KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Randall Ali, Toon van Waterschoot, Marc Moonen, (2019), Integration of a priori and estimated constraints into an MVDR beamformer for speech enhancement
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher ftp://ftp.esat.kuleuven.be/pub/SISTA/rali/Reports/19-108.pdf
Published version	Klik hier als u tekst wilt invoeren.
Journal homepage	
Author contact	your email randall.ali@esat.kuleuven.be Klik hier als u tekst wilt invoeren.
IR	

(article begins on next page)

Integration of a priori and estimated constraints into an MVDR beamformer for speech enhancement

Randall Ali, Toon van Waterschoot and Marc Moonen

***Abstract*—Conventionally, the single constraint of the minimum variance distortionless response (MVDR) beamformer for speech enhancement has been defined using one of two approaches. Either it is based on a priori assumptions such as microphone characteristics, position, speech source location, and room acoustics, or on a relative transfer function (RTF) vector estimate using a data dependent method. Each approach has its respective merits and drawbacks and a decision usually has to be made between one of the approaches. In this paper, an alternative approach of using an integrated MVDR beamformer is investigated, where both the hard constraints from the two conventional approaches are softened to yield two tuning parameters. It will be shown that this integrated MVDR beamformer can be expressed as a convex combination of the conventional MVDR beamformers, a linearly constrained minimum variance (LCMV) beamformer, and an all-zero vector, with real, positive-valued coefficients. By analysing how the tuning parameters affect these coefficients, two tuning rules for a practical implementation of the integrated MVDR are subsequently proposed. An evaluation with simulated and recorded data demonstrates that the integrated MVDR beamformer can be beneficial as opposed to relying on either of the conventional MVDR beamformers.**

***Index Terms*—Speech Enhancement, Multi-Microphone Noise Reduction, Beamforming, Minimum Variance Distortionless Response (MVDR) Beamformer.**

I. Introduction

Devices equipped with multiple microphones such as teleconferencing systems, assistive hearing devices (hearing aids (HA) and cochlear implants (CI) for instance), and automatic speaker recognition systems, are subject to a degradation in performance as the microphones capture a certain degree of noise in addition to the desired signal. This noise inevitably corrupts the desired signal, resulting in poor audio quality and intelligibility. The task of multi-microphone noise reduction therefore involves the extraction of this desired signal from the corrupted mixture of desired signal and noise, which in turn will restore the performance of the device in question. In the case where the desired signal is a speech signal (as will be considered in this paper), noise reduction is also referred to as speech enhancement. An extensive overview of the development of speech enhancement over the past 70 years is provided in [1], with a comprehensive list of references.

One particular algorithm that has persevered and that is still widely used is known as the linearly constrained minimum variance (LCMV) beamformer [2] [3], whereby multiple constraints are imposed upon the minimisation of the total output power of the beamformer. In this paper, a specific case of the LCMV beamformer is considered, known as the minimum variance distortionless response (MVDR) beamformer [4], whereby only one constraint in the direction of the speech source is imposed upon the minimisation of the total output power of the beamformer. An adaptive implementation of the MVDR beamformer has been presented by Frost [5] and a common practical implementation is the generalised sidelobe canceller (GSC) [6].

Regardless of the practical implementation of the MVDR beamformer, the imposed constraint is usually defined from one of two conventional approaches. In the first approach, the constraint is based on a priori assumptions such as microphone characteristics, position, speech source location, and room acoustics (e.g. no reverberation). Therefore, a fixed model is used to characterise the vector of acoustic transfer functions (ATFs) from the speech source to the microphone array. For instance, it is not uncommon in hearing devices to assume the knowledge of the speech source location [7]–[10].

The second approach for defining the constraint for the MVDR beamformer does not involve any such a priori assumptions and alternatively, a data dependent estimate for the ATF vector, or more commonly the relative transfer function (RTF) vector, is performed [11] [12]. The RTF vector is simply defined as the ATF vector normalised to a reference microphone and the estimate is calculated from the second order statistics of speech and noise signals (i.e. the speech-plus-noise correlation matrix and the noise-only correlation matrix).

Each of the mentioned approaches has its merits and drawbacks and may be more suitable for specific acoustic environments. For instance in the first approach, if the speech source is indeed in the direction as defined by the a priori assumptions and the microphone array is properly calibrated, then a robust solution can be obtained. However, this approach may not be so effective when the assumptions are broken, particularly if there is a mismatch between the a priori assumed direction and the actual speech source location [13]. The second approach of estimating the RTF vector on the other

R.Ali and M. Moonen are with KU Leuven, Dept. of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium (email: {randall.ali, marc.moonen}@esat.kuleuven.be).

T. van Waterschoot is with KU Leuven, Dept. of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium and KU Leuven, Dept. of Electrical Engineering (ESAT-ETC), e-Media Research Lab, Andreas Vesaliusstraat 13, 3000 Leuven, Belgium (email: toon.vanwaterschoot@esat.kuleuven.be).

This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of IWT O&O Project nr. 150432 'Advances in Auditory Implants: Signal Processing and Clinical Aspects', KU Leuven Impulsfonds IMP/14/037, KU Leuven C2-16-00449 'Distributed Digital Signal Processing for Ad-hoc Wireless Local Area Audio Networking', and KU Leuven Internal Funds VES/16/032. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. The scientific responsibility is assumed by its authors.

hand, is not subject to such degradations as it is independent of any such a priori assumptions. Nevertheless, for a reliable performance, it is essential that an accurate distinction is made between periods of speech and non-speech activity (for the proper estimation of the speech-plus-noise correlation matrix and the noise-only correlation matrix). While this distinction can be made fairly accurate in favourable acoustic conditions, in adverse acoustic environments such as for a low signal-to-noise-ratio (SNR) and excessive reverberation, the ability to make such a distinction becomes increasingly difficult [14]. The result is that the MVDR beamformer with an RTF vector estimate in this case can result in an undesirable performance.

Therefore the question remains as to how can the mismatch between an assigned RTF vector and the true RTF vector be minimised, or in other words, how can the problem of spatial robustness be addressed? Solutions to this problem have already been proposed in [15] [16]. In this paper, however, the strategy proposed in [17] is considered, where the constraint from a priori assumptions and the constraint from a data dependent estimate were integrated into one cost function by softening these two hard constraints, thereby yielding two tuning parameters. In [17], only a limited analysis was performed on such a cost function and a solution was presented that did not provide much insight into how such a strategy could be applied in practice.

It is therefore the intention of this work to further analyse the potential of the proposed strategy in [17] both theoretically and experimentally. In this paper, the cost function proposed in [17] is reframed in terms of an a priori RTF vector and an estimated RTF vector and subsequently referred to as an integrated MVDR beamformer. It will be shown that this integrated MVDR beamformer can be expressed as a convex combination of the conventional MVDR beamformers, a linearly constrained minimum variance (LCMV) beamformer, and an all-zero vector, with real, positive-valued coefficients. An analysis of how the tuning parameters affect these coefficients both theoretially and through simulations will also demonstrate that the integrated MVDR encompasses a wide range of speech enhancement filters.

Finally, two tuning rules are proposed for a practical implementation of the integrated MVDR beamformer, which make use of a metric of confidence and the relationship between the integrated MVDR and the speech-distortion-weighted MWF (SDW-MWF). An evaluation with recorded data demonstrates that the integrated MVDR beamformer can indeed be beneficial and more accomodating for changes in the acoustic environment as opposed to relying on either of the conventional MVDR beamformers.

The paper is organised as follows. In Section II, the data model and notation are defined. In Section III, a brief review of the two conventional approaches to define the constraints of the MVDR beamformer, and an LCMV approach is provided. In Section IV, the integrated approach is introduced and analysed. In Section V, two tuning rules for a practical implementation are proposed. In Section VI, the integrated MVDR approach is evaluated with simulated and recorded data from an office room. Finally in Section VII, conclusions are drawn.

II. Data Model

A speech enhancement system is considered with an array of M microphones that receives a signal consisting of a desired speech signal that is corrupted by noise in a reverberant environment. In the short-time Fourier transform (STFT) domain, the received signal at frequency, k, and time frame, l, is represented as:

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

where (dropping the dependency on k and l for brevity) $\mathbf{y} = [\mathrm{y}_1\ \mathrm{y}_2\ \dots \mathrm{y}_M]^T$ is a vector containing the respective microphone signals, $\mathbf{x}$ is the speech contribution, represented by s_1, the speech signal in the first microphone of the array, filtered with $\mathbf{h} = [1\ \mathrm{h}_2\ \dots \mathrm{h}_M]^T$, the RTF vector for the array (with the first microphone used as the reference, which translates to the first component of $\mathbf{h}$ being equal to 1). $\mathbf{n} = [\mathrm{n}_1\ \mathrm{n}_2\ \dots \mathrm{n}_M]^T$ is the noise contribution.

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

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

where $\mathbb{E}\{.\}$ is the expectation operator and ^{H} is the Hermitian transpose. It is assumed that the speech signal is uncorrelated with the noise signal, and hence $\mathbf{R}_{\mathbf{yy}} = \mathbf{R}_{\mathbf{xx}} + \mathbf{R}_{\mathbf{nn}}$.

The estimate of the speech component in the first microphone of the array, s_1, is then obtained through the linear filtering of the microphone signals, such that:

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

where $\mathbf{w} = [\mathrm{w}_1\ \mathrm{w}_2\ \dots \mathrm{w}_M]^T$ is the complex-valued filter (beamformer) to be designed.

