrangian multipliers associated with the QCQP are equivalent to the tuning parameters that have been considered in [22]. The additional insight of the QCQP formulation is that these tuning parameters or Lagrangian multipliers can be related to a maximum tolerable speech distortion for the imposition of the a priori or the data-dependent RTF vector. An analysis of this relationship is provided, which facilitates the tuning of the integrated MVDR beamformer from the more intuitive perspective of the maximum tolerable speech distortions as opposed to the combination of filters as in [22]. A general tuning framework will then be discussed along with the suggestion of some particular tuning strategies.

The integrated MVDR beamformer is then evaluated with audio recordings from behind-the-ear hearing aid microphones (the LMA) and three XMs for a single desired speech source in a a re-created cocktail-party scenario. The results demonstrate that the integrated MVDR beamformer can be tuned to yield different

enhanced speech signals, which can find a compromise between relying solely on an a priori RTF vector or a data-dependent RTF vector, and hence may be more suitable for improving speech intelligibility despite changes in the desired speech source position and imperfectly estimated spatial correlation matrices.

The paper is organised as follows. In section 2, the data model is defined. In section 3, the MVDR beamformer as applied to an LMA with XMs is discussed along with the procedures for obtaining the a priori RTF vector based on partial a priori knowledge and the data-dependent RTF vector. Section 4 reformulates the integrated MVDR beamformer as a QCQP and provides an analysis on the effect of the maximum tolerable speech distortions due to the imposition of the a priori RTF vector and the data-dependent RTF vector. In section 5 a general tuning framework is presented, as well as some suggested tuning strategies. In section 6, the integrated MVDR approach is analysed and evaluated with both simulated data as well as experimental data involving the use of behind-the-ear hearing aid microphones and three XMs. Conclusions are then drawn in section 7.

2 Data Model

2.1 Unprocessed Signals

A microphone configuration consisting of an LMA of M_a microphones plus M_e XMs is considered with one desired speech source in a noisy, reverberant[1] environment. In the short-time Fourier transform (STFT) domain, the observed vector of microphone signals at frequency bin k and 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) $\mathbf{y} = [\mathbf{y}_\mathbf{a}^T\ \mathbf{y}_\mathbf{e}^T]^T$, $\mathbf{y}_\mathbf{a} = [\mathrm{y}_{\mathrm{a},1}\ \mathrm{y}_{\mathrm{a},2}\ \cdots \mathrm{y}_{\mathrm{a},M_\mathrm{a}}]^T$ is a vector of the LMA signals, $\mathbf{y}_\mathbf{e} = [\mathrm{y}_{\mathrm{e},1}\ \mathrm{y}_{\mathrm{e},2}\ \cdots \mathrm{y}_{\mathrm{e},M_\mathrm{e}}]^T$ is a vector of the XM signals, $\mathbf{x}$ is the speech contribution, represented by $\mathrm{s}_{\mathrm{a},1}$, the desired speech signal in the first (reference) microphone of the LMA, filtered with $\mathbf{h} = [\mathbf{h}_\mathbf{a}^T\ \mathbf{h}_\mathbf{e}^T]^T$, $\mathbf{h}_\mathbf{a}$ is the RTF vector for the LMA (where the first component of $\mathbf{h}_\mathbf{a}$ is equal to 1 since the first microphone is used as the reference), and $\mathbf{h}_\mathbf{e}$ is the RTF vector for the XM signals. Finally, $\mathbf{n} = [\mathbf{n}_\mathbf{a}^T\ \mathbf{n}_\mathbf{e}^T]^T$ 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.

[1]Reverberation is not explicitly included in the signal model as dereverberation is not addressed in this paper. This paper primarily focuses on noise reduction, although some dereverberation will be achieved as a fortunate by-product of beamforming.

The $(M_\mathrm{a}+M_\mathrm{e})\times(M_\mathrm{a}+M_\mathrm{e})$ spatial correlation matrix for the speech-plus-noise, noise-only, and speech-only signals is defined respectively as:

$$\mathbf{R}_{\mathbf{yy}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} \tag{2}$$

$$\mathbf{R}_{\mathbf{nn}} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\} \tag{3}$$

$$\mathbf{R}_{\mathbf{xx}} = \mathbb{E}\{\mathbf{x}\mathbf{x}^H\} \tag{4}$$

where $\mathbb{E}\{.\}$ is the expectation operator and $\{.\}^H$ is the Hermitian transpose. With the assumption of a single desired speech source from (1), $\mathbf{R}_{\mathbf{xx}}$ can be represented as a rank-1 correlation matrix as follows:

$$\mathbf{R}_{\mathbf{xx}} = \sigma^2_{\mathrm{s}_{\mathrm{a},1}}\mathbf{h}\mathbf{h}^H \tag{5}$$

where $\sigma^2_{\mathrm{s}_{\mathrm{a},1}} = \mathbb{E}\{|\mathrm{s}_{\mathrm{a},1}|^2\}$ is the desired speech power spectral density in the first microphone of the LMA. It is further assumed that the desired speech signal is uncorrelated with the noise signal, and hence $\mathbf{R}_{\mathbf{yy}} = \mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{nn}}$. The speech-plus-noise, noise-only, and speech-only correlation matrix can also be defined solely for the LMA signals respectively as $\mathbf{R}_{\mathbf{y_a y_a}} = \mathbb{E}\{\mathbf{y_a}\mathbf{y_a}^H\}$, $\mathbf{R}_{\mathbf{n_a n_a}} = \mathbb{E}\{\mathbf{n_a}\mathbf{n_a}^H\}$, and $\mathbf{R}_{\mathbf{x_a x_a}} = \mathbb{E}\{\mathbf{x_a}\mathbf{x_a}^H\}$, with $\mathbf{R}_{\mathbf{x_a x_a}}$ also having the same rank-1 structure as in (5). 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 XMs with no bandwidth constraints as well as synchronous sampling rates.

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

$$\mathrm{z}_1 = \mathbf{w}^H\mathbf{y} \tag{6}$$

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

2.2 Pre-whitened-transformed domain

As a pre-processing stage, the unprocessed microphone signals can be firstly transformed with the available a priori RTF vector for the LMA signals and then spatially pre-whitened using the resulting transformed noise-only correlation matrix, yielding a vector of pre-whitened-transformed (PWT) microphone signals. As discussed in [9] and subsequently reviewed in section 3.1, these pre-processing steps essentially compress the M_a LMA signals into one signal. This signal is then used with the pre-processed M_e XM signals to obtain an estimate for the missing part of the RTF vector pertaining to the XMs when there is an available a priori RTF vector for the LMA. Therefore, PWT microphone signals will be adopted for convenience throughout this paper.

To define the transformation operation, an $M_\mathrm{a} \times (M_\mathrm{a}-1)$ blocking matrix $\widetilde{\mathbf{C}}_\mathbf{a}$, and an $M_\mathrm{a} \times 1$ fixed beamformer, $\widetilde{\mathbf{f}}_\mathbf{a}$, are firstly defined such that:

$$\widetilde{\mathbf{C}}_\mathbf{a}^H \widetilde{\mathbf{h}}_\mathbf{a} = \mathbf{0}; \quad \widetilde{\mathbf{f}}_\mathbf{a}^H \widetilde{\mathbf{h}}_\mathbf{a} = 1 \tag{7}$$

where $\widetilde{\mathbf{h}}_\mathbf{a}$ is an available a priori RTF vector (which is some pre-determined estimate or approximation of $\mathbf{h}_\mathbf{a}$), and the notation $\widetilde{(\cdot)}$ refers to quantities based on available a priori knowledge. Using $\widetilde{\mathbf{C}}_\mathbf{a}$ and $\widetilde{\mathbf{f}}_\mathbf{a}$, an $(M_\mathrm{a}+M_\mathrm{e}) \times (M_\mathrm{a}+M_\mathrm{e})$ transformation matrix, $\widetilde{\mathbf{\Upsilon}}$, can be defined as:

$$\widetilde{\mathbf{\Upsilon}} = \left[\begin{array}{c|c} \widetilde{\mathbf{\Upsilon}}_\mathbf{a} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_{M_\mathrm{e}} \end{array}\right] = \left[\begin{array}{c|c} [\widetilde{\mathbf{C}}_\mathbf{a} \;\; \widetilde{\mathbf{f}}_\mathbf{a}] & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I}_{M_\mathrm{e}} \end{array}\right] \tag{8}$$

where $\widetilde{\mathbf{\Upsilon}}_\mathbf{a} = [\widetilde{\mathbf{C}}_\mathbf{a} \;\; \widetilde{\mathbf{f}}_\mathbf{a}]$ and in general $\mathbf{I}_\vartheta$ denotes the $\vartheta \times \vartheta$ identity matrix. Consequently, the transformed speech-plus-noise signals and the transformed noise-only signals are defined respectively as:

$$\widetilde{\mathbf{\Upsilon}}^H \mathbf{y} = \begin{bmatrix} \widetilde{\mathbf{C}}_\mathbf{a}^H \mathbf{y}_\mathbf{a} \\ \widetilde{\mathbf{f}}_\mathbf{a}^H \mathbf{y}_\mathbf{a} \\ \mathbf{y}_\mathbf{e} \end{bmatrix}; \quad \widetilde{\mathbf{\Upsilon}}^H \mathbf{n} = \begin{bmatrix} \widetilde{\mathbf{C}}_\mathbf{a}^H \mathbf{n}_\mathbf{a} \\ \widetilde{\mathbf{f}}_\mathbf{a}^H \mathbf{n}_\mathbf{a} \\ \mathbf{n}_\mathbf{e} \end{bmatrix} \tag{9}$$

This transformation domain is simply the LMA signals that pass through a blocking matrix and a fixed beamformer as is done in the first stage of a typical generalized sidelobe canceller (i.e. the adaptive implementation of an MVDR beamformer) [23], along with the unprocessed XM signals.

A spatial pre-whitening operation can now be defined from the transformed noise-only correlation matrix by using the Cholesky decomposition:

$$\mathbb{E}\{(\widetilde{\mathbf{\Upsilon}}^H \mathbf{n})(\widetilde{\mathbf{\Upsilon}}^H \mathbf{n})^H\} = \mathbf{L}\mathbf{L}^H \tag{10}$$

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

A transformed signal vector can then be pre-whitened by pre-multiplying it with $\mathbf{L}^{-1}$ and will be denoted with an underbar $\underline{(\cdot)}$. Hence the signal model for the unprocessed microphone signals from (1) can be expressed in the PWT domain as[2]:

$$\underline{\mathbf{y}}(k,l) = \mathbf{L}^{-1}(k,l)\widetilde{\mathbf{\Upsilon}}^H(k,l)\mathbf{y}(k,l) \tag{11}$$

$$= \underbrace{\underline{\mathbf{h}}(k,l)\mathrm{s}_{\mathrm{a},1}(k,l)}_{\underline{\mathbf{x}}(k,l)} + \underline{\mathbf{n}}(k,l) \tag{12}$$

where $\underline{\mathbf{y}}$ consists of the PWT LMA and XM signals, i.e. $\underline{\mathbf{y}} = \left[\underline{\mathbf{y}}_\mathbf{a}^T \; \underline{\mathbf{y}}_\mathbf{e}^T\right]^T$, $\underline{\mathbf{n}} = \mathbf{L}^{-1}\widetilde{\mathbf{\Upsilon}}^H\mathbf{n}$, the PWT RTF vector $\underline{\mathbf{h}} = \mathbf{L}^{-1}\widetilde{\mathbf{\Upsilon}}^H\mathbf{h}$, and the respective correlation matrices are:

$$\underline{\mathbf{R}}_{\mathbf{yy}} = \mathbb{E}\{\underline{\mathbf{y}}\,\underline{\mathbf{y}}^H\} = \mathbf{L}^{-1}\widetilde{\mathbf{\Upsilon}}^H\mathbf{R}_{\mathbf{yy}}\widetilde{\mathbf{\Upsilon}}\mathbf{L}^{-H} \tag{13}$$

$$\underline{\mathbf{R}}_{\mathbf{nn}} = \mathbb{E}\{\underline{\mathbf{n}}\,\underline{\mathbf{n}}^H\} = \mathbf{L}^{-1}\widetilde{\mathbf{\Upsilon}}^H\mathbf{R}_{\mathbf{nn}}\widetilde{\mathbf{\Upsilon}}\mathbf{L}^{-H} = \mathbf{I}_{(M_\mathrm{a}+M_\mathrm{e})} \tag{14}$$

$$\underline{\mathbf{R}}_{\mathbf{xx}} = \mathbb{E}\{\underline{\mathbf{x}}\,\underline{\mathbf{x}}^H\} = \sigma^2_{\mathrm{s}_{\mathrm{a},1}}\underline{\mathbf{h}}\,\underline{\mathbf{h}}^H \tag{15}$$

where the expression for $\underline{\mathbf{R}}_{\mathbf{nn}}$ is a direct consequence of (10). With the assumption of the desired speech source and noise being uncorrelated, it also holds that $\underline{\mathbf{R}}_{\mathbf{yy}} = \underline{\mathbf{R}}_{\mathbf{xx}} + \underline{\mathbf{R}}_{\mathbf{nn}}$. In the PWT domain, the estimate of the desired speech signal in the first microphone of the LMA, z_1, which is equivalent to (6), is then obtained through a linear filtering of the PWT microphone signals, such that:

$$\mathrm{z}_1 = \underset{\sim}{\mathbf{w}}^H \underline{\mathbf{y}} \tag{16}$$

where $\underset{\sim}{\mathbf{w}} = \mathbf{L}^H\widetilde{\mathbf{\Upsilon}}^{-1}\mathbf{w}$ is a complex-valued filter[3].

3 MVDR with an LMA and XMs

The MVDR beamformer minimises the noise power spectral density after filtering (minimum variance), with a constraint that the desired speech signal should not be subject to any distortion (distortionless response), which is specified by an appropriate RTF vector for the MVDR beamformer. For the unprocessed microphone signals, the MVDR beamformer problem can be formulated as:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \mathbf{w}^H\mathbf{R}_{\mathbf{nn}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^H\mathbf{h} = 1 \end{aligned} \tag{17}$$

The solution to (17) yields the optimal filter:

$$\mathbf{w}_\mathrm{mvdr} = \frac{\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{R}_{\mathbf{nn}}^{-1}\mathbf{h}} \tag{18}$$

with the desired speech signal estimate, $\mathrm{z}_1 = \mathbf{w}_\mathrm{mvdr}^H\mathbf{y}$. In practice, both $\mathbf{R}_{\mathbf{nn}}$ and $\mathbf{h}$ are unknown and hence must be estimated.