III. MVDR and LCMV review

The minimum power distortionless response (MPDR) beamformer [18] minimises the total output power while preserving the received speech signal in accordance with a constraint that would ideally result in a distortionless response. Using $\mathbf{h}$ from (1) as the constraint, the MPDR problem is given by:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^H\mathbf{R}_{\mathbf{yy}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^H\mathbf{h} = 1 \end{aligned} \tag{4}$$

For a correct constraint, this total output power reduces to the total noise power, so that $\mathbf{R}_{\mathbf{yy}}$ in (4) can be replaced by $\mathbf{R}_{\mathbf{nn}}$ [19]. In practice, however, $\mathbf{R}_{\mathbf{nn}}$ has to be estimated, which is typically done by recursive averaging, along with a voice activity detector (VAD) [20] or a speech presence probability (SPP) estimator [21]. The optimisation problem in (4) can then be replaced accordingly by:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^H\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^H\mathbf{h} = 1 \end{aligned} \tag{5}$$

where the MPDR beamformer is now referred to as the MVDR beamformer and given by:

$$\mathbf{w} = \frac{\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{h}}{\mathbf{h}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{h}} \tag{6}$$

where $\hat{\mathbf{R}}_{\mathbf{nn}}$ is the estimate of $\mathbf{R}_{\mathbf{nn}}$. Finally, the other critical component for the MVDR, the RTF vector, $\mathbf{h}$, is also not typically known in practice. In sections III-A and III-B that follow, strategies are discussed for computing this RTF vector. In section III-C, an LCMV approach using these different RTF vector approximations is also discussed.

A. MVDR with a priori assumptions (MVDR-APR)

One option replacing $\mathbf{h}$ in (6) is to use an a priori assumed RTF vector, $\widetilde{\mathbf{h}} = [1\ \widetilde{\mathrm{h}}_2\ \ldots \widetilde{\mathrm{h}}_{\mathrm{M}}]^T$ as an approximation to $\mathbf{h}$. $\widetilde{\mathbf{h}}$ can be based on a priori assumptions such as microphone characteristics, position, speaker location, and room acoustics (e.g. no reverberation). Similar to (6), the optimal noise reduction filter is then given by:

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

which will be referred to as the MVDR-APR. The speech estimate in the first microphone of the array is subsequently calculated as:

$$\widetilde{\mathrm{z}}_1 = \widetilde{\mathbf{w}}^H\mathbf{y} \tag{8}$$

It is noted that the quantities denoted as $(\,\widetilde{\cdot}\,)$ will be associated with a priori assumed quantities.

In the case where the assumptions are satisfied, for instance if the speech source truly lies in the direction defined by $\widetilde{\mathbf{h}}$, the MVDR-APR can be regarded as robust in adverse acoustic conditions. In other scenarios, however, it can be expected that the speech signal does not always adhere to the assumptions, and hence the MVDR-APR will degrade in performance and potentially suffer from distortions.

B. MVDR with an estimated RTF vector (MVDR-EST)

In order to accommodate for the shortcomings of defining an RTF vector based on a priori assumptions, methods for alternatively estimating the RTF vector from the estimated speech-plus-noise correlation matrix, $\hat{\mathbf{R}}_{\mathbf{yy}}$, and the estimated noise-only correlation matrix, $\hat{\mathbf{R}}_{\mathbf{nn}}$, have been pursued [11] [12]. This estimated RTF vector will therefore not rely on any a priori assumptions and can be used to enhance the speech in more practical scenarios, such as when the speech source is not in the direction specified by the a priori assumptions. Out of these methods, the method of covariance whitening or equivalently that which involves a Generalised Eigenvalue Decomposition (GEVD) has resulted in superior performance [22].

The GEVD of the matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}}\}$ follows as:

$$\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{U}}^{-1} \tag{9}$$

where $\hat{\mathbf{\Sigma}}$ is a diagonal matrix of the generalised eigenvalues, $\{\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{\mathrm{M}}\}$, ordered such that $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \ldots \hat{\sigma}_{\mathrm{M}}$, and $\hat{\mathbf{U}}$ contains the corresponding generalised eigenvectors. As the generalised eigenvectors are defined up to a scaling, it is assumed without loss of generality that $\hat{\mathbf{U}}^H\hat{\mathbf{R}}_{\mathbf{nn}}\hat{\mathbf{U}} = \mathbf{I}_M$, where $\mathbf{I}_M$ is an $M \times M$ identity matrix. Hence, the GEVD is also equivalent to the following joint diagonalisation of $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$:

$$\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{y}}\hat{\mathbf{Q}}^H;\ \hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{Q}}\hat{\mathbf{Q}}^H \tag{10}$$

where $\hat{\mathbf{Q}} = \hat{\mathbf{U}}^{-H}$ is an invertible matrix, and $\hat{\mathbf{\Sigma}}_{\mathbf{y}} = \hat{\mathbf{\Sigma}}$. The corresponding estimate of the speech-only correlation matrix can then be given as:

$$\hat{\mathbf{R}}_{\mathbf{xx}} = \hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{Q}}\underbrace{(\hat{\mathbf{\Sigma}} - \mathbf{I}_M)}_{\hat{\mathbf{\Sigma}}_{\mathbf{x}}}\hat{\mathbf{Q}}^H \tag{11}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ is a diagonal matrix of eigenvalues, $\{\hat{\sigma}_{\mathrm{x}1}, \hat{\sigma}_{\mathrm{x}2}, \ldots, \hat{\sigma}_{\mathrm{xM}}\}$, ordered such that $\hat{\sigma}_{\mathrm{x}1} \geq \hat{\sigma}_{\mathrm{x}2} \geq \ldots \hat{\sigma}_{\mathrm{xM}}$. With a rank-1 approximation to $\hat{\mathbf{R}}_{\mathbf{xx}}$, the estimated RTF vector is then:

$$\hat{\mathbf{h}} = \frac{\hat{\mathbf{Q}}\mathbf{e_1}}{\mathbf{e_1}^T\hat{\mathbf{Q}}\mathbf{e_1}} \tag{12}$$

where the $(M \times 1)$ vector $\mathbf{e_1} = [1\,0\ldots0]^T$. This estimated RTF vector can now be used as the approximation to $\mathbf{h}$ for the MVDR beamformer defined in (6), and is given by:

$$\hat{\mathbf{w}} = \frac{\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}}}{\hat{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}}} \tag{13}$$

which will be referred to as the MVDR-EST. The speech estimate in the first microphone of the array is subsequently calculated as:

$$\hat{\mathrm{z}}_1 = \hat{\mathbf{w}}^H\mathbf{y} \tag{14}$$

It is noted that the quantities denoted as $(\,\hat{\cdot}\,)$ will be associated with estimated quantities.

While such an MVDR beamformer that uses an RTF vector estimate can be effective in many practical scenarios, it is important to acknowledge that a critical requirement is to distinguish between the speech-plus-noise and noise-only periods for an accurate estimation of the respective correlation matrices.

C. LCMV

Given the limitations of the previous MVDR beamformers, it is expected that neither of them would be able to accomodate for dynamic acoustic scenarios. Therefore, a first attempt to resolve this issue would be to consider an LCMV beamformer, where a distortionless response from both the a priori assumed RTF, $\widetilde{\mathbf{h}}$, and the estimated RTF $\hat{\mathbf{h}}$ can be preserved, i.e. multiple hard constraints can be imposed on the minimisation of the total noise power:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^H\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{C}^H\mathbf{w} = \mathbf{b} \end{aligned} \tag{15}$$

where

$$\mathbf{C} = [\widetilde{\mathbf{h}}\ \ \hat{\mathbf{h}}];\ \ \mathbf{b} = [1\ \ 1]^T \tag{16}$$

The corresponding solution to this equality constrained least squares problem can be found by defining the dual function with a Lagrange multiplier vector, $\boldsymbol{\nu}$, and employing the

Karush-Kuhn-Tucker (KKT) conditions to solve the following linear set of equations:

$$\begin{bmatrix} \hat{\mathbf{R}}_{\mathbf{nn}} & \mathbf{C} \\ \mathbf{C}^H & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix} \tag{17}$$

where the left-hand side coefficient matrix is known as the KKT matrix [23]. The result is the well-known equation:

$$\mathbf{w}_{\mathrm{lcmv}} = \hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\,\mathbf{C}\,\mathbf{V}^{-1}\mathbf{b} \tag{18}$$

where

$$\mathbf{V} = [\mathbf{C}^H\,\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\,\mathbf{C}] = \begin{bmatrix} k_{aa} & k_{ab} \\ k_{ba} & k_{bb} \end{bmatrix} \tag{19}$$

and

$$\begin{aligned} k_{aa} &= \widetilde{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}};\;\; k_{bb} = \hat{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}} \\ k_{ab} &= \widetilde{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}};\;\; k_{ba} = \hat{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}} \end{aligned} \tag{20}$$

Upon expansion of (18), it can also be observed that the $\mathbf{w}_{\mathrm{lcmv}}$ can be expressed as a linear combination of the MVDR-APR, $\widetilde{\mathbf{w}}$, and the MVDR-EST, $\hat{\mathbf{w}}$:

$$\mathbf{w}_{\mathrm{lcmv}} = \left[\frac{k_{aa}(k_{bb}-k_{ab})}{k_{aa}k_{bb}-k_{ab}k_{ba}}\right]\widetilde{\mathbf{w}} \quad + \left[\frac{k_{bb}(k_{aa}-k_{ba})}{k_{aa}k_{bb}-k_{ab}k_{ba}}\right]\hat{\mathbf{w}} \tag{21}$$

where the denominator, $(k_{aa}k_{bb} - k_{ab}k_{ba})$ is the determinant of $\mathbf{V}$, real-valued, and always ≥ 0 due to the Cauchy–Schwarz inequality. In the case of a single speaker, while this LCMV solution can potentially provide a distortionless response if $\widetilde{\mathbf{h}}$ or $\hat{\mathbf{h}}$ is a suitable approximation to the true RTF vector, less noise will be reduced as both constraints are always active.

Caution must however be taken with such an approach as the constraints in $\mathbf{C}$ can become redundant in scenarios when $\widetilde{\mathbf{h}}$ and $\hat{\mathbf{h}}$ are linearly dependent or even approximately so, i.e. $\hat{\mathbf{h}} \approx \delta\widetilde{\mathbf{h}}$, where δ is a scalar. This is possible for instance when estimation is reliable and the speech is in the a priori assumed direction. Even though such a case can be deemed as a "fortunate" circumstance of events, the consequence is that $\mathbf{V}$ will become ill-conditioned. On closer observation of (17), it can be seen that $-\mathbf{V}$ is the Schur complement of the $\mathbf{0}$-block in the KKT matrix [23] and the Lagrange multiplier vector is given as $\boldsymbol{\nu} = -\mathbf{V}^{-1}\mathbf{b}$. Therefore, it can be concluded that when $\hat{\mathbf{h}} \approx \delta\widetilde{\mathbf{h}}$, the dual problem is not well defined.