A data-dependent estimate can typically be obtained for $\mathbf{R}_{\mathbf{nn}}$, for instance by recursive averaging, with a voice activity detector [24] or a speech presence probability (SPP) estimator [25]. This data-dependent estimate will be denoted as $\hat{\mathbf{R}}_{\mathbf{nn}}$ and in general the notation $\hat{(\cdot)}$ will refer to any data-dependent estimate.

[2] The dependence on k and l is included here as a reminder and for completeness in the signal model. It will be dropped again unless explicitly required.

[3] Since the sequence of operations from $\mathbf{w}$ to $\underset{\sim}{\mathbf{w}}$ is not exactly that of a PWT signal vector, a slightly different notation is used for this quantity.

In the PWT domain, it can be seen that using $\hat{\mathbf{R}}_{\mathbf{nn}}$ in (10) will result in an estimate for the pre-whitening operator as $\hat{\mathbf{L}}$ and hence from (14), $\hat{\mathbf{R}}_{\mathbf{nn}}$ can be expressed as:

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \widetilde{\boldsymbol{\Upsilon}}^{-H}\hat{\mathbf{L}}\hat{\mathbf{L}}^{H}\widetilde{\boldsymbol{\Upsilon}}^{-1} \tag{19}$$

Replacing $\mathbf{R}_{\mathbf{nn}}$ in (17) with $\hat{\mathbf{R}}_{\mathbf{nn}}$ in (19) then results in the MVDR beamformer problem formulated in the PWT domain as:

$$\begin{aligned} \underset{\underset{\sim}{\mathbf{w}}}{\text{minimise}} \quad & \underset{\sim}{\mathbf{w}}^{H}\underset{\sim}{\mathbf{w}} \\ \text{s.t.} \quad & \underset{\sim}{\mathbf{w}}^{H}\underline{\mathbf{h}} = 1 \end{aligned} \tag{20}$$

where $\underset{\sim}{\mathbf{w}}$ is redefined as $\underset{\sim}{\mathbf{w}} = \hat{\mathbf{L}}^{H}\widetilde{\boldsymbol{\Upsilon}}^{-1}\mathbf{w}$ and $\underline{\mathbf{h}}$ is redefined as $\underline{\mathbf{h}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^{H}\mathbf{h}$. The solution to (20) then yields the optimal filter in the PWT domain:

$$\underset{\sim}{\mathbf{w}}_{\text{mvdr}} = \frac{\underline{\mathbf{h}}}{\underline{\mathbf{h}}^{H}\,\underline{\mathbf{h}}} \tag{21}$$

with the desired speech signal estimate, $\mathrm{z}_1 = \underset{\sim}{\mathbf{w}}_{\text{mvdr}}^{H}\underline{\mathbf{y}}$. As $\mathbf{h}$ is still unknown, however, it means that $\underline{\mathbf{h}}$ is also unknown and an estimate for this component is still required. Using the same $\hat{\mathbf{R}}_{\mathbf{nn}}$, two general approaches for the estimation of $\underline{\mathbf{h}}$ can be considered, either making use of an available a priori RTF vector pertaining to the LMA or making use of only the observable microphone data, i.e. a fully data-dependent estimate. The remainder of this section elaborates on these procedures.

3.1 Using an a priori RTF vector

For a microphone configuration consisting of only an LMA, it is not uncommon to use an a priori RTF vector, $\widetilde{\mathbf{h}}_{\mathbf{a}}$ in place of the true RTF vector. As mentioned earlier, this may be measured a priori or based on several assumptions regarding the spatial scenario and acoustic environment. For the inclusion of XMs into the microphone configuration, however, the notion of an a priori RTF vector is not so straightforward as no immediate prior knowledge with respect to the XMs can be exploited since there are no restrictions on what type of XMs can be used or where they must be placed in the acoustic environment. Hence an a priori RTF vector cannot be prescribed, as was the case for the LMA only. However, since a priori information would typically only be available for the LMA, an a priori RTF vector for a microphone configuration of an LMA with XMs can be defined as follows:

$$\widetilde{\mathbf{h}} = [\ \widetilde{\mathbf{h}}_{\mathbf{a}}^{T}\ \mathbf{h}_{\mathbf{e}}^{T}\ ]^{T} \tag{22}$$

which consists partially of the a priori RTF vector pertaining to the LMA, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, and partially of the RTF vector pertaining to the XM, $\mathbf{h}_{\mathbf{e}}$, which is unknown and remains to be estimated. The estimate of $\mathbf{h}_{\mathbf{e}}$ will be denoted as $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e}}$ to emphasise that it is constrained by the a priori knowledge set by $\widetilde{\mathbf{h}}_{\mathbf{a}}$ but estimated from the observed microphone data. In [9], a procedure involving the generalised eigenvalue decomposition (GEVD) was used for obtaining $\hat{\widetilde{\mathbf{h}}}_{\mathbf{e}}$ which is subsequently reviewed and re-framed in the PWT domain.

In the PWT domain, using (13)-(15), a rank-1 matrix approximation problem can be firstly formulated to estimate the entire RTF vector [9]:

$$\underset{\sigma^2_{\mathrm{s}_{\mathrm{a},1}},\,\mathbf{h}}{\text{minimise}} \quad ||(\hat{\underline{\mathbf{R}}}_{\mathbf{yy}} - \hat{\underline{\mathbf{R}}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s}_{\mathrm{a},1}}\hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^{H}\mathbf{h}\,\mathbf{h}^{H}\widetilde{\boldsymbol{\Upsilon}}\hat{\mathbf{L}}^{-H}||^2_F \tag{23}$$

where $||.||_F$ is the Frobenius norm, and:

$$\hat{\underline{\mathbf{R}}}_{\mathbf{yy}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^{H}\hat{\mathbf{R}}_{\mathbf{yy}}\widetilde{\boldsymbol{\Upsilon}}\hat{\mathbf{L}}^{-H} \tag{24}$$

$$\hat{\underline{\mathbf{R}}}_{\mathbf{nn}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^{H}\hat{\mathbf{R}}_{\mathbf{nn}}\widetilde{\boldsymbol{\Upsilon}}\hat{\mathbf{L}}^{-H} = \mathbf{I}_{(M_{\mathrm{a}}+M_{\mathrm{e}})} \tag{25}$$

where $\hat{\mathbf{R}}_{\mathbf{yy}}$ is the data-dependent estimate of $\mathbf{R}_{\mathbf{yy}}$. From (22), an a priori RTF vector in the PWT domain can be defined as follows:

$$\widetilde{\underline{\mathbf{h}}} = \hat{\mathbf{L}}^{-1}\widetilde{\boldsymbol{\Upsilon}}^{H}[\ \widetilde{\mathbf{h}}_{\mathbf{a}}^{T}\ \mathbf{h}_{\mathbf{e}}^{T}\ ]^{T} = \hat{\mathbf{L}}^{-1}\left[\mathbf{0}^{T}\ \ 1\ \ \mathbf{h}_{\mathbf{e}}^{T}\right]^{T} \tag{26}$$

where $\mathbf{0}$ is a vector of $(M_{\mathrm{a}}-1)$ zeros. Replacing $\mathbf{h}$ with the a priori RTF vector from (22) then results in:

$$\underset{\sigma^2_{\mathrm{s}_{\mathrm{a},1}},\,\mathbf{h}_{\mathbf{e}}}{\text{minimise}} \quad ||(\hat{\underline{\mathbf{R}}}_{\mathbf{yy}} - \hat{\underline{\mathbf{R}}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s}_{\mathrm{a},1}}\widetilde{\underline{\mathbf{h}}}\,\widetilde{\underline{\mathbf{h}}}^{H}||^2_F \tag{27}$$

where now only an estimate is required for $\mathbf{h}_{\mathbf{e}}$, which in turn will define the a priori RTF vector. As discussed in [9], it can be observed that it is only the lower $(M_{\mathrm{e}}+1)\times(M_{\mathrm{e}}+1)$ blocks of $\hat{\underline{\mathbf{R}}}_{\mathbf{yy}}$ and $\hat{\underline{\mathbf{R}}}_{\mathbf{nn}}$ that are required for estimating $\mathbf{h}_{\mathbf{e}}$. Hence (27) can be reduced to:

$$\underset{\sigma^2_{\mathrm{s}_{\mathrm{a},1}},\mathbf{h}_{\mathbf{e}}}{\text{minimise}} \quad ||(\hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{yy}} - \hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{nn}}) - \sigma^2_{\mathrm{s}_{\mathrm{a},1}}\mathbf{J}^{T}\widetilde{\underline{\mathbf{h}}}\,\widetilde{\underline{\mathbf{h}}}^{H}\mathbf{J}||^2_F \tag{28}$$

where $\mathbf{J} = [\ \mathbf{0}_{(M_{\mathrm{e}}+1)\times(M_{\mathrm{a}}-1)}\ |\ \mathbf{I}_{(M_{\mathrm{e}}+1)}\ ]^{T}$ is a selection matrix, $\hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{yy}} = \mathbf{J}^{T}\hat{\underline{\mathbf{R}}}_{\mathbf{yy}}\mathbf{J}$, and $\hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{nn}} = \mathbf{J}^{T}\hat{\underline{\mathbf{R}}}_{\mathbf{nn}}\mathbf{J} = \mathbf{I}_{M_{\mathrm{e}}+1}$. The solution of (28) then follows from a GEVD of the matrix pencil $\{\hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{yy}}, \hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{nn}}\}$ or equivalently from the eigenvalue decomposition (EVD) of $\hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{yy}}$ [26]:

$$\hat{\underline{\underline{\mathbf{R}}}}_{\mathbf{yy}} = \hat{\mathbf{V}}\hat{\boldsymbol{\Gamma}}\hat{\mathbf{V}}^{H} \tag{29}$$

where $\hat{\mathbf{V}}$ is a $(M_\mathrm{e}+1)\times(M_\mathrm{e}+1)$ unitary matrix of eigenvectors and $\hat{\mathbf{\Gamma}}$ is a diagonal matrix with the associated eigenvalues in descending order. The estimate of $\mathbf{h_e}$ then follows from the appropriate scaling of the principal eigenvector, $\hat{\mathbf{v}}_\mathbf{p}$:

$$\begin{bmatrix}\mathbf{0}\\1\\\hat{\tilde{\mathbf{h}}}_\mathbf{e}\end{bmatrix}=\frac{\hat{\mathbf{L}}\mathbf{J}\hat{\mathbf{v}}_\mathbf{p}}{\mathbf{e}_{M_\mathrm{a}}^T\hat{\mathbf{L}}\mathbf{J}\hat{\mathbf{v}}_\mathbf{p}}=\frac{\hat{\mathbf{L}}\mathbf{J}\hat{\mathbf{v}}_\mathbf{p}}{\hat{l}_{M_\mathrm{a}}\hat{\mathrm{v}}_{\mathrm{p},1}} \tag{30}$$

where $\mathbf{e}_{M_\mathrm{a}}$ is an $(M_\mathrm{a}+M_\mathrm{e})$ selection vector consisting of all zeros except for a one in the M_a-th position, $\hat{\mathrm{v}}_{\mathrm{p},1}$ is the first element of $\hat{\mathbf{v}}_\mathbf{p}$, and $\hat{l}_{M_\mathrm{a}}$ is the real-valued $(M_\mathrm{a},M_\mathrm{a})$-th element in $\hat{\mathbf{L}}$. Substitution of this expression into (26) finally yields the a priori RTF vector in the PWT domain as[4]:

$$\underline{\widetilde{\mathbf{h}}}=\frac{1}{\hat{l}_{M_\mathrm{a}}\,\hat{\mathrm{v}}_{\mathrm{p},1}}\begin{bmatrix}\mathbf{0}&\hat{\mathbf{v}}_\mathbf{p}\end{bmatrix}^T \tag{31}$$

Finally, replacing $\underline{\mathbf{h}}$ in (21) with $\underline{\widetilde{\mathbf{h}}}$ from (31) results in the MVDR beamformer based on a priori knowledge pertaining to the LMA:

$$\underline{\widetilde{\mathbf{w}}}_\mathrm{mvdr}=\hat{l}_{M_\mathrm{a}}\,\hat{\mathrm{v}}^*_{\mathrm{p},1}\begin{bmatrix}\mathbf{0}&\hat{\mathbf{v}}_\mathbf{p}\end{bmatrix}^T \tag{32}$$

which will be referred to as MVDR-AP. The corresponding speech estimate is then computed using (16):

$$\widetilde{\mathrm{z}}_1=\hat{l}_{M_\mathrm{a}}\,\hat{\mathrm{v}}_{\mathrm{p},1}\,\hat{\mathbf{v}}_\mathbf{p}^H\begin{bmatrix}\underline{\mathrm{y}}_{\mathrm{a},M_\mathrm{a}}\\\underline{\mathbf{y}}_\mathbf{e}\end{bmatrix} \tag{33}$$

As a consequence of incorporating the a priori information into the rank-1 speech model, it can be seen that it is only necessary to filter the last $(M_\mathrm{e}+1)$ elements of $\underline{\mathbf{y}}$, i.e. $\underline{\mathrm{y}}_{\mathrm{a},M_\mathrm{a}}$ and $\underline{\mathbf{y}}_\mathbf{e}$, with the lower order, $(M_\mathrm{e}+1)$ filter defined by $\hat{l}_{M_\mathrm{a}}\,\hat{\mathrm{v}}^*_{\mathrm{p},1}\,\hat{\mathbf{v}}_\mathbf{p}$.

3.2 Using a data-dependent RTF vector

In the PWT domain, it is (23) that needs to be solved in order to obtain a fully data-dependent estimate of the RTF vector pertaining to the LMA and the XMs. The solution to (23) follows from a GEVD of the matrix pencil $\{\underline{\hat{\mathbf{R}}}_\mathbf{yy},\underline{\hat{\mathbf{R}}}_\mathbf{nn}\}$ or equivalently from the EVD of $\underline{\hat{\mathbf{R}}}_\mathbf{yy}$:

$$\underline{\hat{\mathbf{R}}}_\mathbf{yy}=\hat{\mathbf{Q}}\hat{\mathbf{\Lambda}}\hat{\mathbf{Q}}^H \tag{34}$$

[4]It is acknowledged that there is a slight abuse of notation here as the estimate for $\underline{\widetilde{\mathbf{h}}}$ should be denoted as $\underline{\hat{\widetilde{\mathbf{h}}}}$. However, in favour of legibility and to stress that the estimation is done in accordance to the a priori assumptions set by $\widetilde{\mathbf{h}}_\mathbf{a}$ is why the notation is maintained as $\underline{\widetilde{\mathbf{h}}}$.

where $\hat{\mathbf{Q}}$ is an $(M_\mathrm{a}+M_\mathrm{e})\times(M_\mathrm{a}+M_\mathrm{e})$ unitary matrix of eigenvectors and $\hat{\mathbf{\Lambda}}$ is a diagonal matrix with the associated eigenvalues in descending order. The estimated RTF vector is then given by the principal (first in this case) eigenvector, $\hat{\mathbf{q}}_\mathbf{p}$:

$$\hat{\mathbf{h}}=\frac{\widetilde{\mathbf{\Upsilon}}^{-H}\hat{\mathbf{L}}\,\hat{\mathbf{q}}_\mathbf{p}}{\hat{\eta}_q} \tag{35}$$

where $\hat{\eta}_q=\mathbf{e}_1^T\widetilde{\mathbf{\Upsilon}}^{-H}\hat{\mathbf{L}}\,\hat{\mathbf{q}}_\mathbf{p}$ and $\mathbf{e}_1$ is an $(M_\mathrm{a}+M_\mathrm{e})$ selection vector with a one as the first element and zeros everywhere else. In the PWT domain, this data-dependent RTF vector then becomes:

$$\underline{\hat{\mathbf{h}}}=\hat{\mathbf{L}}^{-1}\widetilde{\mathbf{\Upsilon}}^H\hat{\mathbf{h}}=\frac{\hat{\mathbf{q}}_\mathbf{p}}{\hat{\eta}_q} \tag{36}$$

Replacing $\underline{\mathbf{h}}$ in (21) with $\underline{\hat{\mathbf{h}}}$ from (36) results in the MVDR beamformer that makes use of a data-dependent RTF vector:

$$\underline{\hat{\mathbf{w}}}_\mathrm{mvdr}=\hat{\eta}_q^*\,\hat{\mathbf{q}}_\mathbf{p} \tag{37}$$

which will be referred to as MVDR-DD. The corresponding speech estimate is then computed using (16):

$$\hat{\mathrm{z}}_1=\hat{\eta}_q\,\hat{\mathbf{q}}_\mathbf{p}^H\,\underline{\mathbf{y}} \tag{38}$$

where now all $(M_\mathrm{a}+M_\mathrm{e})$ signals need to be filtered as opposed to only $(M_\mathrm{e}+1)$ signals in (33) when an a priori RTF vector is used. In general, the MVDR-DD could also be used for microphone configurations where there is no a priori knowledge available, such as those consisting of external microphones only.

4 Integrated MVDR beamformer

4.1 Quadratically Constrained Quadratic Program

As opposed to relying on only an a priori RTF vector or a data-dependent RTF vector, the merging or integration of both RTF vectors into a single approach can be framed as a quadratically constrained quadratic program (QCQP), firstly with respect to the unprocessed microphone signals:

$$\begin{aligned}\underset{\mathbf{w}}{\text{minimise}}\quad&\mathbf{w}^H\hat{\mathbf{R}}_\mathbf{nn}\mathbf{w}\\\text{s.t.}\quad&|\mathbf{w}^H\widetilde{\mathbf{h}}-1|^2\le\widetilde{\epsilon}^2\\&|\mathbf{w}^H\hat{\mathbf{h}}-1|^2\le\hat{\epsilon}^2\end{aligned} \tag{39}$$

where $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$ are maximum-tolerated squared deviations from a distortionless response due to $\widetilde{\mathbf{h}}$ or $\hat{\mathbf{h}}$

respectively. The constraints of (39) can also be rewritten in the standard form [27] as follows:

$$\mathbf{w}^H\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H\mathbf{w} - 2\Re\{\widetilde{\mathbf{h}}^H\mathbf{w}\} + 1 - \widetilde{\epsilon}^2 \leq 0 \quad (40)$$

$$\mathbf{w}^H\hat{\mathbf{h}}\hat{\mathbf{h}}^H\mathbf{w} - 2\Re\{\hat{\mathbf{h}}^H\mathbf{w}\} + 1 - \hat{\epsilon}^2 \leq 0 \quad (41)$$

where $\Re\{.\}$ denotes the real part of its argument. As the matrices $\hat{\mathbf{R}}_{\mathbf{nn}}$, $\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H$, and $\hat{\mathbf{h}}\hat{\mathbf{h}}^H$ are all positive semidefinite, it is then evident that the QCQP of (39) is convex [27]. In the PWT domain, (39) is equivalently:

$$\begin{aligned} \underset{\underline{\mathbf{w}}}{\text{minimise}} \quad & \underline{\mathbf{w}}^H\underline{\mathbf{w}} \\ \text{s.t.} \quad & |\underline{\mathbf{w}}^H\underline{\widetilde{\mathbf{h}}} - 1|^2 \leq \widetilde{\epsilon}^2 \\ & |\underline{\mathbf{w}}^H\underline{\hat{\mathbf{h}}} - 1|^2 \leq \hat{\epsilon}^2 \end{aligned} \quad (42)$$

where $\underline{\widetilde{\mathbf{h}}}$ and $\underline{\hat{\mathbf{h}}}$ are given in (31) and (36) respectively. Whereas in (20), the hard constraint of $\underline{\mathbf{h}}$ is replaced by either $\underline{\widetilde{\mathbf{h}}}$ or $\underline{\hat{\mathbf{h}}}$, (42) can be interpreted as the relaxation of the hard constraints imposed by $\underline{\widetilde{\mathbf{h}}}$ or $\underline{\hat{\mathbf{h}}}$ by the specified deviations $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$ respectively. In the following, the quantities $|\underline{\mathbf{w}}^H\underline{\widetilde{\mathbf{h}}} - 1|^2$ and $|\underline{\mathbf{w}}^H\underline{\hat{\mathbf{h}}} - 1|^2$ are referred to as speech distortions and $\widetilde{\epsilon}^2$ and $\widetilde{\epsilon}^2$ are the respective maximum tolerable speech distortions. Furthermore, the first inequality constraint in (42) will be referred to as the a priori constraint (APC), and the second inequality constraint will be referred to as the data-dependent constraint (DDC).

The QCQP of (39) is in fact a subset of the more general QCQP considered in [28,29] and as well as an extension to the parametrised multi-channel Wiener filter [30]. In [28,29], the inequality constraints considered are a set of a priori measured RTF vectors, and in [30], only one inequality constraint is considered. The difference in (39) from both of these approaches is that two inequality constraints are considered, one that relies on a priori knowledge, and the other which is fully estimated from the data.

The Lagrangian of (42) is given by:

$$\begin{aligned} \mathcal{L}(\underline{\mathbf{w}}, \alpha, \beta) = \underline{\mathbf{w}}^H\underline{\mathbf{w}} + \alpha(|\underline{\mathbf{w}}^H\underline{\widetilde{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2) \\ + \beta(|\underline{\mathbf{w}}^H\underline{\hat{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2) \end{aligned} \quad (43)$$

where α and β are Lagrangian multipliers. Taking the partial derivative of (43) with respect to $\underline{\mathbf{w}}$ and setting to zero results in what will be referred to as the integrated MVDR beamformer, MVDR-INT:

$$\begin{aligned} \underline{\mathbf{w}}_{\text{int}} = (\mathbf{I}_{(M_{\text{a}}+M_{\text{e}})} + \alpha\underline{\widetilde{\mathbf{h}}}\underline{\widetilde{\mathbf{h}}}^H + \beta\underline{\hat{\mathbf{h}}}\underline{\hat{\mathbf{h}}}^H)^{-1}(\alpha\underline{\widetilde{\mathbf{h}}}\underline{\widetilde{\mathbf{h}}}^H \\ + \beta\underline{\hat{\mathbf{h}}}\underline{\hat{\mathbf{h}}}^H)\mathbf{e}_1 \end{aligned} \quad (44)$$

where the actual values of α and β depend on the prescribed maximum tolerable speech distortions $\widetilde{\epsilon}^2$ and $\widetilde{\epsilon}^2$. It can also be observed that (44) is in fact identical (in the PWT domain) to the integrated MVDR beamformer considered in [22] and hence can be written as a linear combination of $\underline{\widetilde{\mathbf{w}}}_{\text{mvdr}}$ and $\underline{\hat{\mathbf{w}}}_{\text{mvdr}}$ with complex weightings[5] [22]:

$$\underline{\mathbf{w}}_{\text{int}} = g_{ap}\,(\alpha,\beta)\,\underline{\widetilde{\mathbf{w}}}_{\text{mvdr}} + g_{dd}\,(\alpha,\beta)\,\underline{\hat{\mathbf{w}}}_{\text{mvdr}} \quad (45)$$

where $\underline{\widetilde{\mathbf{w}}}_{\text{mvdr}}$ and $\underline{\hat{\mathbf{w}}}_{\text{mvdr}}$ are given in (32) and (37) respectively, and the complex weightings are given by:

$$g_{ap}\,(\alpha,\beta) = \left[\frac{\alpha k_{aa}[1+\beta(k_{bb}-k_{ab})]}{\text{D}}\right] \quad (46)$$

$$g_{dd}\,(\alpha,\beta) = \left[\frac{\beta k_{bb}[1+\alpha(k_{aa}-k_{ba})]}{\text{D}}\right] \quad (47)$$

where

$$\text{D} = \alpha k_{aa} + \beta k_{bb} + \alpha\beta(k_{aa}k_{bb} - k_{ab}k_{ba}) + 1 \quad (48)$$

and

$$k_{aa} = \underline{\widetilde{\mathbf{h}}}^H\underline{\widetilde{\mathbf{h}}};\quad k_{bb} = \underline{\hat{\mathbf{h}}}^H\underline{\hat{\mathbf{h}}}; \quad (49)$$

$$k_{ab} = \underline{\widetilde{\mathbf{h}}}^H\underline{\hat{\mathbf{h}}};\quad k_{ba} = \underline{\hat{\mathbf{h}}}^H\underline{\widetilde{\mathbf{h}}}. \quad (50)$$

Using the expressions for $\underline{\widetilde{\mathbf{w}}}_{\text{mvdr}}$ and $\underline{\hat{\mathbf{w}}}_{\text{mvdr}}$ from (32) and (37) respectively, the resulting speech estimate from the MVDR-INT is then:

$$\text{z}_{\text{int}} = g_{ap}^*\,(\alpha,\beta)\,\widetilde{\text{z}}_1 + g_{dd}^*\,(\alpha,\beta)\,\hat{\text{z}}_1 \quad (51)$$

where $\widetilde{\text{z}}_1$ and $\hat{\text{z}}_1$ are defined in (33) and (38) respectively. Hence the integrated beamformer output is simply a linear combination of the two speech estimates which relied on either a priori information or not.

Once appropriate values are chosen for $\widetilde{\epsilon}^2$ and $\hat{\epsilon}^2$, then a package for specifying and solving convex programs such as CVX [31, 32] can be used for solving (42). Alternatively, more computationally efficient methods may be applied such as those proposed in [28, 29], one of which is highlighted in Algorithm 1. Here, a gradient ascent method [33] for solving (42) is described, which is based on solving the dual problem:

$$\begin{aligned} \underset{(\alpha,\beta)}{\text{maximise}} \quad & \mathscr{D}(\alpha,\beta) \\ \text{s.t.} \quad & \alpha \geq 0;\quad \beta \geq 0 \end{aligned} \quad (52)$$

[5]It can also be expressed as a convex combination of various beamformers as discussed in [22].

where $\mathscr{D}(\alpha,\beta) = \inf_{\underline{\mathbf{w}}_{\text{int}}} \mathcal{L}(\underline{\mathbf{w}}_{\text{int}},\alpha,\beta)$ is the infimum of $\mathcal{L}(\underline{\mathbf{w}}_{\text{int}},\alpha,\beta)$ and referred to as the dual function. As the dual function is concave [27], a gradient ascent procedure can be used to update the values of α and β using the gradients, $\frac{\partial\mathscr{D}(\alpha,\beta)}{\partial\alpha} = |\underline{\mathbf{w}}_{\text{int}}^H\widetilde{\underline{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2$ and $\frac{\partial\mathscr{D}(\alpha,\beta)}{\partial\beta} = |\underline{\mathbf{w}}_{\text{int}}^H\hat{\underline{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2$, i.e. the gradients of the dual function with respect to the particular Lagrange multiplier are the respective constraints. This then gives rise to Algorithm 1 [29], which makes use of the simplified expressions for $\underline{\mathbf{w}}_{\text{int}}$ with the complex-valued weightings as opposed to computing (44) directly. The Lagrangian multipliers, α and β, are then updated via the gradient ascent procedure with the step size γ, whose value can be controlled using a backtracking method [34]. The algorithm continues until the respective gradients are within some specified tolerance, δ.