As an extreme case, if $\hat{\mathbf{h}} = \widetilde{\mathbf{h}}$ ($\delta = 1$), the weights of $\widetilde{\mathbf{w}}$ and $\hat{\mathbf{w}}$ in (21) will be indeterminate forms of $\frac{0}{0}$. A distinct example is as follows: for $\widetilde{\mathbf{h}} = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^T$ and $\hat{\mathbf{h}} = \begin{bmatrix}1 & 1+r\cos\theta & 1+r\sin\theta\end{bmatrix}^T$ (with r being some distance and $\theta \in [0,\, 2\pi]$), it is readily shown that the hard constraints $\mathbf{C}^T\mathbf{w} = \begin{bmatrix}1 & 1\end{bmatrix}^T$ are equivalent with $\begin{bmatrix}1 & 1 & 1 \\ 0 & \cos\theta & \sin\theta\end{bmatrix}\mathbf{w} = \begin{bmatrix}1 & 0\end{bmatrix}^T$ where the r cancels out. As a result, even for infinitesimally small r (i.e. for $\hat{\mathbf{h}}$ arbitrarily close to $\widetilde{\mathbf{h}}$), the LCMV beamformer as well as its noise reduction performance will be independent of r but dependent on θ, which is indeed undesirable.

One solution to this problem would be the diagonal loading of $\mathbf{V}$, which has been proven to avoid instabilities in the LCMV context [24]. The procedure consists of adding a scaled identity, $\epsilon\mathbf{I}$, to $\mathbf{V}$, resulting in a regularised version of (21):

$$\mathbf{w}_{\mathrm{lcmv}}^{\epsilon} = \left[\frac{k_{aa}(k_{bb}^{\epsilon}-k_{ab})}{k_{aa}^{\epsilon}k_{bb}^{\epsilon}-k_{ab}k_{ba}}\right]\widetilde{\mathbf{w}} + \left[\frac{k_{bb}(k_{aa}^{\epsilon}-k_{ba})}{k_{aa}^{\epsilon}k_{bb}^{\epsilon}-k_{ab}k_{ba}}\right]\hat{\mathbf{w}} \tag{22}$$

where $k_{aa}^{\epsilon} = k_{aa} + \epsilon$ and $k_{bb}^{\epsilon} = k_{bb} + \epsilon$. When $\hat{\mathbf{h}} = \delta\widetilde{\mathbf{h}}$ (22) will tend toward a scaled version of the MVDR-APR provided that $\epsilon << k_{aa}$.

IV. Integrated approach

A. Formulation

The fact that the MVDR-APR, the MVDR-EST, and LCMV beamformers impose hard constraints is the underlying reason why these approaches may still have a limited performance in dynamic acoustic scenarios. Consequently, an alternative approach may be considered such as that proposed in [17] where these hard constraints are softened. Such a beamformer can then merge the benefits of the MVDR-APR and the MVDR-EST beamformers in order to yield a more versatile one. For instance, if the speech source moves outside of the a priori assumed direction, more weight can be given to the estimated RTF vector to accommodate for the loss in performance that would otherwise result from using the a priori assumed RTF vector alone. On the other hand, if the estimated RTF vector becomes unreliable, less weight can be given to it and the system can revert to using the a priori assumed RTF vector, which may have an improved performance if the speech source is indeed close to the a priori assumed direction.

Similar to [17], if the case is considered where $\widetilde{\mathbf{h}}$ is defined according to some assumptions as in Section III-A and $\hat{\mathbf{h}}$ is estimated as in III-B, an integrated MVDR cost function can be given as:

$$\min_{\mathbf{w}} \quad \mathbf{w}^H\hat{\mathbf{R}}_{\mathbf{nn}}\mathbf{w} + \alpha\,|\mathbf{w}^H\widetilde{\mathbf{h}} - 1|^2 + \beta\,|\mathbf{w}^H\hat{\mathbf{h}} - 1|^2 \tag{23}$$

where $\alpha \in [0,\, \infty)$ and $\beta \in [0,\, \infty)$ are tuning parameters that control how much of the respective RTF vectors are imposed. This cost function is simply the combination of the MVDR-APR and MVDR-EST cost function, except that the constraints have been softened by α and β. The solution to (23) is then given by:

$$\mathbf{w}_{\mathrm{int}} = (\hat{\mathbf{R}}_{\mathbf{nn}} + \alpha\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H + \beta\hat{\mathbf{h}}\hat{\mathbf{h}}^H)^{-1}(\alpha\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H + \beta\hat{\mathbf{h}}\hat{\mathbf{h}}^H)\mathbf{e_1} \tag{24}$$

This will be referred to as the MVDR-INT, and can be re-written as:

$$\mathbf{w}_{\mathrm{int}} = (\hat{\mathbf{R}}_{\mathbf{nn}} + \mathbf{C}\boldsymbol{\Gamma}\mathbf{C}^H)^{-1}(\mathbf{C}\boldsymbol{\Gamma}\mathbf{C}^H)\mathbf{e_1} \tag{25}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}\{\alpha, \beta\}$. Applying the matrix inversion lemma to the inverse in (25) yields:

$$\begin{aligned} (\hat{\mathbf{R}}_{\mathbf{nn}} + \mathbf{C}\boldsymbol{\Gamma}\mathbf{C}^H)^{-1} = \hat{\mathbf{R}}_{\mathbf{nn}}^{-1} - \hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{C}\big[\boldsymbol{\Gamma}^{-1} \\ + (\mathbf{C}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\mathbf{C})\big]^{-1}\mathbf{C}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1} \end{aligned} \tag{26}$$

Substitution of (26) into (25) eventually results in a simple expression for the MVDR-INT:

$$\mathbf{w}_{\text{int}} = \frac{\alpha k_{aa}[1+\beta(k_{bb}-k_{ab})]}{\mathrm{D}}\widetilde{\mathbf{w}} + \frac{\beta k_{bb}[1+\alpha(k_{aa}-k_{ba})]}{\mathrm{D}}\hat{\mathbf{w}} \tag{27}$$

where:

$$\mathrm{D} = \alpha k_{aa} + \beta k_{bb} + \alpha\beta(k_{aa}k_{bb} - k_{ab}k_{ba}) + 1 \tag{28}$$

It can now be observed from (27) that the MVDR-INT beamformer is a linear combination of the MVDR-APR and the MVDR-EST beamformers, whose complex-valued weightings are defined by α, β, and the constants from (20).

In the limiting case where $\alpha \to \infty$ and $\beta \to \infty$, (23) is equivalent to (15) with the LCMV solution given by (21). As was discussed in Section III-C, in scenarios where $\hat{\mathbf{h}} \approx \delta\widetilde{\mathbf{h}}$, the LCMV suffers from an ill-conditioning. Hence the limit as $(\alpha, \beta, \hat{\mathbf{h}}) \to (\infty, \infty, \delta\widetilde{\mathbf{h}})$ (or equivalently $(\frac{1}{\alpha}, \frac{1}{\beta}, \hat{\mathbf{h}}) \to (0, 0, \delta\widetilde{\mathbf{h}})$) will also result in an ill-conditiong problem in function of (27). For the LCMV, this problem was resolved with diagonal loading resulting in (22). Hence the ill-conditioning for the MVDR-INT can also be resolved by regularising (27) as follows:

$$\mathbf{w}^{\epsilon}_{\text{int}} = \frac{\alpha k_{aa}[1+\beta(k^{\epsilon}_{bb}-k_{ab})]}{\mathrm{D}^{\epsilon}}\widetilde{\mathbf{w}} + \frac{\beta k_{bb}[1+\alpha(k^{\epsilon}_{aa}-k_{ba})]}{\mathrm{D}^{\epsilon}}\hat{\mathbf{w}} \tag{29}$$

where

$$\mathrm{D}^{\epsilon} = \alpha k_{aa} + \beta k_{bb} + \alpha\beta(k^{\epsilon}_{aa}k^{\epsilon}_{bb} - k_{ab}k_{ba}) + 1 \tag{30}$$

For $\alpha \to \infty$ and $\beta \to \infty$, (29) then indeed reduces to (22).

In order to realise practical values for α and β to analyse the MVDR-INT further (for instance it is ambiguous as to what exactly a practical value is for $\alpha \to \infty$ or $\beta \to \infty$), it is suggested to make the following normalisations:

$$\alpha = \frac{\bar{\alpha}}{k_{aa}}; \qquad \beta = \frac{\bar{\beta}}{k_{bb}} \tag{31}$$

where $\bar{\alpha}$ and $\bar{\beta}$ are still real-valued constants. Substituting (31) into (29) results in:

$$\bar{\mathbf{w}}^{\epsilon}_{\text{int}} = \frac{\bar{\alpha}[1+\bar{\beta}(\frac{k^{\epsilon}_{bb}-k_{ab}}{k_{bb}})]}{\bar{\mathrm{D}}^{\epsilon}}\widetilde{\mathbf{w}} + \frac{\bar{\beta}[1+\bar{\alpha}(\frac{k^{\epsilon}_{aa}-k_{ba}}{k_{aa}})]}{\bar{\mathrm{D}}^{\epsilon}}\hat{\mathbf{w}} \tag{32}$$

where

$$\bar{\mathrm{D}}^{\epsilon} = \bar{\alpha} + \bar{\beta} + \bar{\alpha}\bar{\beta}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb} - k_{ab}k_{ba}}{k_{aa}k_{bb}}) + 1 \tag{33}$$

While it is evident that the MVDR-INT is a linear combination of the MVDR-APR and the MVDR-EST beamformers, it should be highlighted that by the substitution of (22), (32) can also be expressed as:

$$\bar{\mathbf{w}}^{\epsilon}_{\text{int}} = \frac{\bar{\alpha}}{\bar{\mathrm{D}}^{\epsilon}}\widetilde{\mathbf{w}} + \frac{\bar{\beta}}{\bar{\mathrm{D}}^{\epsilon}}\hat{\mathbf{w}} + \frac{\bar{\alpha}\bar{\beta}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}{\bar{\mathrm{D}}^{\epsilon}}\mathbf{w}^{\epsilon}_{\text{lcmv}} + \frac{1}{\bar{\mathrm{D}}^{\epsilon}}\mathbf{w}_{\text{z}} \tag{34}$$

where $\mathbf{w}_{\text{z}}$ is the all-zero $(M \times 1)$ vector. (34) reveals that the MVDR-INT is a convex combination of the MVDR-APR beamformer, the MVDR-EST beamformer, the LCMV beamformer, and the all-zero vector. Furthermore, as opposed to (32) which consists of complex-valued coefficients, in (34), the coefficients are all positive and real-valued since $(k^{\epsilon}_{aa}k^{\epsilon}_{bb} - k_{ab}k_{ba}) > 0$ from the Cauchy–Schwarz inequality and the regularisation that was introduced. The MVDR-INT can therefore be truly regarded as a more global and versatile beamformer that encompasses wide range of filters bounded by the MVDR-APR, MVDR-EST, LCMV, and the all-zero vector. Consequently, it suggests that (34) can be tuned for an optimal performance depending on the acoustic environment. As such, some interesting cases of different tunings of (34) are discussed in the following.

B. Limiting cases of the tuning parameters

In terms of a contingency strategy as in [25], the fall back mechanism will be to revert to an MVDR-APR when the MVDR-EST becomes unreliable. One means by which this can be achieved is to always keep the a priori constraint active, i.e. set $\bar{\alpha} \to \infty$ while only tuning $\bar{\beta}$. This additionally simplifies the practical implementation of the filter as only one tuning parameter needs to be considered. Considering the limit of $\bar{\alpha} \to \infty$ in (34) results in:

$$\lim_{\bar{\alpha}\to\infty} \bar{\mathbf{w}}^{\epsilon}_{\text{int}} = \frac{1}{1+\bar{\beta}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\widetilde{\mathbf{w}} + \frac{\bar{\beta}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}{1+\bar{\beta}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\mathbf{w}^{\epsilon}_{\text{lcmv}} \tag{35}$$

which is a convex combination of the MVDR-APR and the LCMV. Clearly, if $\bar{\beta} = 0$, (35) reverts to (7), satisfying the notion of a contingency strategy. In scenarios where the a priori assumptions are not satisfied, increasing values of $\bar{\beta}$ in (35) will attempt to maintain a distortionless response and tend toward the LCMV. This however, may be at the cost of having less noise reduced as $\widetilde{\mathbf{h}}$ is always active. Nevertheless, this compromised performance will be an improvement as opposed to simply using only an MVDR-APR, provided that $\hat{\mathbf{h}}$ is accurate.

A contrary filter can also be derived, where alternatively all of the weight is placed on the estimated RTF vector, i.e. $\bar{\beta} \to \infty$. Considering this limit in (34) results in:

$$\lim_{\bar{\beta}\to\infty} \bar{\mathbf{w}}^{\epsilon}_{\text{int}} = \frac{1}{1+\bar{\alpha}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\hat{\mathbf{w}} + \frac{\bar{\alpha}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}{1+\bar{\alpha}(\frac{k^{\epsilon}_{aa}k^{\epsilon}_{bb}-k_{ab}k_{ba}}{k_{aa}k_{bb}})}\mathbf{w}^{\epsilon}_{\text{lcmv}} \tag{36}$$

which is now a convex combination of the MVDR-EST and the LCMV. If $\bar{\alpha} = 0$, (36) reverts to (13), indicating that this filter has the potential to perform better than the MVDR-APR regardless of the speech source location, provided that $\hat{\mathbf{h}}$ is accurate. If $\hat{\mathbf{h}}$ is not accurate as it will inevitably become in some scenarios, then $\bar{\alpha}$ can be increased to introduce more of the a priori assumed RTF vector and the MVDR-INT will tend toward the LCMV beamformer. If the speech source is such that it adheres to some a priori assumed conditions, for instance, if it is close to the direction defined by $\widetilde{\mathbf{h}}$, then increasing $\bar{\alpha}$ can result in a filter that has a better performance than that of the MVDR-EST when $\hat{\mathbf{h}}$ is inaccurate.

C. Equivalence to the Speech Distortion Weighted MWF

In the previous sections, it was demonstrated that the MVDR-INT encompasses a number of different types of spatial filters. It is worthwhile to highlight that the MVDR-INT can also incorporate a postfilter on top of these spatial filters for further noise suppression. This can be observed by considering the case where either $\bar{\alpha} = 0$ or $\bar{\beta} = 0$, which consequently results in an equivalence between the MVDR-INT and the speech-distortion-weighted MWF (SDW-MWF) [26]. In fact, the ability of the MVDR-INT to also incorporate a postfilter is also evident from (34) as the all-zero vector, $\mathbf{w}_{\mathrm{z}}$ is also included as one of the terms in the convex combination.

In [22], it was shown that the solution for the GEVD-based SDW-MWF is given by:

$$\mathbf{w}_{\mathrm{sdw-mwf}} = (\mathbf{R}_{\mathbf{xx,r1}} + \mu\hat{\mathbf{R}}_{\mathbf{nn}})^{-1}\mathbf{R}_{\mathbf{xx,r1}}\mathbf{e_1} \tag{37}$$

where $\mu \in [0\ \infty)$ is a parameter to trade-off the amount of speech distortion against the amount of noise reduction, and $\mathbf{R}_{\mathbf{xx,r1}}$ is a rank-1 approximation to the speech-only correlation matrix, $\mathbf{R}_{\mathbf{xx}}$. In the case of an a priori assumed RTF vector, $\mathbf{R}_{\mathbf{xx,r1}} = \widetilde{\mathbf{R}}_{\mathbf{xx,r1}} = \hat{\Phi}_{\mathrm{s1}}\widetilde{\mathbf{h}}\widetilde{\mathbf{h}}^H$, whereas for the estimated RTF vector, $\mathbf{R}_{\mathbf{xx,r1}} = \hat{\mathbf{R}}_{\mathbf{xx,r1}} = \hat{\Phi}_{\mathrm{s1}}\hat{\mathbf{h}}\hat{\mathbf{h}}^H$, where $\hat{\Phi}_{\mathrm{s1}}$ is an estimated speech power in the reference microphone. Considering either of these cases, it is well known that (37) can be decomposed into [19]:

$$\mathbf{w}_{\mathrm{sdw-mwf}} = \begin{cases} \frac{\hat{\Phi}_{\mathrm{s1}}}{\hat{\Phi}_{\mathrm{s1}}+\mu(\widetilde{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}})^{-1}}\ \widetilde{\mathbf{w}} & \text{if } \mathbf{R}_{\mathbf{xx,r1}} = \widetilde{\mathbf{R}}_{\mathbf{xx,r1}} \\ \frac{\hat{\Phi}_{\mathrm{s1}}}{\hat{\Phi}_{\mathrm{s1}}+\mu(\hat{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}})^{-1}}\ \hat{\mathbf{w}} & \text{if } \mathbf{R}_{\mathbf{xx,r1}} = \hat{\mathbf{R}}_{\mathbf{xx,r1}} \end{cases} \tag{38}$$

The conditions of $\mathbf{R}_{\mathbf{xx,r1}} = \widetilde{\mathbf{R}}_{\mathbf{xx,r1}}$ and $\mathbf{R}_{\mathbf{xx,r1}} = \hat{\mathbf{R}}_{\mathbf{xx,r1}}$ are equivalent to either $\bar{\beta} = 0$ or $\bar{\alpha} = 0$ respectively, hence for the MVDR-INT in (34), it can be deduced that:

$$\bar{\mathbf{w}}_{\mathrm{int}}^{\epsilon} = \begin{cases} \frac{\bar{\alpha}}{\bar{\alpha}+k_{aa}(\widetilde{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\widetilde{\mathbf{h}})^{-1}}\ \widetilde{\mathbf{w}} & \text{if } \bar{\beta} = 0 \\ \frac{\bar{\beta}}{\bar{\beta}+k_{bb}(\hat{\mathbf{h}}^H\hat{\mathbf{R}}_{\mathbf{nn}}^{-1}\hat{\mathbf{h}})^{-1}}\ \hat{\mathbf{w}} & \text{if } \bar{\alpha} = 0 \end{cases} \tag{39}$$

Upon comparing (38) and (39), it is observed that both solutions are indeed equivalent provided that:

$$\bar{\alpha} = \frac{\hat{\Phi}_{\mathrm{s1}}k_{aa}}{\mu} \ \text{ when } \ \bar{\beta} = 0 \tag{40}$$

$$\bar{\beta} = \frac{\hat{\Phi}_{\mathrm{s1}}k_{bb}}{\mu} \ \text{ when } \ \bar{\alpha} = 0 \tag{41}$$

A summary of the various tunings for $\bar{\alpha}$ and $\bar{\beta}$ as applied to (34) and the corresponding speech enhancement strategies is presented in Table I.

TABLE I
TABLE 1: SUMMARY OF SPEECH ENHANCEMENT STRATEGIES RESULTING FROM (34) FOR DIFFERENT TUNINGS OF $\bar{\alpha}$ AND $\bar{\beta}$.

Tuning Parameters	Speech Enhancement Strategy	Equation
$\bar{\alpha} \to \infty$, $\bar{\beta} = 0$	MVDR-APR	(7)
$\bar{\alpha} = 0$, $\bar{\beta} \to \infty$	MVDR-EST	(13)
$\bar{\alpha} \to \infty$, $\bar{\beta} \to \infty$	LCMV	(22)
$\bar{\alpha} \to \infty$, $\bar{\beta} \in [0, \ \infty)$	MVDR-APR + LCMV	(35)
$\bar{\alpha} \in [0, \ \infty)$, $\bar{\beta} \to \infty$	MVDR-EST + LCMV	(36)
$\bar{\alpha} = 0$, $\bar{\beta} \in [0, \ \infty)$	SDW-MWF	(39)
$\bar{\beta} = 0$, $\bar{\alpha} \in [0, \ \infty)$	SDW-MWF	(39)

V. TUNING STRATEGY

Following the theoretical anaylysis of the MVDR-INT, the next issue to be addressed is how should the parameters $\bar{\alpha}$ and $\bar{\beta}$ be tuned in a practical situation? In this section two potential tuning rules are proposed which make use of a metric of confidence as well as the relationship between the MVDR-INT and the SDW-MWF. The dependence of the quantities on time is reintroduced in this section to emphasize that $\bar{\alpha}$ and $\bar{\beta}$ are updated in each time frame (the dependence on frequency is still omitted as all frequencies are updated in the same manner).