Algorithm 1 Gradient ascent method for solving the QCQP of (42)

Initialise α, β, $\underline{\mathbf{w}}_{\text{int}}$. Set tolerance, δ. $n = 0$
while $(|\underline{\mathbf{w}}_{\text{int}}^H(n)\widetilde{\underline{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2) > \delta$ OR $(|\underline{\mathbf{w}}_{\text{int}}^H(n)\hat{\underline{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2) > \delta$ **do**
  $g_{ap}(n) = g_{ap}\,(\alpha(n-1),\beta(n-1))$ from (46)
  $g_{dd}(n) = g_{dd}\,(\alpha(n-1),\beta(n-1))$ from (47)
  $\underline{\mathbf{w}}_{\text{int}}(n) = g_{ap}(n)\widetilde{\underline{\mathbf{w}}}_{\text{mvdr}} + g_{dd}(n)\hat{\underline{\mathbf{w}}}_{\text{mvdr}}$ from (45)
  Set γ according to a backtracking method.
  $\alpha(n) = \max\{\alpha(n-1) + \gamma(|\underline{\mathbf{w}}_{\text{int}}^H(n)\widetilde{\underline{\mathbf{h}}} - 1|^2 - \widetilde{\epsilon}^2), 0\}$
  $\beta(n) = \max\{\beta(n-1) + \gamma(|\underline{\mathbf{w}}_{\text{int}}^H(n)\hat{\underline{\mathbf{h}}} - 1|^2 - \hat{\epsilon}^2), 0\}$
  $n = n + 1$
end while

4.2 Effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$

As the QCQP of (42) in principle is to be solved for every time frame and frequency bin, it can therefore lead to quite a versatile beamformer as the parameters, $\widetilde{\epsilon}$ and $\hat{\epsilon}$ can be set independently for each frequency in every time frame in order to define the inequality constraints. So although (42) is a well known QCQP for which there are several methods available to find the solution, it still remains unclear as to what would be a reasonable strategy for setting or tuning $\widetilde{\epsilon}$ and $\hat{\epsilon}$ in practice. As opposed to [22], where tuning rules were developed for the Lagrangian multipliers, here a strategy is outlined for tuning $\widetilde{\epsilon}$ and $\hat{\epsilon}$, which will in turn compute the appropriate Lagrangian multipliers (for instance as outlined in Algorithm 1), as this is believed to be a more insightful procedure.

In order to develop a strategy for tuning $\widetilde{\epsilon}$ and $\hat{\epsilon}$, it will be useful to observe the constraints of (42) in more detail. The derivations that follow will reveal that the space spanned by $\widetilde{\epsilon}$ and $\hat{\epsilon}$ can be divided into four distinct regions as illustrated in Fig. 1, where each of these regions corresponds to a particular set of constraints being active.

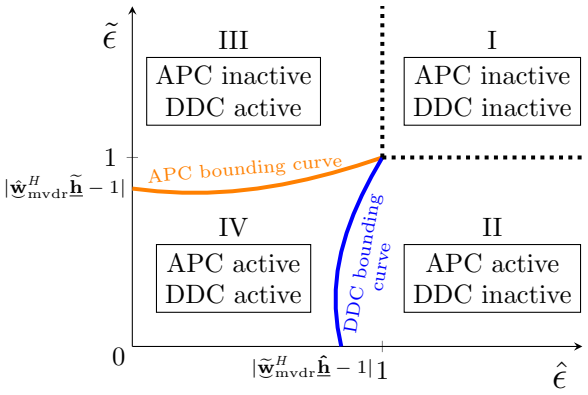


Figure 1 Depiction of the four regions for which the a priori constraint (APC) and the data-dependent constraint (DDC) may be active or inactive within the space spanned by the maximum tolerable speech distortion parameters, $\widetilde{\epsilon}$ and $\hat{\epsilon}$. The curve dividing the regions II and IV is the DDC bounding curve defined when the equality is satisfied from (56). The curve dividing the regions III and IV is the APC bounding curve defined when the equality is satisfied from (60).

Firstly, substitution of $\underline{\mathbf{w}}_{\text{int}} = \mathbf{0}$ into the APC and DDC from (42) shows that when $\widetilde{\epsilon} > 1$ and $\hat{\epsilon} > 1$, both the APC and the DDC are inactive. This condition therefore defines the upper-right region (region I) of Fig. 1 and indeed corresponds to a complete attenuation of the microphone signals, i.e. a zero output signal.

For the case when $\hat{\epsilon} \to \infty$, i.e. when the DDC is inactive, then $\beta \to 0$. If the APC is still active however, it becomes[6]:

$$|\underline{\mathbf{w}}_{\text{int}}^H\widetilde{\underline{\mathbf{h}}} - 1| \leq \widetilde{\epsilon} \tag{53}$$

Furthermore, if $0 \leq \widetilde{\epsilon} \leq 1$, then it can be deduced that:

$$\lim_{\substack{\hat{\epsilon}\to\infty \\ 0\leq\widetilde{\epsilon}\leq 1}} \underline{\mathbf{w}}_{\text{int}} = (1-\widetilde{\epsilon})\widetilde{\underline{\mathbf{w}}}_{\text{mvdr}} \tag{54}$$

Substitution of (54) into (53) readily makes this evident, recalling that $\widetilde{\underline{\mathbf{w}}}_{\text{mvdr}}^H\widetilde{\underline{\mathbf{h}}} = 1$. It is worthwhile to also note that by using (46), the relationship between α and $\widetilde{\epsilon}$ for $0 \leq \widetilde{\epsilon} \leq 1$ is then given as:

$$\lim_{\substack{\hat{\epsilon}\to\infty \\ 0\leq\widetilde{\epsilon}\leq 1}} \alpha = \frac{1}{k_{aa}}\frac{(1-\widetilde{\epsilon})}{\widetilde{\epsilon}} \tag{55}$$

[6] The square root has been taken on both sides of the inequality from (42) in order to simplify the derivations that follow.

In regards to the DDC, as $\hat{\epsilon}$ is decreased (from $\hat{\epsilon} \to \infty$), it remains inactive until $|\underline{\mathbf{w}}^H \hat{\underline{\mathbf{h}}} - 1| = \hat{\epsilon}$. By substitution of (54) into the DDC of (42), the value of $\hat{\epsilon}$ at which the DDC becomes active, $\hat{\epsilon}_o$, is given by:

$$\hat{\epsilon}_o = |\widetilde{\underline{\mathbf{w}}}^H_{\text{mvdr}} \hat{\underline{\mathbf{h}}}(1 - \widetilde{\epsilon}) - 1| \tag{56}$$

In the limits of $\widetilde{\epsilon}$, when $\widetilde{\epsilon} = 1$, $\hat{\epsilon}_o = 1$, and when $\widetilde{\epsilon} = 0$, $\hat{\epsilon}_o = |\widetilde{\underline{\mathbf{w}}}^H_{\text{mvdr}} \hat{\underline{\mathbf{h}}} - 1|$, where depending on $\hat{\underline{\mathbf{h}}}$, $|\widetilde{\underline{\mathbf{w}}}^H_{\text{mvdr}} \hat{\underline{\mathbf{h}}} - 1| < 1$ or $|\widetilde{\underline{\mathbf{w}}}^H_{\text{mvdr}} \hat{\underline{\mathbf{h}}} - 1| \geq 1$. The range of values obtained for $\hat{\epsilon}_o$ from (56) within the domain $0 \leq \widetilde{\epsilon} \leq 1$ define what will be referred to as the DDC bounding curve as depicted in Fig. 1. Hence region II in Fig. 1 is enclosed by the DDC bounding curve, $\widetilde{\epsilon} = 0$ and, $\widetilde{\epsilon} = 1$, representing the space where the APC is active and the DDC is inactive.

A similar analysis can be followed starting from the case when $\widetilde{\epsilon} \to \infty$, i.e. when the APC is inactive and hence $\alpha \to 0$. If the DDC is still active however, it becomes:

$$|\underline{\mathbf{w}}^H_{\text{int}} \hat{\underline{\mathbf{h}}} - 1| \leq \hat{\epsilon} \tag{57}$$

When $0 \leq \hat{\epsilon} \leq 1$, then the following relationships can be deduced:

$$\lim_{\substack{\widetilde{\epsilon} \to \infty \\ 0 \leq \hat{\epsilon} \leq 1}} \underline{\mathbf{w}}_{\text{int}} = (1 - \hat{\epsilon}) \hat{\underline{\mathbf{w}}}_{\text{mvdr}} \tag{58}$$

$$\lim_{\substack{\widetilde{\epsilon} \to \infty \\ 0 \leq \hat{\epsilon} \leq 1}} \beta = \frac{1}{k_{bb}} \frac{(1 - \hat{\epsilon})}{\hat{\epsilon}} \tag{59}$$

Finally, for the APC, as $\widetilde{\epsilon}$ is decreased (from initially $\widetilde{\epsilon} \to \infty$), the value, $\widetilde{\epsilon}_o$, at which this constraint becomes active is given by:

$$\widetilde{\epsilon}_o = |\hat{\underline{\mathbf{w}}}^H_{\text{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}) - 1| \tag{60}$$

In the limits of $\hat{\epsilon}$, when $\hat{\epsilon} = 1$, $\widetilde{\epsilon}_o = 1$, and when $\hat{\epsilon} = 0$, $\widetilde{\epsilon}_o = |\hat{\underline{\mathbf{w}}}^H_{\text{mvdr}} \widetilde{\underline{\mathbf{h}}} - 1|$, where depending on $\widetilde{\underline{\mathbf{h}}}$, $|\hat{\underline{\mathbf{w}}}^H_{\text{mvdr}} \widetilde{\underline{\mathbf{h}}} - 1| < 1$ or $|\hat{\underline{\mathbf{w}}}^H_{\text{mvdr}} \widetilde{\underline{\mathbf{h}}} - 1| \geq 1$. The range of values obtained for $\widetilde{\epsilon}_o$ from (60) within the domain $0 \leq \hat{\epsilon} \leq 1$ define what will be referred to as the APC bounding curve as depicted in Fig. 1. Hence region III in Fig. 1 is enclosed by the APC bounding curve, $\hat{\epsilon} = 0$ and, $\hat{\epsilon} = 1$, representing the space where the APC is inactive and the DDC is active.

Finally, in the lower-left region, region IV, both the APC and the DDC become active within the area enclosed by the APC and DDC bounding curve. It should be kept in mind that Fig. 1 is only an illustration and that the shape of the area for which the APC and DDC are both active can change depending on the RTF vectors, $\widetilde{\underline{\mathbf{h}}}$ and $\hat{\underline{\mathbf{h}}}$. For instance, Fig. 1 shows $|\hat{\underline{\mathbf{w}}}^H_{\text{mvdr}} \widetilde{\underline{\mathbf{h}}} - 1| < 1$ and $|\widetilde{\underline{\mathbf{w}}}^H_{\text{mvdr}} \hat{\underline{\mathbf{h}}} - 1| < 1$ (points on the axes), whereas it is possible that either of these points may be greater than or equal to one.

5 Confidence Metric and Tuning

5.1 Confidence Metric

One of the ingredients towards developing a tuning strategy for setting appropriate values for $\widetilde{\epsilon}$ and $\hat{\epsilon}$ is that of a confidence metric, which is indicative of the confidence in the accuracy of the data-dependent RTF vector. In [22], it was proposed that a principal generalised eigenvalue resulting from the data-dependent estimation procedure be used as such a confidence metric. In the following, it is proposed again to use such a metric, however, due to the formulation in the PWT domain, the principal eigenvalue, $\hat{\lambda}_1$ from the EVD in (34) will be used. It can be shown that $\hat{\lambda}_1$ is equivalent to the resulting posterior SNR when the MVDR-DD is applied and therefore serves as a reasonable metric for making a decision with respect to the accuracy of the data-dependent RTF. For the MVDR-DD in (37), the resulting posterior SNR is given by:

$$\widehat{\text{SNR}}_{\text{DD}} = \frac{\hat{\underline{\mathbf{w}}}^H_{\text{mvdr}} \hat{\underline{\mathbf{R}}}_{\mathbf{yy}} \hat{\underline{\mathbf{w}}}_{\text{mvdr}}}{\hat{\underline{\mathbf{w}}}^H_{\text{mvdr}} \hat{\underline{\mathbf{R}}}_{\mathbf{nn}} \hat{\underline{\mathbf{w}}}_{\text{mvdr}}} \tag{61}$$

where it is recalled that $\hat{\underline{\mathbf{R}}}_{\mathbf{nn}} = \mathbf{I}_{(M_{\text{a}}+M_{\text{e}})}$. Substitution of (34) and (37)[7] into (61) results in $\widehat{\text{SNR}}_{\text{DD}} = \hat{\lambda}_1$.

As in [22], $\hat{\lambda}_1$ can then be used in a logisitic function to define the confidence metric, $\text{F}(l)$[8]:

$$\text{F}(l) = \frac{1}{1 + e^{-\rho(10 \log_{10}(\hat{\lambda}_1(l)) - \lambda_{\text{t}})}} \tag{62}$$

where $\text{F}(l) \in [0, \ 1]$, ρ controls the gradient of the transition from 0 to 1, and λ_{t} is a threshold (in dB), beyond which $\text{F}(l) \to 1$. Hence, as $10 \log_{10}(\hat{\lambda}_1(l))$ increases beyond λ_{t}, then $\text{F}(l) \to 1$, indicating high confidence in the accuracy of the data-dependent RTF vector. On the other hand, as $10 \log_{10}(\hat{\lambda}_1(l))$ decreases below λ_{t}, then $\text{F}(l) \to 0$, indicating low confidence in the accuracy of the data-dependent RTF vector.

5.2 Tuning strategy

With the depiction of the space spanned by $\widetilde{\epsilon}$ and $\hat{\epsilon}$ from Fig. 1 in mind, a general two-step procedure can be followed to establish a particular tuning strategy:

[7] Recall that $\hat{\underline{\mathbf{w}}}_{\text{mvdr}}$ can be equivalently expressed as $\hat{\underline{\mathbf{w}}}_{\text{mvdr}} = \hat{\eta}^*_q \hat{\mathbf{Q}} \mathbf{e}_1$.

[8] The time index is reintroduced here to reinforce that these quantities are to be computed in each time frame. All frequencies are still treated equivalently.

1. Choose two points on the $\{\hat{\epsilon}, \widetilde{\epsilon}\}$ plane: ϵ_{AP} and ϵ_{DD}. The coordinates of ϵ_{AP}, $\{\hat{\epsilon}_{AP}, \widetilde{\epsilon}_{AP}\}$, will specify the maximum tolerable speech distortions for the case when there is no confidence in the accuracy of the data-dependent RTF vector. The coordinates of ϵ_{DD}, $\{\hat{\epsilon}_{DD}, \widetilde{\epsilon}_{DD}\}$, on the other hand, will specify the maximum tolerable speech distortions for the case when there is complete confidence in the accuracy of the data-dependent RTF vector.
2. Define an appropriate path in order to connect ϵ_{AP} and ϵ_{DD}, where the variation along this path would be a function of the confidence metric, $\mathrm{F}(l)$. As $\mathrm{F}(l)$ changes in each time-frequency segment, different values of $\hat{\epsilon}$ and $\widetilde{\epsilon}$ will be chosen along this path and subsequently used in the QCQP from (42).