A. Metric of confidence

Firstly, a metric of confidence is required to indicate how accurate the RTF vector was estimated (i.e. the accuracy of $\hat{\mathbf{h}}(l)$), which can be subsequently used to assign more or less weight to $\bar{\beta}(l)$. One potential option for such a metric can be the generalised eigenvalues of $\hat{\mathbf{\Sigma}}(l)$ from (9), or equivalently, $\hat{\mathbf{\Sigma}}_{\mathbf{y}}(l)$ from (10).

As only a single speaker is considered, and a rank-1 speech-only correlation matrix is assumed, it is only the principal generalised eigenvalue, $\hat{\sigma}_1(l)$, and its corresponding generalised eigenvector that is of interest. $\hat{\sigma}_1(l)$ can in fact, be interpreted as an a posteriori SNR measure, and hence it is expected that higher values of this a posteriori SNR measure will correspond to a more accurate estimate of the RTF vector. In two extreme cases of voice activity detection, $\hat{\sigma}_1(l)$ will be given as (recall (11)):

$$\hat{\sigma}_1(l) \approx \begin{cases} \sigma_{\mathrm{x1}}(l) + 1 & \text{for perfect VAD} \\ \frac{1}{\sigma_{\mathrm{x1}}(l)+1} & \text{worst case imperfect VAD} \end{cases} \tag{42}$$

where $\sigma_{\mathrm{x1}}(l)$ is the expected (not estimated) eigenvalue from the rank-1 speech-only correlation matrix. In the case of the perfect VAD, where $\hat{\mathbf{R}}_{\mathbf{yy}} \approx \mathbb{E}\{\mathbf{y}\mathbf{y}^H\}$ and $\hat{\mathbf{R}}_{\mathbf{nn}} \approx \mathbb{E}\{\mathbf{n}\mathbf{n}^H\}$, larger values of $\hat{\sigma}_1(l)$, i.e. $\hat{\sigma}_1(l) \gg 1$ correspond to the situation where $\sigma_{\mathrm{x1}}(l) \gg 1$, indicative that speech was present and correctly classified, which would lead to an accurately estimated RTF vector. On the other hand for a worst case imperfect VAD, where $\hat{\mathbf{R}}_{\mathbf{yy}} \approx \mathbb{E}\{\mathbf{n}\mathbf{n}^H\}$ and $\hat{\mathbf{R}}_{\mathbf{nn}} \approx \mathbb{E}\{\mathbf{y}\mathbf{y}^H\}$, since the speech is inaccurately classified, regardless the value of $\sigma_{\mathrm{x1}}(l)$, $\hat{\sigma}_1(l) \leq 1$. It can also be observed that in general, when the speech is classified as noise, smaller values of $\hat{\sigma}_1(l)$ will be expected. In the case where the noise is classified as speech, this can result in larger values of $\hat{\sigma}_1(l)$ but is however, not as detrimental to overall performance in comparison to the case when speech is classified as noise [27].

A logisitic function can then be used to map $\hat{\sigma}_1(l)$ to a value, $\mathrm{F}(l) \in [0\ \ 1]$, where 0 indicates a low confidence in the accuracy of the estimated RTF and 1 a high confidence in the

accuracy of the estimated RTF. The relationship between $\mathrm{F}(l)$ and $\hat{\sigma}_1(l)$ is given by:

$$\mathrm{F}(l) = \frac{1}{1 + e^{-\rho(\hat{\sigma}_1(l) - \sigma_t)}} \tag{43}$$

where ρ controls the gradient of the transition from 0 to 1, and σ_t is a threshold generalised eigenvalue, beyond which $\mathrm{F}(l) \rightarrow 1$. Therefore, using higher values for σ_t would correspond to a more conservative approach where $\mathrm{F}(l)$ would tend mostly to 0 and lower values for σ_t would correspond to a more aggressive approach where $\mathrm{F}(l)$ would tend mostly to 1.

B. Tuning rules

1) SDW tuning rule: The relationship between the MVDR-INT and the SDW-MWF from Section IV-C suggests that (40) and (41) can be used for tuning $\bar{\alpha}(l)$ and $\bar{\beta}(l)$ respectively. However, in order to control the respective weights in accordance with how well the RTF vector has been estimated, the previously proposed metric of confidence can be incorporated such that $\bar{\alpha}(l)$ and $\bar{\beta}(l)$ are tuned as follows:

$$\bar{\alpha}^{\mathrm{sdw}}(l) = (1 - \mathrm{F}(l)) \frac{\hat{\Phi}_{\mathrm{s1}}(l) k_{aa}(l)}{\mu} \tag{44}$$

$$\bar{\beta}^{\mathrm{sdw}}(l) = \mathrm{F}(l) \frac{\hat{\Phi}_{\mathrm{s1}}(l) k_{bb}(l)}{\mu} \tag{45}$$

For this tuning, μ must be chosen (which is a familiar trade-off choice) and $\hat{\Phi}_{\mathrm{s1}}(l)$ must be computed, for instance as in [28]. Furthermore, it should be noted that if $\hat{\mathbf{h}} = \widetilde{\mathbf{h}}$, substitution of (44) and (45) into (23) would yield the cost function for the SDW-MWF for all values of $\mathrm{F}(l)$. Otherwise, there is a trade-off between the MVDR-APR and the MVDR-EST, with more emphasis placed on the respective beamformer as prescribed by $\mathrm{F}(l)$. With values computed for $\bar{\alpha}^{\mathrm{sdw}}(l)$ and $\bar{\beta}^{\mathrm{sdw}}(l)$, these parameters can then be substituted in (32) to compute the MVDR-INT beamformer. This tuning rule is summarised in Algorithm 1.

Algorithm 1 SDW tuning rule for the MVDR-INT at each time frame for a particular frequency.

Set value for μ, ϵ, ρ, and σ_t
for $l = 1$ to L **do**
(1) Estimate $\hat{\Phi}_{\mathrm{s1}}(l)$ from the reference microphone.
(2) Compute $\widetilde{\mathbf{w}}(l)$ from (7) and $\hat{\mathbf{w}}(l)$ from (13)
(3) Compute F (l) from (43).
(4) Compute $\bar{\alpha}^{\mathrm{sdw}}(l)$ from (44) and $\bar{\beta}^{\mathrm{sdw}}(l)$ from (45).
(5) Compute $\bar{\mathbf{w}}^{\epsilon}_{\mathrm{int}}(l)$ from (32)
end for

2) Contingency tuning rule: A second tuning rule may also be considered that is in line with the contingency noise reduction strategy of [25]. For this strategy, only $\bar{\beta}$ needs to be tuned, which can be computed as $\bar{\beta}^{\mathrm{sdw}}(l)$ from (45). This can then be substituted along with $\widetilde{\mathbf{w}}$ from (7), $\hat{\mathbf{w}}(l)$ from (13), and $\mathbf{w}^{\epsilon}_{\mathrm{lcmv}}$ from (22) into (35). In such a strategy, smaller values of $\bar{\beta}^{\mathrm{sdw}}(l)$ will tend to the MVDR-INT beamformer, while larger values of $\bar{\beta}^{\mathrm{sdw}}(l)$ will tend to the LCMV beamformer. This contingency tuning rule is summarised in Algorithm 2.

Algorithm 2 Contingency tuning rule for the MVDR-INT at each time frame for a particular frequency.

Set value for μ, ϵ, ρ, and σ_t
for $l = 1$ to L **do**
(1) Estimate $\hat{\Phi}_{\mathrm{s1}}(l)$ from the reference microphone.
(2) Compute $\widetilde{\mathbf{w}}(l)$ from (7) and $\hat{\mathbf{w}}(l)$ from (13)
(3) Compute $\mathbf{w}^{\epsilon}_{\mathrm{lcmv}}(l)$ from (22)
(4) Compute F (l) from (43).
(5) Compute $\bar{\beta}^{\mathrm{sdw}}(l)$ from (45).
(6) Compute $\bar{\mathbf{w}}^{\epsilon}_{\mathrm{int}}(l)$ from (35) (the contingency version)
end for

VI. Evaluation and Discussion

In order to evaluate the MVDR-INT strategy, both simulated data as well as recorded data from a typical office room were used. In the simulated case, scenarios with accurately and inaccurately estimated RTF vectors were considered in order to understand how the tuning parameters, $\bar{\alpha}$ and $\bar{\beta}$ affect the MVDR-INT beamformer. For the recorded data, the tuning rules as described in Section V were applied to the MVDR-INT beamformer and the resulting performance was compared to that from the MVDR-APR, MVDR-EST, and LCMV beamformers.

For the processing of the algorithms, the weighted overlap-and-add (WOLA) method [29], with a discrete Fourier transform (DFT) size of 512, $50\,\%$ overlap, a square-root Hanning window, and a sampling frequency of $16\,\mathrm{kHz}$ was used. Depending on the scenario under consideration, either a perfect or imperfect means of voice activity detection was used to retrieve $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$.

Although the processing described in this paper was done in the frequency domain, the respective quantities were converted back into the time domain for evaluation. The metrics used to evaluate the various experiments were the change in unweighted output SNR from the input to the output and the short-time objective intelligibility (STOI) measure [30]. The change in unweighted SNR (Δ SNR) was computed as follows:

$$\Delta\mathrm{SNR} = \underbrace{10\log_{10}\left(\frac{\mathbb{E}\{|Z_{\mathrm{x},1}[n]|^2\}}{\mathbb{E}\{|Z_{\mathrm{n},1}[n]|^2\}}\right)}_{\text{SNR output}} - \underbrace{10\log_{10}\left(\frac{\mathbb{E}\{|X_1[n]|^2\}}{\mathbb{E}\{|N_1[n]|^2\}}\right)}_{\text{SNR input}} \tag{46}$$

where n is the discrete time index, $Z_{\mathrm{x},1}[n]$ and $Z_{\mathrm{n},1}[n]$ are the individually processed speech-only and processed noise-only components in the discrete time domain resulting from the particular algorithm, and $X_1[n]$ and $N_1[n]$ are the unprocessed speech-only and unprocessed noise-only components in the discrete time domain at the reference microphone.

The STOI metric used $X_1[n]$ as the reference signal in order to evaluate the intelligibility of the processed signal, $Z_1[n]$. As opposed to the absolute values, a change in STOI (Δ STOI) from its respective value resulting from using the unprocessed signal was computed in order to reflect the relative improvements of the various algorithms.

A. Simulated Data

The simulation environment, depicted in Figure 1, consisted of a reverberant room with reverberation time (RT) $0.25\,\mathrm{s}$ and dimensions $4.3\,\mathrm{m} \times 6.9\,\mathrm{m} \times 2.6\,\mathrm{m}$, a linear microphone array, and a single speech source within a spherically diffuse noise field. The array consisted of 4 omnidirectional microphones with an inter-element spacing of 4 cm. For the speech source signal, six sentences separated by silence from the English Hearing-In-Noise Test (HINT) database [31] were used. The diffuse noise field was generated using uncorrelated excerpts of multitalker babble noise from Audiotec [32] and the method outlined in [33]. Uncorrelated white noise was also added to each of the microphone signals such that the ratio of the speech signal power in the first microphone of the array to the uncorrelated white noise power was 30 dB. The room impulse responses were obtained using the randomised image method [34] and implemented from [35]. $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ were then estimated accordingly via recursive averaging with an averaging time of $3\,\mathrm{s}$.

Two general scenarios in which to observe the effect of the tuning parameters of the MVDR-INT were considered: (i) where the RTF vector was accurately estimated and (ii) where the RTF vector was not accurately estimated. For each scenario, the speech source was placed 1 m away from the centre of the microphone array and was swept through a series of angles from 0^o to 180^o (as indicated by the bold curved arrow in figure 1). For each speech source angle, the evaluation metrics previously discussed were then calculated for the MVDR-APR, MVDR-EST and MVDR-INT for several values of $\bar{\alpha}$ and $\bar{\beta}$ as a function of the angle of the speech source. The regularised version of the MVDR-INT from (32) was used for all evaluations with the regularisation parameter, $\epsilon = 10^{-2}k_{aa}$.

For the MVDR-APR, the a priori RTF vector , $\widetilde{\mathbf{h}}$ was defined with respect to the endfire direction. A white noise signal of $20\,\mathrm{s}$ was played at the position of the speech source in the endfire direction in order to compute a rank-1 correlation matrix per frequency. The first column normalised with respect to a reference microphone was then used as the definition for $\widetilde{\mathbf{h}}$. For these simulations, the first microphone (m_1 in Fig. 1) was used as the reference microphone. For the MVDR-EST, $\hat{\mathbf{h}}$ was computed from (12) in section III-B.

1) Accurately estimated RTF vector : In this scenario, the input SNR at the reference microphone was set to 4 dB, and a perfect VAD was used to estimate the correlation matrices, $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$. The use of a perfect VAD is idealistic of course, however, it was used to ensure that the RTF vector would have indeed been accurately estimated.

Fig. 2 displays the subsequent results from the simulations in this scenario. On the left column, Figs. 2 (a)-(b) show the performance metrics of the MVDR-APR, MVDR-EST and the MVDR-INT for the tuning parameters such that $\bar{\alpha} \geq \bar{\beta}$. On the right column, figures 2 (c)-(d) show the same performance metrics for the MVDR-APR and MVDR-EST, but with the MVDR-INT tuned such that $\bar{\beta} > \bar{\alpha}$. A global legend for all of the plots in this section is provided in Table II, indicating the exact values of $\bar{\alpha}$ and $\bar{\beta}$ used for the MVDR-INT.

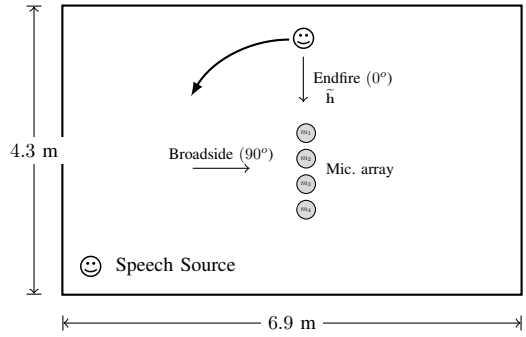


Fig. 1. Plan view of simulation environment for an accurately estimated RTF vector with a reverberation time of $0.25\,\mathrm{s}$. The arrow on the speech source indicates that simulations were done for different speech source angles, including that corresponding to $\widetilde{\mathbf{h}}$. Not shown is the simulated diffuse noise field. The room height was $2.6\,\mathrm{m}$.

TABLE II
GLOBAL LEGEND FOR THE PLOTS OF THE DIFFERENT NOISE REDUCTION ALGORITHMS IN FIG. 2 AND FIG. 3.

Plot Style	Noise Reduction Strategy	$\bar{\alpha}$	$\bar{\beta}$
	MVDR-EST	-	-
	MVDR-APR	-	-
	MVDR-INT	100	0.1
	MVDR-INT	100	1
	MVDR-INT	100	5
	MVDR-INT	100	10
	MVDR-INT	100	100
	MVDR-INT	0.1	100
	MVDR-INT	1	100
	MVDR-INT	5	100
	MVDR-INT	10	100

From all of the plots, it can be observed that the MVDR-EST is superior to the MVDR-APR as expected with higher Δ SI-SNR and Δ STOI for all speech source angles. The MVDR-APR on the other hand only has a performance similar to the MVDR-EST at a speech source angle of 0^o, i.e. the direction from which $\widetilde{\mathbf{h}}$ was defined and a clear drop in performance or all other speech source angles. It is noted as well that for the MVDR-EST, there is a general reduction in performance closer to the broadside direction. This is in accordance with the fact that superdirective beamformers achieve higher gains at endfire directions [36] and is not surprising, considering the symmetry of the beam patterns created by the array.

Focusing on the left column of Fig. 2, the various plots of the MVDR-INT are for $\bar{\alpha}$ fixed to 100 and $\bar{\beta} = \{0.1, 1, 5, 10, 100\}$, i.e. $\bar{\alpha} \geq \bar{\beta}$. Firstly, it can be seen that as $\bar{\alpha}$ is increasingly greater than $\bar{\beta}$, there is a convergence to the MVDR-APR as demonstrated by the extreme case where $\bar{\alpha} = 100$ and $\bar{\beta} = 0.1$. For non-zero values of $\bar{\beta}$, it is observed that there is an improved performance over the MVDR-APR beyond a source location of $30°$.

On the right column of Fig. 2, the various plots of the MVDR-INT are now for $\bar{\beta}$ fixed to 100 and $\bar{\alpha} = \{0, 0.1, 1, 5, 10\}$, i.e. $\bar{\beta} > \bar{\alpha}$. It can be seen that as $\bar{\beta}$ is increasingly greater than $\bar{\alpha}$, there is a convergence to the MVDR-EST as demonstrated by the extreme case where

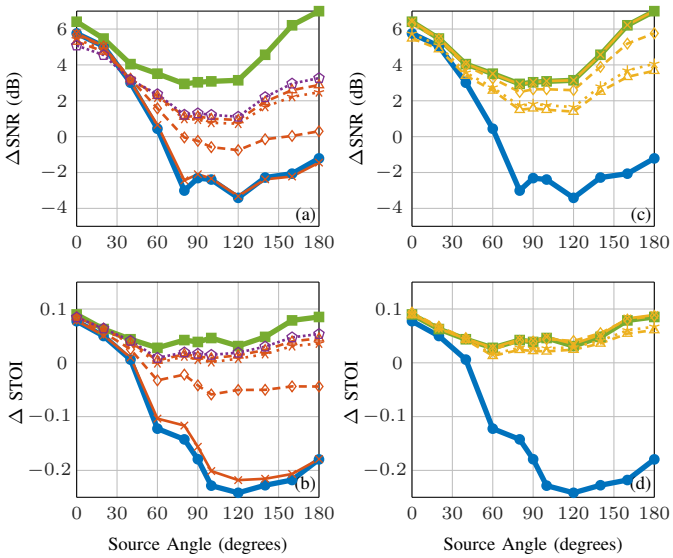


Fig. 2. Performance of the MVDR-APR, MVDR-EST and MVDR-INT with different values of $\bar{\alpha}$ and $\bar{\beta}$, for a scenario where $\hat{\mathbf{h}}$ was an accurately estimated RTF vector. The input SNR was 4 dB. The left column, (a)-(b), displays the MVDR-INT with $\bar{\alpha} \geq \bar{\beta}$ and the right column, (c)-(d), displays the MVDR-INT with $\bar{\beta} > \bar{\alpha}$. The corresponding legend is shown in Table II.

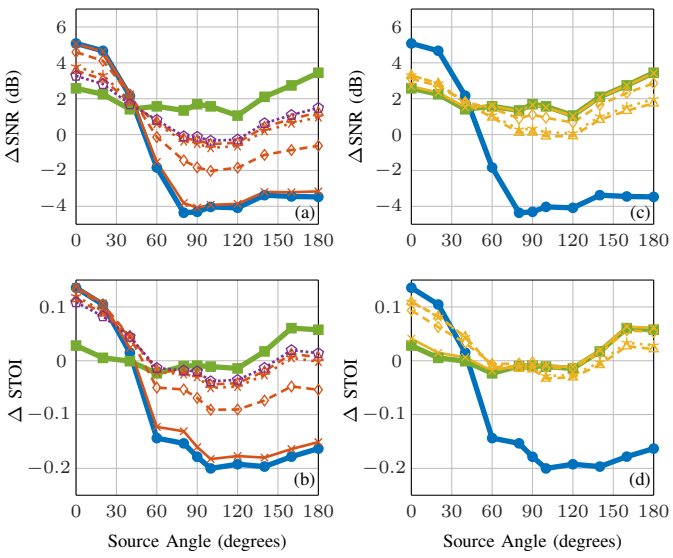


Fig. 3. Performance of the MVDR-APR, MVDR-EST and MVDR-INT with different values of $\bar{\alpha}$ and $\bar{\beta}$, for a scenario where $\hat{\mathbf{h}}$ was an inaccurately estimated RTF vector. The input SNR was -3 dB. The left column, (a)-(b), displays the MVDR-INT with $\bar{\alpha} \geq \bar{\beta}$ and the right column, (c)-(d), displays the MVDR-INT with $\bar{\beta} > \bar{\alpha}$. The corresponding legend is shown in Table II.