Figure 2 depicts three examples of how such a general tuning strategy can be interpreted in the $\{\hat{\epsilon}, \widetilde{\epsilon}\}$ plane, where a linear path has been used to connect the points, ϵ_{AP} and ϵ_{DD}. Before further elaborating on Fig. 2, however, one possible tuning strategy will be briefly outlined. In this strategy, ϵ_{AP} and ϵ_{DD}, are chosen by making use of the relationship between the integrated MVDR and the so-called speech distortion weighted multi-channel Wiener filter (SDW-MWF) [35,36]. Although ϵ_{AP} and ϵ_{DD} can in general be chosen without making use of this relation, it is done to highlight how the speech distortion parameter, μ, from the SDW-MWF is related to the maximum tolerable speech distortion parameters of the integrated MVDR, especially as this μ is a well-established trade-off parameter. For the path connecting ϵ_{AP} and ϵ_{DD}, a linear path will be defined using the confidence metric, $\mathrm{F}(l)$.

In the PWT domain, the cost function for the SDW-MWF is given by:

$$\underset{\underbrace{\mathbf{W}}}{\text{minimise}} \quad \mu\, \underbrace{\mathbf{w}}^H \underbrace{\mathbf{w}} + \sigma^2_{\mathrm{s_{a,1}}} |\underbrace{\mathbf{w}}^H \underline{\mathbf{h}} - 1|^2 \tag{63}$$

which consists of two terms, the first corresponding to the noise power spectral density after filtering and the second corresponding to the speech distortion. The speech distortion parameter, $\mu \in (0,\ \infty)$ is used to trade-off between the amount of noise reduction and speech distortion, where larger values of μ put more emphasis on reducing the noise and smaller values put more emphasis on reducing the speech distortion. Two separate SDW-MWF formulations can then be considered for $\underline{\widetilde{\mathbf{h}}}$ and $\underline{\hat{\mathbf{h}}}$ respectively:

$$\underset{\underbrace{\mathbf{W}}}{\text{minimise}} \quad \widetilde{\mu}\, \underbrace{\mathbf{w}}^H \underbrace{\mathbf{w}} + \sigma^2_{\mathrm{s_{a,1}}} |\underbrace{\mathbf{w}}^H \underline{\widetilde{\mathbf{h}}} - 1|^2 \tag{64}$$

$$\underset{\underbrace{\mathbf{W}}}{\text{minimise}} \quad \hat{\mu}\, \underbrace{\mathbf{w}}^H \underbrace{\mathbf{w}} + \sigma^2_{\mathrm{s_{a,1}}} |\underbrace{\mathbf{w}}^H \underline{\hat{\mathbf{h}}} - 1|^2 \tag{65}$$

where $\widetilde{\mu} \in (0,\ \infty)$ and $\hat{\mu} \in (0,\ \infty)$ are the separate speech distortion parameters for each cost function. The solutions to (64) and (65) are then respectively given by:

$$\widetilde{\mathbf{w}}_{\mathrm{sdw}} = (\widetilde{\mu}\mathbf{I}_{(M_{\mathrm{a}}+M_{\mathrm{e}})} + \hat{\sigma}^2_{\mathrm{s_{a,1}}} \underline{\widetilde{\mathbf{h}}}\, \underline{\widetilde{\mathbf{h}}}^H)^{-1} \hat{\sigma}^2_{\mathrm{s_{a,1}}} \underline{\widetilde{\mathbf{h}}}\, \underline{\widetilde{\mathbf{h}}}^H \mathbf{e}_1 \tag{66}$$

$$\hat{\mathbf{w}}_{\mathrm{sdw}} = (\hat{\mu}\mathbf{I}_{(M_{\mathrm{a}}+M_{\mathrm{e}})} + \hat{\sigma}^2_{\mathrm{s_{a,1}}} \underline{\hat{\mathbf{h}}}\, \underline{\hat{\mathbf{h}}}^H)^{-1} \hat{\sigma}^2_{\mathrm{s_{a,1}}} \underline{\hat{\mathbf{h}}}\, \underline{\hat{\mathbf{h}}}^H \mathbf{e}_1 \tag{67}$$

where $\hat{\sigma}^2_{\mathrm{s_{a,1}}}$ is an estimate of $\sigma^2_{\mathrm{s_{a,1}}}$. On comparing the $\underbrace{\mathbf{w}}_{\mathrm{int}}$ in (44) to (66) and (67), it can be observed that there is a relationship between the integrated MVDR beamformer and the SDW-MWF. By considering the expressions written as an MVDR beamformer followed by a single-channel post filter [36], it can be deduced that [22]:

$$\alpha = \frac{\hat{\sigma}^2_{\mathrm{s_{a,1}}}}{\widetilde{\mu}} \ \text{ when } \ \beta = 0 \tag{68}$$

$$\beta = \frac{\hat{\sigma}^2_{\mathrm{s_{a,1}}}}{\hat{\mu}} \ \text{ when } \ \alpha = 0 \tag{69}$$

Proceeding to define the coordinates of ϵ_{AP}, (68) is substituted into (55) to obtain a value for $\widetilde{\epsilon}$ as:

$$\widetilde{\epsilon}_{AP} = \frac{\widetilde{\mu}}{\widetilde{\mu} + \hat{\sigma}^2_{\mathrm{s_{a,1}}} k_{aa}} \tag{70}$$

Hence the range of values for $\widetilde{\mu}$ are essentially compressed into a range of values for $\widetilde{\epsilon}_{AP}$ such that $0 \leq \widetilde{\epsilon}_{AP} \leq 1$. This means that $\widetilde{\epsilon}_{AP}$ can be chosen to be within this range without having to specify $\widetilde{\mu}$. However, (70) serves to clarify how the choice of $\widetilde{\epsilon}_{AP}$ is related to the cost function of (64).

Using the value of $\widetilde{\epsilon}_{AP}$ in (56) then yields a range of choices for $\hat{\epsilon}_{AP}$ such that $\hat{\epsilon}_{AP} \leq \hat{\epsilon}_o$:

$$\hat{\epsilon}_{AP} \leq |\underbrace{\widetilde{\mathbf{w}}}^H_{\mathrm{mvdr}} \underline{\hat{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1| \tag{71}$$

If $\hat{\epsilon}_{AP} = |\underbrace{\widetilde{\mathbf{w}}}^H_{\mathrm{mvdr}} \underline{\hat{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1|$, then ϵ_{AP} lies on the DDC bounding curve of Fig. 1. For all values of $\hat{\epsilon}$ such that $\hat{\epsilon} > |\underbrace{\widetilde{\mathbf{w}}}^H_{\mathrm{mvdr}} \underline{\hat{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1|$, the DDC remains inactive and hence setting a value of $\hat{\epsilon}$ within this region will always result in the same achievable[9] speech distortion defined by $|\underbrace{\widetilde{\mathbf{w}}}^H_{\mathrm{mvdr}} \underline{\hat{\mathbf{h}}}(1 - \widetilde{\epsilon}_{AP}) - 1|$. Furthermore, when the DDC is inactive, then (68) holds, so that values of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ in region II from Fig. 1 would result in the SDW-MWF from (66).

[9] Achievable here is meant to differentiate between the actual speech distortion that is obtained and the maximum tolerable value that was specified.

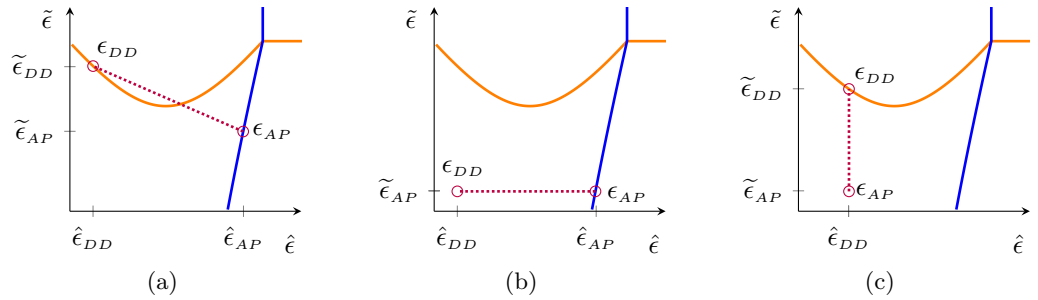


Figure 2 Depiction of three different tuning strategies (a) trading off the maximum tolerable speech distortions between the APC and DDC, (b) fixed maximum tolerable speech distortion for the APC but variable maximum tolerable speech distortion for the DDC, and (c) fixed maximum tolerable speech distortion for the DDC but variable maximum tolerable speech distortion for the APC.

Similarly, by firstly substituting (69) in (59) and making use of (60), the coordinates $\{\hat{\epsilon}_{DD}, \widetilde{\epsilon}_{DD}\}$ of ϵ_{DD} can be defined as:

$$\hat{\epsilon}_{DD} = \frac{\hat{\mu}}{\hat{\mu} + \hat{\sigma}^2_{\mathrm{s_{a,1}}} k_{bb}} \tag{72}$$

$$\widetilde{\epsilon}_{DD} \leq |\hat{\underset{\sim}{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1| \tag{73}$$

Now if $\widetilde{\epsilon}_{DD} = |\hat{\underset{\sim}{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1|$, then ϵ_{DD} lies on the APC bounding curve of Fig. 1. Additionally, for all values of $\widetilde{\epsilon}$ such that $\widetilde{\epsilon} > |\hat{\underset{\sim}{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1|$, the APC remains inactive and hence setting a value of $\widetilde{\epsilon}$ within this region will always result in the same achievable speech distortion defined by $|\hat{\underset{\sim}{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}_{DD}) - 1|$. Furthermore, when the APC is inactive, then (69) holds, so that values of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ in region III from Fig. 1 would result in the SDW-MWF from (67).

The insight of Fig. 1 and additional value of the MVDR-INT as compared to the SDW-MWF is now apparent. Given the two SDW-MWF solutions from (66) and (67), it is not immediately clear how to optimally interpolate between them by using a linear combination of the filters themselves. In Fig. 1, however, it can be seen that an optimal interpolation between (66) and (67), i.e. between regions II and III, can be achieved through the specification of the maximum tolerable speech distortion parameters, $\widetilde{\epsilon}$ and $\hat{\epsilon}$ along some path from region II to region III. In essence, the MVDR-INT has introduced region IV, which serves as a bridge for connectings regions II and III, thereby facilitating the use of both the priori and data-dependent RTF vectors. This then corresponds to the second step of the general procedure for tuning, where ϵ_{AP} and ϵ_{DD} are to be connected. Here, it is proposed to use the confidence metric, $\mathrm{F}(l)$ to perform a linear interpolation between ϵ_{AP} and ϵ_{DD} to yield the values for $\widetilde{\epsilon}$ and $\hat{\epsilon}$ respectively as:

$$\hat{\epsilon} = (1 - \mathrm{F}(l))\ \hat{\epsilon}_{AP} + \mathrm{F}(l)\ \hat{\epsilon}_{DD} \tag{74}$$

$$\widetilde{\epsilon} = (1 - \mathrm{F}(l))\ \widetilde{\epsilon}_{AP} + \mathrm{F}(l)\ \widetilde{\epsilon}_{DD} \tag{75}$$

which are subsequently squared to be used in the QCQP from (42). Consequently, as the confidence in the accuracy of the data-dependent RTF vector increases, the maximum tolerable speech distortions will be specified by values tending towards $\{\hat{\epsilon}_{DD}, \widetilde{\epsilon}_{DD}\}$. On the contrary, as this confidence decreases, maximum tolerable speech distortions will be specified by values tending towards $\{\hat{\epsilon}_{AP}, \widetilde{\epsilon}_{AP}\}$.

Returning focus to Fig. 2, the three examples of a tuning strategy can now be understood. A particular realisation of the APC and the DDC bounding curves has been plotted and the intersecting point of both curves corresponds to the $\{1, 1\}$ coordinate (recall Fig. 1). In the tuning of Fig. 2 (a), as $\mathrm{F}(l)$ increases, the path along the dotted line is taken from ϵ_{AP} to arrive at ϵ_{DD} which gradually sets a larger value of $\widetilde{\epsilon}$ for the APC and a smaller value of $\hat{\epsilon}$ for the DDC. Depending on the particular realisation of the APC and DDC bounding curves, it may be that such a path can entirely lie within the area enclosed by these curves or part of it may lie outside as shown in Fig. 2 (a). The latter is in fact a fortunate circumstance because the achieved speech distortion corresponding to the inactive constraint will actually be lower than what was prescribed by the tuning. In the case of Fig. 2 (a) for instance, when the linear path is above the APC bounding curve, it means that $\widetilde{\epsilon} > |\hat{\underset{\sim}{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}) - 1|$ (recall (60)). Since beyond $|\hat{\underset{\sim}{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}) - 1|$ the APC continues to be inactive, the actual speech distortion that would be achieved in relation to this constraint would correspond to $|\hat{\underset{\sim}{\mathbf{w}}}^H_{\mathrm{mvdr}} \widetilde{\underline{\mathbf{h}}}(1 - \hat{\epsilon}) - 1|$, which is by definition less than $\widetilde{\epsilon}$. Hence, although there is a linear path from ϵ_{AP} to ϵ_{DD}, at the point where this linear path intersects with the APC bounding curve, the

actual speech distortions that would be achieved are those that continue along the APC bounding curve in order to arrive at ϵ_{DD}.

The tunings depicted in Fig. 2 (b) and (c) are representative of strategies where the maximum tolerable speech distortion is fixed for one of the constraints, and only the maximum tolerable speech distortion for the other constraint is tuned. In Fig. 2 (b), ϵ_{DD} is defined by setting $\widetilde{\epsilon}_{DD} = \widetilde{\epsilon}_{AP}$, so that the maximum tolerable speech distortion for the APC is fixed. $\hat{\epsilon}$ is then tuned according to (74). This is representative of a case where the APC is always active and the DDC is only included if there is confidence in the accuracy of the data-dependent RTF vector. Fig. 2 (c) depicts an opposite strategy, where now ϵ_{AP} is set by setting $\hat{\epsilon}_{AP} = \hat{\epsilon}_{DD}$, so that the maximum tolerable speech distortion for the DDC is fixed.