$\bar{\beta} = 100$ and $\bar{\alpha} = 0.1$. As $\bar{\alpha}$ is increased, however, the Δ STOI seems to be maintained, with some minimal reduction in the Δ SNR. This reduction would have been due to the fact that there is extra noise from the direction defined by $\tilde{\mathbf{h}}$ for greater values of $\bar{\alpha}$.

2) Inaccurately estimated RTF vector: Fig. 3 displays simulation results in the second scenario where the RTF vector estimate has not been accurate. Here the noise was scaled such that the input SNR at the reference microphone was -3 dB. An imperfect VAD [37], using the Minimum Mean Square Error (MMSE) method and with a mean talkspurt length of 500 ms was used to compute $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$. Once again, for each speech source angle, the Δ SI-SNR and the Δ STOI were computed for the MVDR-APR, MVDR-EST and the regularised version of the MVDR-INT from (32) for several tunings of $\bar{\alpha}$ and $\bar{\beta}$.

Immediately it can be observed that the MVDR-APR maintains its typical performance characteristic as in Fig. 2 for Δ SNR and Δ STOI metrics, where there is a maximum improvement at the endfire direction that tapers off in all other source angles. The MVDR-EST on the other hand suffers from a uniform reduction in performance across all source angles in comparison to Fig. 2 when the RTF vector was accurately estimated. The loss is particularly apparent in the endfire direction with respect to the MVDR-APR, which implies that the RTF vector was poorly estimated. Although there seems to be an improvement over the MVDR-APR outside of the endfire direction, a robust performance is not expected in this region due to the unpredictability of the imperfect VAD.

With respect to the MVDR-INT in both the left and right columns of Fig. 3, it is seen again that in the extreme case of $\bar{\alpha} = 100$ and $\bar{\beta} = 0.1$ it converges to the MVDR-APR and for $\bar{\alpha} = 0.1$ and $\bar{\beta} = 100$, it converges to the MVDR-EST. In all other tunings of $\bar{\alpha}$ and $\bar{\beta}$ displayed (i.e. the same tunings as described for Fig. 2), a similar performance is achieved that is a compromise between the MVDR-APR and the MVDR-EST. A benefit is indeed achieved in the endfire region over the MVDR-EST as $\bar{\alpha}$ is increasingly greater than $\bar{\beta}$, which introduces more of the MVDR-APR beamformer.

These simulations have provided some insight into the question of how to set the $\bar{\alpha}$ and $\bar{\beta}$ for the MVDR-INT. In the following section, the MVDR-INT will be applied to recordings from a typical office room, which use the proposed tuning strategies from Section V.

B. Recorded Data

Audio recordings were made in an office room of dimensions $5.4\,\mathrm{m} \times 3.5\,\mathrm{m} \times 2.5\,\mathrm{m}$ with an approximate reverberation time of 0.3 s. The audio signals (both speech and noise) were played through an RME UCX sound card, Samson SERVO 200 amplifier and a JBL CONTROL 1 PRO loudspeaker. These signals were then captured by a linear microphone array consisting of four omni-directional AKG CK32 microphones with an inter-element spacing of 4 cm. The microphones were connected to a Behringer EURORACK MX 3242X mixer and acquired through Simulink, Matlab via a Speedgoat real-time acquisition system. The speech and noise were recorded separately and added afterwards to create a desired input SNR.

As in the simulations, this experiment served to evaluate the performance of the various algorithms on the basis of a speech source in different locations. Hence, the speech source, SS1, was instantaneously moved between locations (as depicted in Fig. 4) - (i) the endfire direction (0^o, which would be set to the a priori direction), (ii) -45^o, (iii) the broadside direction (-90^o), and finally again to (iv) the endfire direction. For each of these locations 6 random sentences separated by silence from the HINT database were used, each lasting for a duration

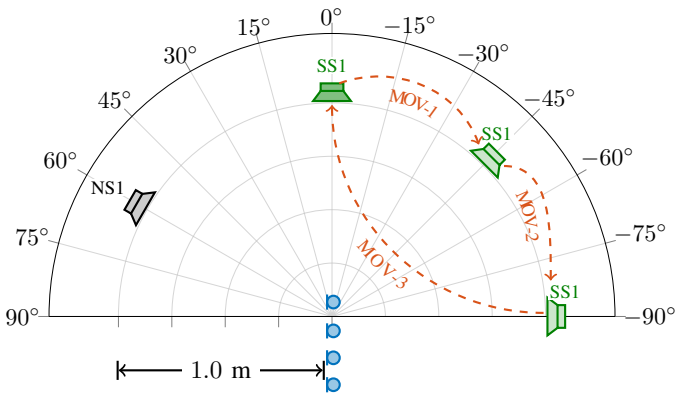


Fig. 4. Scenario for recordings captured by a 4-element microphone array in an office room. The speech source was played through SS1, which was moved from 0^o, -45^o, -90^o, and then back to 0^o. The speech source remained in each position for approximately 20 s. Speech-shaped noise was played through the NS1 at 60^o.

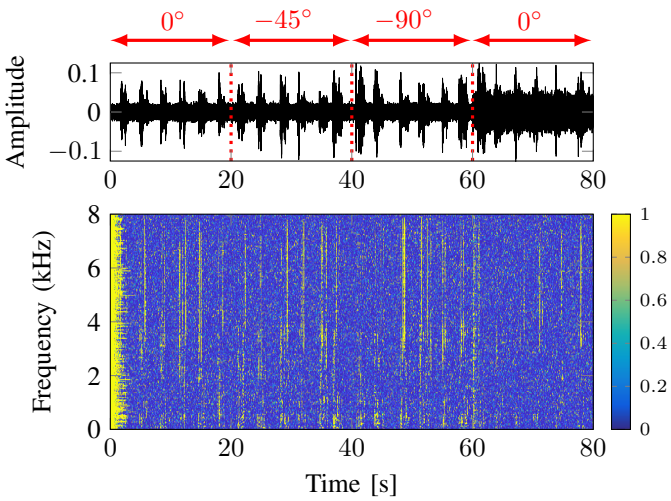


Fig. 5. (Top) Input noisy signal of the reference microphone. The arrows above the plot indicate the 20 s intervals for which the speech source was at a particular angle in accordance with Fig. 4. (Bottom) Corresponding probability if speech is present in this signal after applying an SPP estimator from [21].

of approximately 20 s. A localised noise source of speech-shaped noise was also recorded at an angle of 60^o relative to the endfire direction. In addition to recording the speech and the noise separately, a white noise signal played at the endfire position was also recorded and used to compute $\widetilde{\mathbf{h}}$ in a similar manner to that of Section VI-A. The uppermost microphone in Fig. 4 was used as the reference microphone.

Using the individual speech and noise recordings, a noisy signal was created such that for the first 60 s (i.e. during the instantaneous movements from the endfire to the broadside direction), the input SNR at the reference microphone was approximately 4 dB and for the last 20 s (when the source was back in the endfire direction), the noise was increased so that the input SNR at the reference microphone in this segment was approximately -3 dB.

The upper plot of Fig. 5 illustrates this noisy input signal of the reference microphone and the lower plot of Fig. 5 depicts the corresponding probability if speech is present in the STFT domain after applying the SPP estimator from [21]. Using this result, periods for which the speech was active were extracted if the SPP ≥ 0.5, and periods of noise were extracted if the SPP < 0.5. $\hat{\mathbf{R}}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}$ were then computed accordingly via recursive averaging with an averaging time of 1 s. On inspection of the lower plot of Fig. 5, it can be observed that in the final 20 s, where the input SNR is lower, particularly in the lower frequencies, the speech presence is not as distinct as compared for previous times, which is indicative of some misclassification.

Fig. 6 displays the resulting performance of the MVDR-APR, MVDR-EST, and the regularised LCMV from (22), along with the MVDR-INT, for the different tuning rules described in Section V-B when $\mu = 0.001$. $\hat{\Phi}_{\mathrm{s1}}(l)$ was computed using the method from [28] as implemented in [37] but with the noise estimation update computed as in [21]. The SDW tuning rule from Algorithm 1 will be referred to as MVDR-INT-sdw and the contingency tuning rule from Algorithm 2 will be referred to as MVDR-INT-cnt. All algorithms were evaluated using the metrics mentioned at the beginning of this section and were computed over 5 s time frames with a 50 % overlap. For the logistic function of (43), the parameters were set such that $\rho = 10$, and $\sigma_t = 20$. For the regularisation parameter, $\epsilon = 10^{-2} k_{aa}$.

Firstly, focusing on the MVDR-APR and the MVDR-EST, as expected, the MVDR-APR exhibits a robust performance when the speech is in the endfire direction (a priori assigned direction), i.e. during the time segments of the first 20 s and the final 20 s. For all other time segments, where the speech was outside of the endfire direction, the metrics are indicative of a reduced performance. The MVDR-EST on the other hand maintains its performance for changes in the speech source location during the first 60 s, but then suffers a reduction in performance in the final 20 s, where $\hat{\mathbf{h}}$ would have been an inaccurate estimate due to the lower input SNR. For the MVDR-EST, it should be noted that although the Δ STOI seems similar from 40 s to 80 s, the absolute STOI value of at the reference microphone is much lower in the final 20 s, and hence the MVDR-EST indeed exhibits a poor performance in these final 20 s. The LCMV is able to attain a performance that is in between that of the MVDR-APR and the MVDR-EST. However, as indicated by the Δ SNR, when the speech is outside of the endfire direction, less noise is reduced in comparison to the MVDR-EST , and when $\hat{\mathbf{h}}$ was inaccurate, less noise is reduced in comparison to the MVDR-APR.

For the first 60 s, it can be observed that the performance of the MVDR-INT-sdw is similar to that of the MVDR-EST in terms of both Δ SNR and Δ STOI. In the final 20 s, when $\hat{\mathbf{h}}$ was inaccurate, the MVDR-INT-sdw however moves away from the performance of the MVDR-EST and approaches that of the MVDR-APR, truly merging the benefits of both the MVDR-APR and MVDR-EST. The MVDR-INT-cnt, on the other hand, maintains the performance of the MVDR-APR when the speech source is in the endfire direction, but otherwise has a performance that is similar to the LCMV. Hence the MVDR-INT-cnt also reduces less noise when the speech is

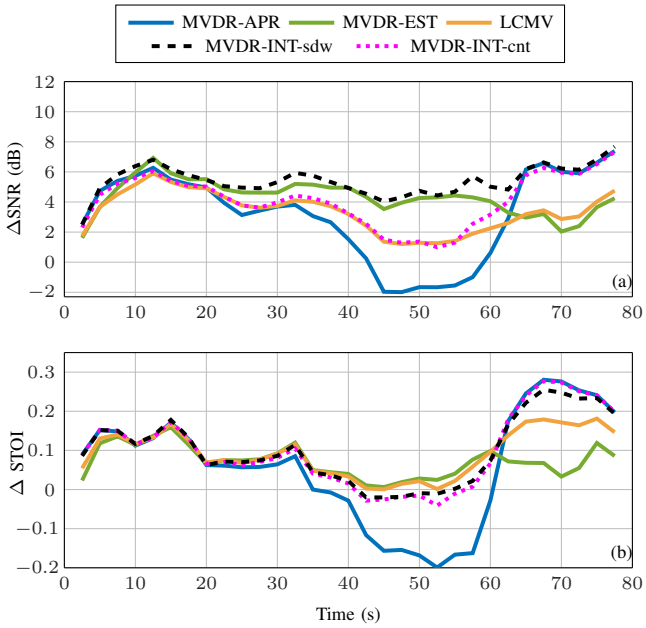


Fig. 6. Performance of the MVDR-APR, MVDR-EST, LCMV, MVDR-INT-sdw, and MVDR-INT-cnt when $\mu = 0.001$ for the scenario of Fig. 4.

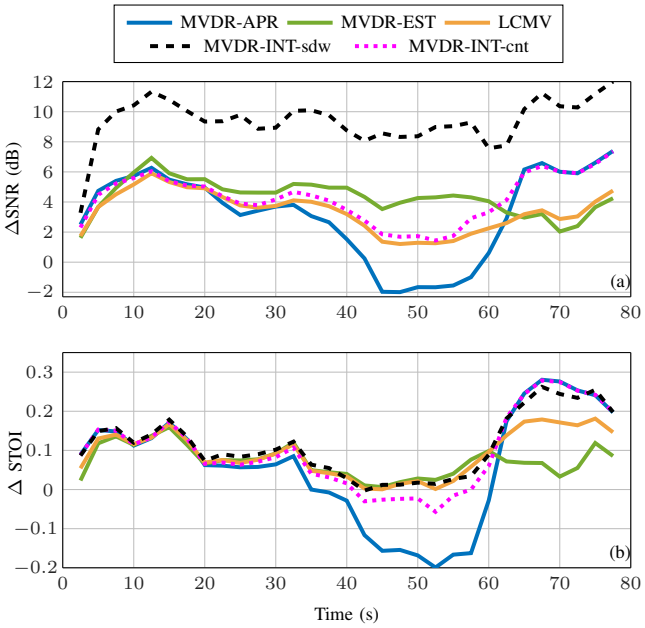


Fig. 7. Performance of the MVDR-APR, MVDR-EST, LCMV, MVDR-INT-sdw, and MVDR-INT-cnt when $\mu = 0.1$ for the scenario of Fig. 4.

not in the a priori assumed location, but nevertheless merges the benefits of the MVDR-APR and the LCMV.

Fig. 7 displays the results for the same experiment as in Fig. 6, except that $\mu = 0.1$ for the MVDR-INT-sdw and MVDR-INT-cnt. This larger value of μ introduces a more aggressive post filter, which accounts for the increased Δ SNR for the MVDR-INT-sdw. The Δ STOI remains similar to that when $\mu = 0.001$, although some audible artifacts can now be heard, which is an expected result for more aggressive postfilters. The performance of the MVDR-INT-cnt remains relatively similar to that when $\mu = 0.001$.

The results of Fig. 6 and Fig. 7 can be further explained by observing the values of the metric of confidence, $\mathrm{F}(k,l)$ in Fig. 8. In general, there was more confidence for the higher frequencies and periods where the speech source was outside of the a priori assumed direction, resulting in larger values being assigned to $\bar{\beta}$ and hence imposing more of the MVDR-EST. In the final 20 s in particular, there was less confidence in the lower frequency region, resulting in lower values being assigned to $\bar{\beta}$, which would have imposed more of the MVDR-APR. This behaviour is indeed consistent with the observations from the SPP of Fig. 5, where it was the lower frequencies in the final 20 s that suffered from some misclassification.

Both tuning rules have demonstrated that the MVDR-INT can indeed be a useful strategy as it has exhibited a performance that is more accomodating for changes in the acoustic environment as compared to using either the MVDR-APR, MVDR-EST, or LCMV only. The resulting audio signals from this experiment and a repeated version that uses a babble noise may be listened to for a subjective evaluation at [38].

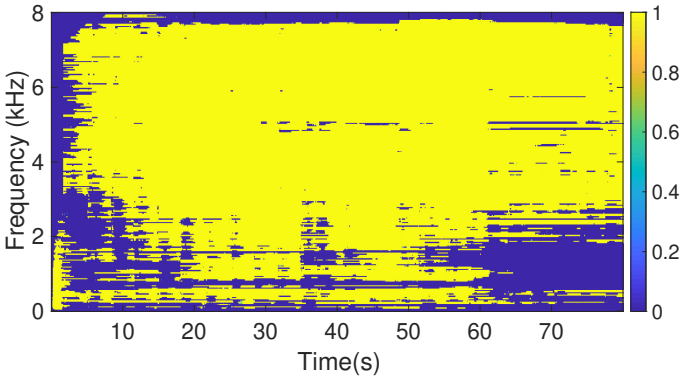


Fig. 8. Metric of confidence, $\mathrm{F}(k,l)$ from the results of Fig. 6 and Fig. 7.

VII. Conclusion

In an MVDR beamformer, the relative transfer function (RTF) vector constraint is usually defined from either (i) a priori assumptions regarding microphone characteristics, position, speech source location and room acoustics (MVDR-APR) or (ii) an estimate using the speech-plus-noise correlation matrix and noise-only correlation matrix (MVDR-EST). Each of these approaches has their respective merits and drawbacks in certain acoustic scenarios and a decision usually has to be made as to which of the approaches to follow for a particular application. In this paper, an analysis and evaluation has been carried out on an alternative approach of using an integrated MVDR (MVDR-INT) beamformer, where both the hard constraints from the two conventional approaches are softened to yield two tuning parameters, $\bar{\alpha}$ and $\bar{\beta}$.

It was found that the MVDR-INT could be expressed as a convex combination of the MVDR-APR beamformer, the

MVDR-EST beamformers, a linearly constrained minimum variance (LCMV) beamformer, and an all-zero vector, with real, positive-valued coefficients. The effect of different tuning combinations of $\bar{\alpha}$ and $\bar{\beta}$ on the behaviour of the MVDR-INT beamformer was then explored theoretically and confirmed through simulations in cases where an RTF vector estimate was accurate and where it was inaccurate, demonstrating that the MVDR-INT encompasses a wide range of speech enhancement filters.

Using a metric of confidence and the relation of the MVDR-INT to the speech-distortion-weighted Multi-channel Wiener Filter, two tuning rules were proposed for a practical implementation of the MVDR-INT. An evaluation with recorded data from an office room demonstrated that the MVDR-INT beamformer can indeed provide a more versatile beamformer by merging the benefits from the MVDR-APR and MVDR-EST beamformers.

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 692–730, April 2017.

[2] M. H. Er and A. Cantoni, "Derivative Constraints for Broad-band Element Space Antenna Array Processors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 6, pp. 1378–1393, 1983.

[3] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, April 1988.

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

[5] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug 1972.

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

[7] M. Kompis and N. Dillier, "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.*, vol. 109, no. 3, pp. 1123–1133, 2001.

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

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

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

[11] I. Cohen, "Relative Transfer Function Identification Using Speech Signals," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, 2004.

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

[13] C. Pan, J. Chen, and J. Benesty, "Performance Study of the MVDR Beamformer as a Function of the Source Incidence Angle," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, no. 1, pp. 67–79, 2014.

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

[15] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Amer.*, vol. 54, no. 3, pp. 771–785, 1973.

[16] S. A. Vorobyov, A. B. Gershman, and Z. Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, 2003.

[17] A. Spriet, S. Doclo, M. Moonen, and J. Wouters, "A unification of adaptive multi-microphone noise reduction systems," in *Proc. 2006 Int. Workshop Acoustic Echo Noise Control (IWAENC '06)*, Paris, France, Sept. 2006.

[18] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, 1987.

[19] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. Ray Liu, Eds. John Wiley & Sons, Inc., 2010, ch. 10, pp. 269–302.

[20] S. Van Gerven and F. Xie, "A comparative study of speech detection methods," *Proc. Fifth European Conference on Speech*, pp. 1095–1098, 1997.

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

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

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[24] S. Chakrabarty and E. A. P. Habets, "On the Numerical Instability of an LCMV Beamformer for a Uniform Linear Array," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 272–276, 2016.

[25] R. Ali and M. Moonen, "A contingency multi-microphone noise reduction strategy based on linearly constrained multi-channel wiener filtering," in *Proc. 2016 Int. Workshop Acoustic Signal Enhancement (IWAENC '16)*, Xi'an, China, Sept 2016.

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

[27] ——, "The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and Generalized Sidelobe Cancellation," *Signal Processing*, vol. 85, no. 6, pp. 1073–1088, 2005.

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

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

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

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

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

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

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

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

[36] G. Elko, "Superdirectional microphone arrays," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds. Kluwer Academic Press, 2010, ch. 10, pp. 181–235.

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

[38] R.Ali. (2019). [Online]. Available: ftp://ftp.esat.kuleuven.be/pub/SISTA/rali/Reports/INT_MVDR_LMA/Audio_Data