6 Evaluation and Discussion

In order to gain further insight into the behaviour of the integrated MVDR beamformer using the QCQP formulation, a simulation was firstly considered involving only an LMA without XMs. As will be demonstrated, observing such a scenario facilitates the visualisation of the theoretical beam patterns that would be generated under different tuning strategies. Following this simulation, recorded data from an acoustic scenario involving behind-the-ear dummy[10] hearing aid microphones along with XMs in a cocktail party scenario was then analysed and evaluated.

6.1 Beam patterns for a linear microphone array

As the notion of a traditional beam pattern is not immediately extended to the case of an LMA with XMs[11], the following beam patterns are generated using an LMA only.

For visualising the beam patterns, a linear LMA consisting of 4 microphones and 5 cm spacing was considered. Two anechoic RTF vectors, simulating an a priori RTF vector, $\widetilde{\mathbf{h}}_{\mathbf{a}}$, and a data-dependent RTF vector, $\hat{\mathbf{h}}_{\mathbf{a}}$, were computed according to a far-field approximation, i.e. $[1\ e^{-j2\pi f\tau_2(\theta)}\ e^{-j2\pi f\tau_3(\theta)}\ e^{-j2\pi f\tau_4(\theta)}]^T$, where f is the frequency (Hz) which was set to 3 kHz, $\tau_m(\theta) = \frac{(m-1)0.05\cos(\theta)}{c}$ is the relative time delay between the m^{th} microphone and the reference microphone (the microphone closest to the desired speech source) of the LMA, θ is the angle of the desired speech source, and $c = 345\,\mathrm{m\,s^{-1}}$ is the speed of sound. For $\widetilde{\mathbf{h}}_{\mathbf{a}}$, $\theta = 0°$ and for $\hat{\mathbf{h}}_{\mathbf{a}}$, $\theta = 60°$. Using this definition of $\widetilde{\mathbf{h}}_{\mathbf{a}}$, $\widetilde{\mathbf{C}}_{\mathbf{a}}$, and $\widetilde{\mathbf{f}}_{\mathbf{a}}$ were defined accordingly from (7) and $\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}}$ from (8). With $\mathbf{R}_{\mathbf{n_a n_a}} = \mathbf{I}_{M_a}$, the pre-whitening operation from (10) was then computed but with $\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}}$ instead of $\widetilde{\boldsymbol{\Upsilon}}$, and hence denoted as $\mathbf{L}_{\mathbf{a}}$. In the PWT domain, the respective RTF vectors are given by $\underline{\widetilde{\mathbf{h}}}_{\mathbf{a}} = \mathbf{L}_{\mathbf{a}}^{-1}\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}}^{H}\widetilde{\mathbf{h}}_{\mathbf{a}}$ and $\underline{\hat{\mathbf{h}}}_{\mathbf{a}} = \mathbf{L}_{\mathbf{a}}^{-1}\widetilde{\boldsymbol{\Upsilon}}_{\mathbf{a}}^{H}\hat{\mathbf{h}}_{\mathbf{a}}$. The optimal PWT domain filters, $\underline{\widetilde{\mathbf{w}}}_{\mathbf{a}}$, and $\underline{\hat{\mathbf{w}}}_{\mathbf{a}}$ were then computed as in (21), but using either $\underline{\widetilde{\mathbf{h}}}_{\mathbf{a}}$ or $\underline{\hat{\mathbf{h}}}_{\mathbf{a}}$. Finally, (74) and (75) were used to $\widetilde{\epsilon}$ and $\hat{\epsilon}$, after which (42) was then solved using CVX [31, 32] to yield the integrated MVDR beamformer for the LMA only, denoted as $\underline{\mathbf{w}}_{\mathbf{a},\mathrm{int}}$. The beam patterns were computed as $|\underline{\mathbf{w}}_{\mathbf{a},\mathrm{int}}^{H}\underline{\mathbf{h}}(\theta)|$, where $\underline{\mathbf{h}}(\theta)$ is the PWT domain RTF vector corresponding to an angle, θ.

Figure 3 illustrates the resulting beam patterns for two tuning strategies for different values of $\mathrm{F}(l)$ (in this case $l = 1$ and hence the dependence on l is omitted). The left-hand plot of Fig. 3 corresponds to a tuning strategy similar to that depicted in Fig. 2 (a), where there is a trade-off between the two constraints. For this strategy, $\widetilde{\mu} = \hat{\mu} = 0.2$ and $\hat{\sigma}^2_{\mathrm{s_{a,1}}} = 1$, which means that ϵ_{AP} and ϵ_{DD} were fairly close to the x-axis and y-axis respectively. As F increases, the beam pattern is clearly seen to evolve from focusing on the a priori direction of 0° to eventually that of the data-dependent direction of 60°. As a linear path is followed, at the midpoint, both $\widetilde{\epsilon}$ and $\hat{\epsilon}$ are of a similarly larger values, which explains the nature of the lower magnitude in the beam pattern during the transition.

The right-hand plot of Fig. 3 corresponds to a tuning strategy as depicted in Fig. 2 (b), i.e. when the APC is always active. As F increases, it can be observed that the beam in the a priori direction of 0° is maintained, while more gain is attributed to the data-dependent direction of 60°. In this particular case, however, it is noted that although the response at 60° is in accordance with the maximum tolerable speech distortion prescribed, there is a slight tilt of the beam towards 68° as compared to if only the DDC was active. Nevertheless, this can still be a useful tuning strategy for cases when a high confidence is placed on the a priori RTF vector.

6.2 Effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$

In this section the effect of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ on the behaviour of the integrated MVDR beamformer for the case of an LMA and XMs is further investigated using recorded audio data. A batch processing framework will be applied so as to observe an average performance at a single frequency. In the following section, the processing will be done using a Weighted Overlap and Add (WOLA) framework [37] and a broadband performance will be assessed.

[10] This means that only the microphone signals alone, without any processing are captured.

[11] The complication arises in that some of the XMs can be in the nearfield with respect to the desired source. A visualisation can nevertheless be created, but will have to be considered within a plane or volume with cartesian coordinates.

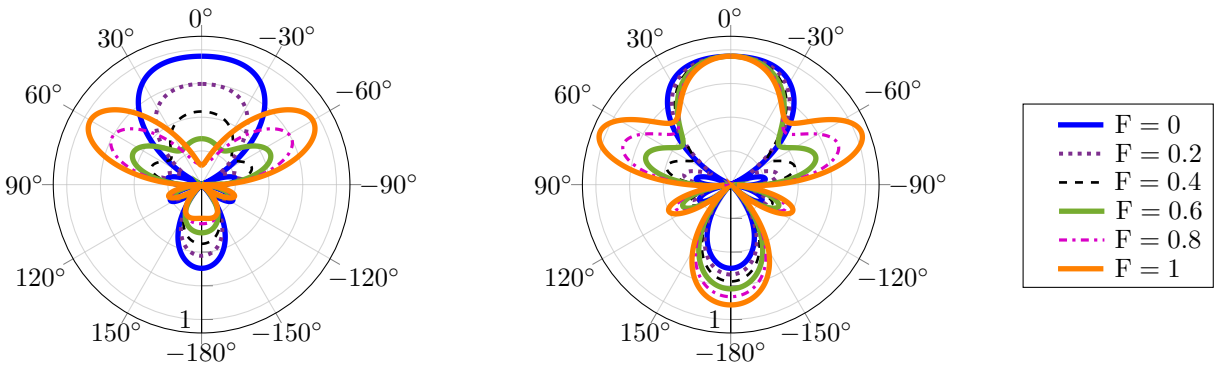


Figure 3 Beam patterns as a function of the confidence metric, F, at a frequency of 3 kHz for different tunings of the integrated MVDR-LMA-XM beamformer as applied to a microphone configuration consisting of an LMA only. (Left) A tuning strategy similar to that depicted in Fig. 2 (a) and (Right) A tuning strategy similar to that depicted in Fig. 2 (b). $\mathrm{F} = 0$ corresponds to the position ϵ_{AP} and $\mathrm{F} = 1$ corresponds to the position ϵ_{DD} from Figs. 2. As F increases, the path from ϵ_{AP} to ϵ_{DD} is followed resulting in the depicted beam patterns.

Audio recordings of speech and noise were made in the laboratory room as depicted in Fig. 4, which has a reverberation time of approximately 1.5 s. A Neumann KU-100 dummy head was placed in a central location of the room and equipped with two (i.e. left and right) behind-the-ear hearing aids, each consisting of two microphones spaced approximately 1.3 cm apart. Hence, in the following, the LMA is considered as having a total of four microphones, i.e. the stacked left ear and right ear microphones. The first microphone of the left ear hearing aid was used as the reference microphone. Three omnidirectional XMs (two AKG CK32 microphones and one AKG CK97O microphone) were placed at heights of 1 m from the floor and at varying distances from the dummy head as shown in Fig. 4. A Genelec 8030C loudspeaker, was placed at 1 m and different azimuth angles from the dummy head to generate a speech signal from a male speaker [38]. The loudspeaker and the dummy head were placed at a height of approximately 1.3 m from the floor (only angles 0° and 60° were used as shown in Fig. 4). For the noise, a cocktail party scenario was re-created. With the same configuration of the dummy head and external microphones from Fig. 4, participants stood outside of a 1 m circumference from the dummy head in a random manner (i.e. all participants were not confined to a particular corner in the room). Beverages in glasses as well as snacks were served while the participants engaged in conversation. At any given time, there were nine male participants and six female participants present in the room. A recording of such a scenario was made for approximately one hour, but a random sample was used in the following analysis.

As opposed to a free-field a priori RTF vector, a more suitable a priori RTF vector for the behind-the-ear hearing aid microphones was obtained from pre-measured impulse responses in the scenario as depicted in Fig. 4. The impulse responses were computed from an exponential sine-sweep measurement with the loudspeaker position at 0° (the azimuth direction directly in front of the dummy head) and 1 m so that the a priori RTF vector would be defined in accordance with a source located at 0° and 1 m from the dummy

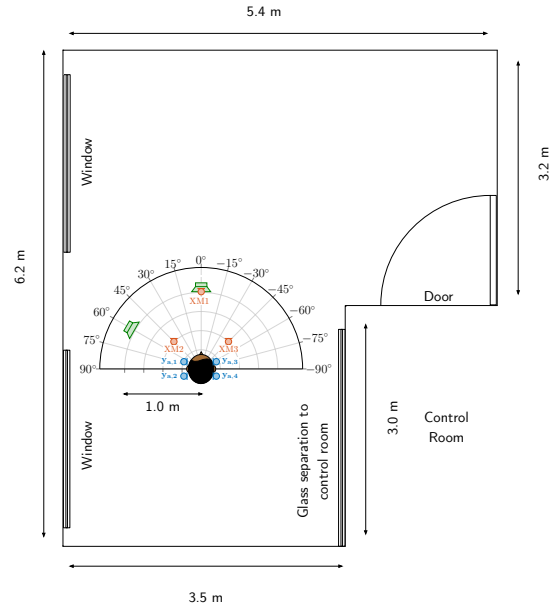


Figure 4 Spatial scenario for the audio recordings. Separate recordings were made of speech signals from the loudspeakers positioned at 0° and 60°. These were then mixed with a cocktail party type noise as explained in the text to create the noisy microphone signals.

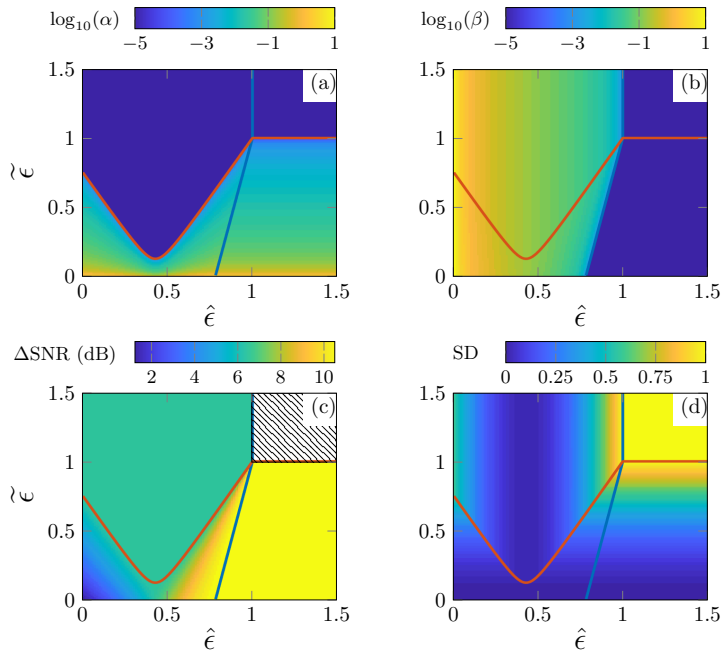


Figure 5 Behaviour of the integrated MVDR-LMA-XM beamformer at a frequency of 2 kHz as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ for the case when the desired speech source is at 0°, i.e., in the direction of the a priori constraint. (a) Lagrangian multiplier, $\log_{10}(\alpha)$, (b) Lagrangian multiplier, $\log_{10}(\beta)$, (c) ΔSNR, (d) Speech Distortion (SD). The APC and DDC bounding curves analogous to those from Fig. 1 are also shown.

head. The initial section of these impulse responses corresponding to the direct component was extracted, with a length according to the size of the Discrete Fourier Transform (DFT) window to be used in the STFT domain processing. This direct component was then smoothed with a Tukey window and converted to the frequency domain. In each frequency bin, these smoothed frequency domain impulse responses were then scaled with respect to the smoothed frequency domain impulse response of the reference microphone. This was then used as $\widetilde{\mathbf{h}}_{\mathbf{a}}(k)$, and was kept the same for each time frame.

A scenario was firstly considered for the desired speech source located at 0° in Fig. 4, i.e. the location where the a priori RTF vector was defined. A 4 s sample of the desired speech signal was mixed with a random sample of the cocktail party noise at a broadband input SNR of 0 dB. For the batch processing framework with a DFT size of 256 samples, $\mathbf{R}_{\mathbf{yy}}$ and $\mathbf{R}_{\mathbf{nn}}$ were estimated by time averaging across the entire length of the signal in the respective speech-plus-noise or noise-only frames. Using the SPP [25] from the first microphone of the left ear hearing aid, frames for which the speech was active were chosen if the resulting SPP > 0.85. The RTF vectors, $\underline{\widetilde{\mathbf{h}}}$ and $\underline{\hat{\mathbf{h}}}$, were computed according to the procedures described in sections 3.1 and 3.2. Using CVX [31,32], the MVDR-INT from (42) was then evaluated for a range of $0 < \widetilde{\epsilon} < 1.5$ and $0 < \hat{\epsilon} < 1.5$ at a frequency of 2 kHz.

Figures 5 (a) and (b) display the resulting (base-10 log) values of the Lagrangian multipliers α and β respectively as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$, along with the APC and DDC bounding curves. These plots support the theoretical analysis of the space spanned by $\widetilde{\epsilon}$ and $\hat{\epsilon}$ from Fig. 1. In Fig. 5 (a), it is clearly observed that as the value of $\widetilde{\epsilon}$ exceeds the APC bounding curve, then $\alpha \to 0$ so that the APC is inactive while the DDC remains active. Similarly, in Fig. 5 (b), as the value of $\hat{\epsilon}$ exceeds the DDC bounding curve, then $\beta \to 0$ so that the APC remains active and the DDC is inactive. The regions where both constraints are active, and when neither are active can also be observed.

Figures 5 (c) and (d) are plots of the corresponding change in SNR (ΔSNR) from the reference microphone as well as the speech distortion which was computed as follows:

$$\Delta\text{SNR} = 10\log_{10}\left(\frac{|\underline{\mathbf{w}}_{\text{int}}^H\underline{\mathbf{h}}|^2}{\underline{\mathbf{w}}_{\text{int}}^H\underline{\mathbf{w}}_{\text{int}}}\right) - 10\log_{10}\left(\frac{1}{\mathbf{e}_1^T\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{e}_1}\right) \tag{76}$$

$$\text{SD} = |\underline{\mathbf{w}}_{\text{int}}^H\underline{\mathbf{h}} - 1|^2 \tag{77}$$

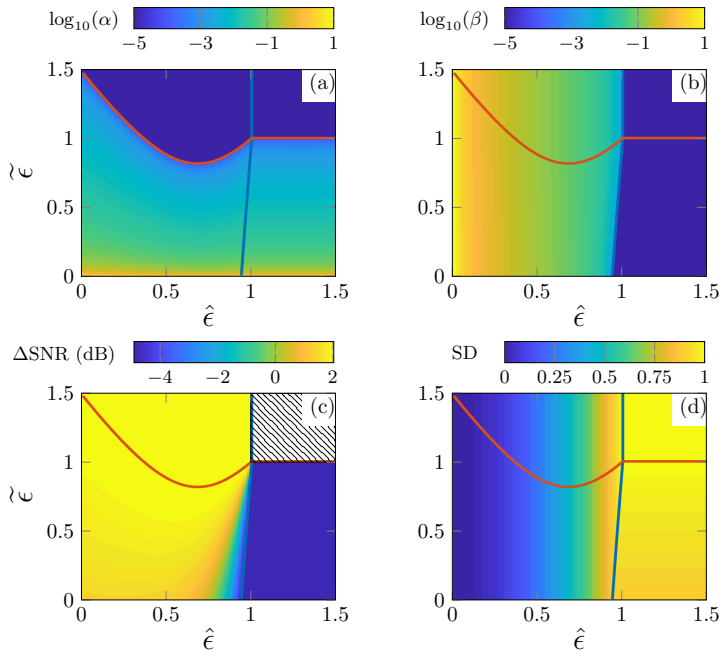


Figure 6 Behaviour of the integrated MVDR-LMA-XM beamformer at a frequency of 2 kHz as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$ for the case when the source is at 60°, i.e., not in the direction of the a priori constraint. (a) Lagrangian multiplier, $\log_{10}(\alpha)$, (b) Lagrangian multiplier, $\log_{10}(\beta)$, (c) ΔSNR, (d) Speech Distortion (SD). The APC and DDC bounding curves analogous to those from Fig. 1 are also shown.

where the first term of the ΔSNR is the output SNR and the second term is the input SNR at the unprocessed reference microphone[12] and in this scenario $\underline{\mathbf{h}} = \underline{\widetilde{\mathbf{h}}}$. The true value of $\underline{\mathbf{h}}$ is unknown, hence the results of Figs. 5 (c) and (d) are suggestive for the case when the true RTF vector corresponds to that of the a priori assumed RTF vector. In Fig. 5 (c), since $\underline{\mathbf{w}} \to \mathbf{0}$ in the region where $\hat{\epsilon} \geq 1$ and $\widetilde{\epsilon} \geq 1$, it is purposefully hatched so as to indicate that in this region an output SNR is undefined.

As expected, it can be observed that the best ΔSNR is achieved for the region where the DDC is inactive and the APC is active, with a compromise within the region where the two constraints are active. An interesting observation here is the poor ΔSNR in the region where $\widetilde{\epsilon} \to 0$ and $\hat{\epsilon} \to 0$. Even though the maximum tolerable speech distortions have been specified to be quite small, in this case $\underline{\widetilde{\mathbf{h}}}$ and $\underline{\hat{\mathbf{h}}}$ can be parallel, which can lead to redundant constraints and an ill-conditioning problem as discussed in [22]. In terms of the SD, fairly low distortions are achieved when either of the constraints are active or when both are active. As both $\widetilde{\epsilon} \to 1$ and $\hat{\epsilon} \to 1$, the speech distortion increases, which is expected from (70) and (72), i.e. the SDW-MWF parameters, $\widetilde{\mu}$ and $\hat{\mu}$. As $\widetilde{\mu} \to \infty$, $\widetilde{\epsilon} \to 1$, and as $\hat{\mu} \to \infty$, $\hat{\epsilon} \to 1$, which accounts for the increasing speech distortion from Fig. 5 (d). Another point to highlight in Fig. 5 (d) is that a low speech distortion is also achieved in the region where the APC bounding curve is a minimum, regardless the value of $\widetilde{\epsilon}$. As discussed in section 5.2, for a value of $\widetilde{\epsilon} > \widetilde{\epsilon}_o$ (where $\widetilde{\epsilon}_o$ is the value of $\widetilde{\epsilon}$ on the APC bounding curve from (60)), the achievable distortion would in fact correspond to $\widetilde{\epsilon}_o$ on the APC bounding curve, which is quite low in this minimum region.

Figure 6 now displays a similar set of results, however for the case when the desired speech source was located at 60° as depicted in Fig. 4. As the a priori RTF vector was based on a speaker located at 0°, this scenario represented a mismatch between the a priori RTF vector and the true RTF vector. The same procedure as previously described was also followed to obtain the MVDR-INT filters.

Figures 6 (a) and (b) display the resulting values of the (base-10 log) Lagrangian multipliers α and β respectively as a function of $\widetilde{\epsilon}$ and $\hat{\epsilon}$, along with the APC and DDC bounding curves. The nature of these plots is quite similar to that of Figs. 5 (a) and (b) in terms of how α and β vary with respect to the bounding curves. In comparison to Figs. 5 (a) and (b), Figs. 6 (a) and (b) also highlight the fact that these bounding curves can have quite different appearances.

[12]The numerator of this term is 1 since the first component of the RTF vector for the unprocessed microphone signals is 1.

Figures 6 (c) and (d) display the corresponding ΔSNR and SD respectively, however with $\underline{\mathbf{h}} = \underline{\hat{\mathbf{h}}}$ in (76) and hence the results are suggestive for the case when the true RTF vector corresponds to that of the data dependent RTF vector. Now it can be observed that the best ΔSNR is achieved for the region where the APC is inactive and the DDC is active, with a compromise within the region where the two constraints are active. For the SD, fairly low speech distortions are achieved for small values of $\hat{\epsilon}$ as expected. For small values of $\widetilde{\epsilon}$ and large values of $\hat{\epsilon}$, i.e. toward the region where only the APC is active, it can be observed that the speech distortion increases, which is a direct result of the speech source not being in the a priori defined direction of 0°. Once again, it can also be seen that the speech distortion generally increases as both $\widetilde{\epsilon} \to 1$ and $\hat{\epsilon} \to 1$.

The results of Figs. 5 and 6 provide some more insight into the behaviour of the MVDR-INT and demonstrate that in some scenarios a better performance can be achieved when either only the APC or only the DDC is active. Furthermore, it was observed that there were transition regions where a compromise could be achieved between these limits of performance when either only the APC or only the DDC is active. Therefore it suggests that tuning strategies such as those depicted in Fig. 2 would indeed be appropriate means of obtaining an optimal filter as opposed to relying on only an APC or DDC.

6.3 Performance of tuning strategies

The audio recordings as previously described for the scenario depicted in Fig. 4 were also used to observe the performance of the tuning strategies. A desired speech signal was created where the desired speech source was initially located at 0° for a duration of 5 s, and then instantaneously moved to 60° for another 6 s. This was then mixed with a random sample of the cocktail party noise at a broadband input SNR of 2 dB. The same a priori RTF vector pertaining to the hearing aid microphones, $\widetilde{\mathbf{h}}_{\mathbf{a}}(k)$, as previously described was used, i.e., $\widetilde{\mathbf{h}}_{\mathbf{a}}(k)$ was computed for a source located at 0° and 1 m from the dummy head.

For the STFT processing, the WOLA method, with a DFT size of 256 samples, 50 % overlap, a square-root hanning window, and a sampling frequency of 16 kHz were used. By using the SPP [25] computed on XM2, frames were classified as containing speech if the SPP > 0.8, otherwise the frames were classified as noise-only. All RTF vector estimates were performed in frames which were classified as containing speech. All the relevant correlation matrices were also estimated using a forgetting factor corresponding to an averaging time of 300 ms. $\mathbf{R}_{\mathbf{nn}}$ was only estimated when the SPP < 0.8.

For the MVDR-INT, two tuning strategies were considered - (i) the tradeoff between the maximum tolerable speech distortions for the APC and DDC, corresponding to Fig. 2 (a), which will be referred to as MVDR-INT-3a and (ii) where the maximum tolerable speech distortion for the APC is constant, but the maximum tolerable speech distortion for the DDC varies, corresponding to Fig. 2 (b), and which will be referred to as MVDR-INT-3b. For both tunings, $\widetilde{\mu} = \hat{\mu} = 0.001$, and $\hat{\sigma}^2_{\mathrm{s_{a,1}}}$ was computed using the method from [39] as implemented in [40] but with the noise estimation update computed as in [25]. A different setting was used for the confidence metric, $\mathrm{F}(l)$ in (62) for each of the tunings such that for the MVDR-INT-3a, $\rho = 1$ and $\lambda_{\mathrm{t}} = 5\,\mathrm{dB}$, and for MVDR-INT-3b, $\rho = 1$ and $\lambda_{\mathrm{t}} = 10\,\mathrm{dB}$, i.e. a higher thresholding was used for the MVDR-AP tuning. With all parameters assigned, the QCQP problem from (42) was solved using the gradient ascent procedure as described in Algorithm 1.

The metrics used to evaluate the following experiments were the speech intelligibility-weighted SNR [41] (SI-SNR), the short-time objective intelligibility (Δ STOI) [42], and the normalised speech-to-reverberation modulation energy ratio for cochlear implants, (SRMR-CI) [43]. The SI-SNR improvement in relation to the reference microphone was calculated as:

$$\Delta\text{SI-SNR} = \sum_i I_i \left(\mathrm{SNR}_{\mathrm{i,out}} - \mathrm{SNR}_{\mathrm{i,in}}\right) \tag{78}$$

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, $\mathrm{SNR}_{i,\mathrm{in}}$ is the input SNR (dB), and $\mathrm{SNR}_{i,\mathrm{out}}$ 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 [44]. The input SNR was computed accordingly using the unprocessed speech-only and unprocessed noise-only components (in the discrete time domain) at the reference microphone, 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 STOI metric, the reference signal used was the unprocessed desired speech source convolved with 256 samples (i.e, same length as the DFT size) of the (pre-measured) impulse response from the desired speech signal location to the reference microphone. As the room was quite reverberant, however, a true reference signal is somewhat ambiguous to define, and hence the non-intrusive metric, SRMR-CI, suitable for hearing instruments, in particular cochlear implants was also used.

Figure 7 displays the performance of the various algorithms, where all the metrics have been computed

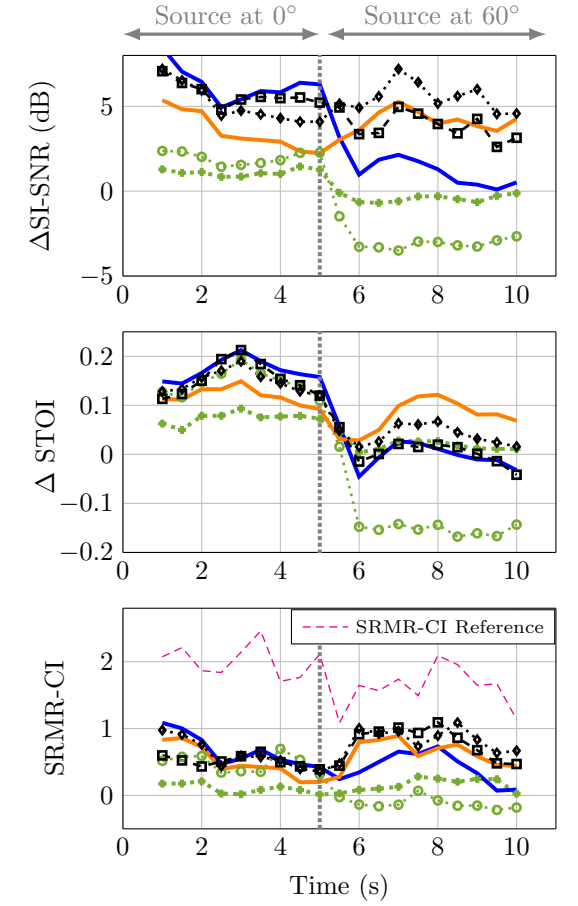


Figure 7 Performance of the MVDR-AP, MVDR-DD, and two tunings of the integrated MVDR beamformer, MVDR-INT-3a and MVDR-INT-3b along with XM1 and XM2 from Fig. 4. The vertical lines are indicative of the time at which the source moves from 0° to 60°.

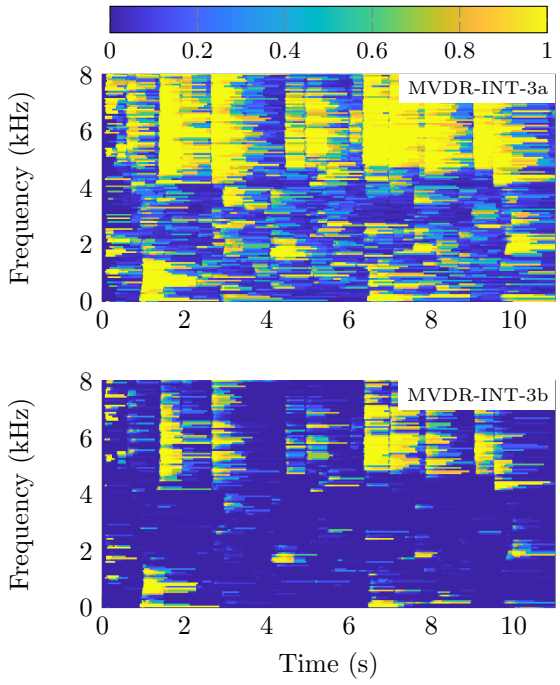


Figure 8 Confidence metrics of the evaluation performed in Fig. 7 for (Top) MVDR-INT-3a and (Bottom) MVDR-INT-3b.

in 2 s time frames with a 25 % overlap. The relative improvements of the SI-SNR and the STOI metrics in relation to the reference microphone have been plotted. The metrics for XM1 and XM2 from Fig. 4 are also plotted. In order to contextualise the values of the SRMR-CI metric, an additional plot of the performance for the reference signal (that which was used for the STOI metric) is displayed. From all the metrics, as expected the MVDR-AP performs better than the MVDR-DD in the first 5 s as the speech source was at 0°, i.e. the a priori direction. However, in the latter 6 s, when the speech source was at 60°, the MVDR-DD achieves a better performance.

With respect to the XMs, it can also be seen that the performance of XM1 decreases after 5 s as the source moves to the location of 60°, while XM2 has more of a consistent performance across the different speech locations. In terms of the Δ SI-SNR, the performance of all of the other algorithms are better than either of the XMs, which demonstrates that simply listening to the XM only would not always immediately yield satisfactory performance.

Within the first 5 s, the MVDR-INT-3a is able to find a compromise between the MVDR-AP and MVDR-DD in terms of all metrics. In the final 6 s, although the Δ STOI is once again in between the MVDR-AP and MVDR-DD, the performance in terms of Δ SI-SNR and SRMR-CI is in fact better than either of the MVDR-AP or the MVDR-DD. This is a direct consequence of the nature of the integrated MVDR-LMA-XM beamformer as different linear combinations of the MVDR-AP and the MVDR-DD are effectively applied to different time-frequency segments, yielding a broadband SI-SNR that could be better than either the MVDR-AP or MVDR-DD.

For the MVDR-INT-3b, within the first 5 s, the performance in terms of all the metrics is closer to that of the MVDR-AP which is expected as the APC is kept active at all times. In the following 6 s, the STOI metric indicates that the speech intelligibility has not changed from that of the MVDR-AP. However, an improvement can be observed in both Δ SI-SNR and SRMR-CI metrics as some frequency bins would have also had the DDC active.

The corresponding confidence metrics across all time frames and frequencies for the MVDR-INT-3a and the MVDR-INT-3b are displayed in Fig. 8. The upper plot corresponds to the confidence metric of MVDR-INT-3a, and reveals that much of the confidence has been placed on the higher frequencies, presumably because there was less noise in this region. Therefore a smaller value of $\hat{\epsilon}$ and a larger value of $\widetilde{\epsilon}$ would have been assigned to the DDC and APC respectively, i.e. the MVDR-INT-3a in this region would have tended toward the MVDR-DD. Several regions of uncertainty are also observed where the MVDR-INT-3a would then find a compromise between the MVDR-AP and the MVDR-DD. In the lower plot of Fig. 8, the confidence metric for the MVDR-INT-3b shows a much more conservative behaviour due to the larger threshold of λ_t. It is observed that there are now many regions where there is little confidence, and hence a larger value of $\hat{\epsilon}$ and a smaller value of $\widetilde{\epsilon}$ would have been assigned to the DDC and APC respectively, i.e. the MVDR-INT-3b in these regions would have tended toward the MVDR-AP. More confidence is now only placed in the higher frequency region and there are still some regions of uncertainty so that a compromise can be achieved. The resulting audio signals from this section[13] may also be listened to for a subjective evaluation at [45].

7 Conclusion

An integrated MVDR beamformer that merges the benefits from using an available a priori relative transfer function (RTF) vector and a data-dependent RTF vector was developed for a microphone configuration consisting of a local microphone array (LMA) and multiple external microphones (XMs). The framework has been presented in a pre-whitened-transformed (PWT) domain, which consists of an initial transformation of the microphone signals through a blocking matrix and a fixed beamformer, followed by a pre-whitening operation, facilitating convenient processing operations. In the PWT domain, procedures for obtaining an a priori RTF vector and data-dependent RTF vector have also been derived, where the a priori RTF vector is based on an a priori RTF vector pertaining to the LMA only.

With the two RTF vectors, an integrated MVDR beamformer was proposed by formulating a quadratically constrained quadratic program (QCQP), with two constraints, one of which is related to the maximum tolerable speech distortion for the imposition of the a priori RTF vector and the other related to the maximum tolerable speech distortion for the imposition of the data-dependent RTF vector. It was shown how the space spanned by each of these maximum tolerable speech distortions could be divided into four separate regions, each of which corresponded to a particular set of constraints being active or inactive. This insight then facilitated the development of a general tuning framework where the maximum tolerable speech distortions are chosen in accordance with the confidence in the accuracy of the data-dependent RTF vector. A particular set of tuning rules was also proposed, which made use of a relationship to the speech distortion weighted multi-channel Wiener filter.

The potential of the integrated MVDR beamformer was demonstrated by using audio data from an LMA of behind-the-ear hearing aid microphones and three XMs for a single desired speech source within a recreated cocktail-party scenario. A narrowband evaluation confirmed the theoretical behaviour of the integrated MVDR as a function of the maximum tolerable speech distortion parameters. A broadband evaluation has shown that the integrated MVDR beamformer can be tuned to yield different enhanced speech signals, which may be suitable for improving speech intelligibility despite changes in the desired speech source position and imperfectly estimated spatial correlation matrices.

Acknowledgements

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 also 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.

Abbreviations

APC: a priori constraint; DDC: data-dependent constraint; EVD: eigenvalue decomposition; GEVD: generalised eigenvalue decomposition; LMA: local microphone array; MVDR: minimum variance distortionless response; MVDR-DD: fully data-dependent MVDR beamformer; MVDR-AP: MVDR beamformer based on a priori knowledge; MVDR-INT: integrated MVDR beamformer; MWF: multi-channel Wiener filter; PWT: pre-whitened-transformed; QCQP: quadratically constrained quadratic program; RTF: relative transfer function; SNR: signal-to-noise ratio; SI-SNR: speech intelligibility-weighted SNR; SPP: speech presence probability; SRMR-CI: speech-to-reverberation modulation energy ratio for cochlear implants; STOI: short-time objective intelligibility; XM: external microphone; WOLA: weighted overlap and add.

Author's contributions

RA, TvW, and MM conceptualised and analysed the QCQP framework and tuning strategy. RA drafted the manuscript, implemented the algorithms in software, and conducted the experiments. All authors have interpreted the results and reviewed the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The microphone data analysed in the current study as well as audio samples of the processed signals are available at [45]. Further materials are also available from the corresponding author upon request.

[13] Audio samples are also uploaded for the case when the SPP was computed on the reference microphone.

References

1. Brandstein, M., Ward, D.B.: Microphone Arrays: Signal Processing, Techniques and Applications. Springer, New York (2001)
2. Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A.: A consolidated perspective on multi-microphone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
3. Vincent, E., Virtanen, T., Gannot, S.: Audio Source Separation and Speech Enhancement. John Wiley & Sons, Ltd (2018)
4. Szurley, J., Bertrand, A., van Dijk, B., Moonen, M.: Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(5), 952–966 (2016)
5. Gößling, N., Doclo, S.: Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field. In: Proc. 2016 Int. Workshop Acoustic Signal Enhancement (IWAENC '18), Tokyo, Japan, pp. 146–150 (2018)
6. Gößling, N., Doclo, S.: RTF-based binaural MVDR beamformer exploiting an external microphone in a diffuse noise field. In: Proc. ITG Conference on Speech Communication, Oldenburg, Germany, pp. 1–5 (2018)
7. Cvijanovic, N., Sadiq, O., Srinivasan, S.: Speech enhancement using a remote wireless microphone. IEEE Trans. on Consumer Electronics **59**(1), 167–174 (2013)
8. Yee, D., Kamkar-Parsi, H., Martin, R., Puder, H.: A Noise Reduction Post-Filter for Binaurally-linked Single-Microphone Hearing Aids Utilizing a Nearby External Microphone. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(1), 5–18 (2017)
9. Ali, R., Bernardi, G., van Waterschoot, T., Moonen, M.: Methods of extending a generalized sidelobe canceller with external microphones. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(9), 1349–1364 (2019)
10. Bertrand, A., Moonen, M.: Robust distributed noise reduction in hearing aids with external acoustic sensor nodes. EURASIP Journal on Advances in Signal Processing **2009** (2009)
11. Hassani, A.: Distributed signal processing algorithms for multi-task wireless acoustic sensor networks. PhD thesis, KU Leuven (October 2017)
12. Markovich-Golan, S., Gannot, S., Cohen, I.: Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks. IEEE Trans. Audio Speech Lang. Process. **21**(2), 343–356 (2013)
13. Capon, J.: High-resolution frequency-wavenumber spectrum analysis. Proc. of the IEEE **57**(8), 1408–1418 (1969)
14. Van Trees, H.L.: Optimum Array Processing. John Wiley & Sons, Ltd (2001)
15. Gannot, S., Burshtein, D., Weinstein, E.: Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
16. Greenberg, J.E., Zurek, P.M.: Evaluation of an adaptive beamforming method for hearing aids. J. Acoust. Soc. Amer. **91**(3), 1662–1676 (1992)
17. Kates, J.M., Weiss, M.R.: A comparison of hearing-aid array-processing techniques. J. Acoust. Soc. Amer. **99**(5), 3138–3148 (1996)
18. Kompis, M., Dillier, N.: Performance of an adaptive beamforming noise reduction scheme for hearing aid applications. I. Prediction of the signal-to-noise-ratio improvement. J. Acoust. Soc. Amer. **109**(3), 1123–1133 (2001)
19. Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., van Dijk, B., van Wieringen, A., Wouters, J.: Speech understanding in background noise with the two-microphone adaptive beamformer BEAM in the Nucleus Freedom Cochlear Implant System. Ear and hearing **28**(1), 62–72 (2007)
20. Cohen, I.: Relative Transfer Function Identification Using Speech Signals. IEEE Trans. Speech Audio Process. **12**(5), 451–459 (2004)
21. Markovich-Golan, S., Gannot, S.: Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method. In: Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '15), Brisbane, Australia, pp. 544–548 (2015)
22. Ali, R., Van Waterschoot, T., Moonen, M.: Integration of a priori and estimated constraints into an MVDR beamformer for speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(12), 2288–2300 (2019)
23. Griffiths, L., Jim, C.: An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. **30**(1), 27–34 (1982)
24. Van Gerven, S., Xie, F.: A comparative study of speech detection methods. Fifth European Conference on Speech, 1095–1098 (1997)
25. Gerkmann, T., Hendriks, R.C.: Noise power estimation based on the probability of speech presence. In: Proc. 2011 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '11), pp. 145–148 (2011)
26. Markovsky, I.: Low Rank Approximation: Algorithms, Implementation, Applications. Springer (2012)
27. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York, NY, USA (2004)
28. Liao, W.C., Luo, Z.Q., Merks, I., Zhang, T.: An effective low complexity binaural beamforming algorithm for hearing aids. Proc. 2015 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '15) (2), 1–5 (2015)
29. Liao, W.C., Hong, M., Merks, I., Zhang, T., Luo, Z.Q.: Incorporating spatial information in binaural beamforming for noise suppression in hearing aids. Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '15), 5733–5737 (2015)
30. Souden, M., Benesty, J., Affes, S.: On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio Speech Lang. Process. **18**(2), 260–276 (2010)
31. Grant, M., Boyd, S.: CVX: Matlab Software for Disciplined Convex Programming, version 2.1. `http://cvxr.com/cvx` (2014)
32. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. In: Blondel, V., Boyd, S., Kimura, H. (eds.) Recent Advances in Learning and Control. Lecture Notes in Control and Information Sciences, pp. 95–110. Springer (2008)
33. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
34. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York, NY, USA (2006)
35. Spriet, A., Moonen, M., Wouters, J.: Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction. Signal Processing **84**(12), 2367–2387 (2004)
36. Doclo, S., Gannot, S., Moonen, M., Spriet, A.: Handbook on Array Processing and Sensor Networks, pp. 269–302. John Wiley & Sons, Ltd (2010). Chap. 10: Acoustic Beamforming for Hearing Aid Applications
37. Crochiere, R.: A weighted overlap-add method of short-time Fourier analysis/Synthesis. IEEE Trans. Acoust., Speech, Signal Process. **28**(1), 99–102 (1980)
38. Veaux, C., Yamagishi, J., MacDonald, K.: CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (2016). `http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html`
39. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. **33**(2), 443–445 (1985)
40. Brookes, M., et al.: Voicebox: Speech processing toolbox for Matlab. Imperial College, London (1997). `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox`
41. Greenberg, J.E., Peterson, P.M., Zurek, P.M.: Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. J. Acoust. Soc. Amer. **94**(5), 3009–3010 (1993)
42. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An Algorithm for Intelligibility Prediction of Time – Frequency Weighted Noisy Speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
43. Santos, J.F., Falk, T.H.: Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 2197–2206 (2014)
44. ANSI-S3.5-1997: American National Standard Methods for calculation of the speech intelligibility index. J. Acoust. Soc. Amer. (June 1997)
45. R.Ali: (2020). `ftp://ftp.esat.kuleuven.be/pub/SISTA/rali/Reports/public_data_mvdrint`