KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Authors: Arnout Roebben, Jitske Van Peer, Jesper Jensen, Toon van Waterschoot, Jan Wouters, and Marc Moonen Title: Hybrid Algorithm Design for Noise Reduction with the Generalised Eigenvalue Decomposition Multi-Channel Wiener Filter
Archived version	■ Submitted version ☐ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher) ☐ Published ☐ Published (Open Access)
Acknowledgements	In this manuscript, the following funding is acknowledged: ☐ EU H22020 ☐ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ☐ Other



Hybrid Algorithm Design for Noise Reduction with the Generalised Eigenvalue Decomposition Multi-Channel Wiener Filter

Arnout Roebben, *Graduate Student Member, IEEE*, Jitske Van Peer, Jesper Jensen, Toon van Waterschoot, *Member, IEEE*, Jan Wouters, and Marc Moonen, *Fellow, IEEE*

***Abstract*—In many speech recording applications the recorded desired speech is corrupted by noise, such that a noise reduction algorithm is required. This noise reduction algorithm can be designed following a model-based, data-driven, or hybrid design approach. In a model-based design, a filter is obtained by optimising a cost function under the assumption of a specific signal model. In a data-driven design, on the contrary, no signal model is assumed, and a neural network is optimised, according to a selected cost function, to directly estimate filter coefficients or the desired speech signal. In a hybrid design, model-based and data-driven designs are fused aiming to combine the benefits of both, i.e., achieving the advantage of both the interpretable signal model and the neural network performance. In this paper, hybrid algorithm design is applied to multi-microphone noise reduction with the generalised eigenvalue decomposition multichannel Wiener filter (GEVD-MWF), where the neural network estimates the generalised eigenvectors and eigenvalues, the correlation matrices, or the smoothing variables of the GEVD-MWF model-structure. The neural network is optimised end-to-end on the noise reduction task. It is shown that the hybrid GEVD-MWF algorithms can outperform model-based and even data-driven counterparts through leveraging both the signal model and the neural network-based parameter estimation, i.e., that the hybrid algorithms seem to benefit from using the model structure. The hybrid algorithms also produce (interpretable) parameters that show a degree of similarity with the spectrogram of the clean speech, despite these parameters being optimised end-to-end.**

***Index Terms*—Speech enhancement, Noise reduction (NR), Multi-microphone, Generalised eigenvalue decomposition multi-channel Wiener filter (GEVD-MWF), Hybrid design**

I. Introduction

In many speech recording applications, such as in hearing instruments and in teleconferencing devices, the recorded desired speech is corrupted by noise, such that a noise reduction algorithm is required [1], [2]. This noise reduction algorithm can either be single-microphone, thereby only exploiting spectro-temporal information, or multi-microphone, thereby exploiting both spectro-temporal and spatial information [1], [3]. Single- and multi-microphone algorithms are generally designed following one of three design approaches, namely a model-based design approach, a data-driven design approach, or a hybrid design approach [2].

In a model-based design, a signal model is formulated, which is substituted into a cost function, together with the filter that is optimised [2], [3]. This then leads to an optimal filter with respect to both the formulated signal model and cost function. This signal model and cost function are oftentimes chosen such that a closed-form solution arises [1], [3]–[7]. For example, if the desired speech is modelled as an acoustic transfer function (ATF) applied to a source signal, then minimising the output noise power, while preserving the sound originating from the ATF, leads to the minimum variance distortionless response (MVDR) beamformer [3], [4]. Minimising the mean squared error (MSE) between the filtered microphone signals and the desired speech, under the assumption of additive noise that is uncorrelated to the desired speech, leads to the multi-channel Wiener filter (MWF) [1], [3], [7]. The MWF based on a generalised eigenvalue decomposition (GEVD), hence referred to as GEVD-MWF, is a variant of the MWF that makes use of GEVD operations to guarantee positive semi-definiteness of the correlation matrices involved and mostly results in improved performance [7].

In a data-driven design, no specific signal model is assumed [2]. Microphone signals are passed to a neural network, which either directly returns an estimate of the desired speech signal [8]–[12], or a filter (so-called mask) to be applied to the microphone signals [13]–[19]. Variants that combine both approaches also exist [20], [21]. The parameters of the neural network are obtained by minimising a cost function on a training set. This then leads to a neural network optimised for the selected cost function and training set [2], [22].

Data-driven algorithms typically achieve higher performance than model-based algorithms as they are not restricted by model assumptions and allow for a larger diversity in cost functions [2], [22]. However, model-based algorithms do not require training sets, tend to be more robust to deviations from these training sets, and have higher interpretability as the parameters are tied to the signal model [2], [22].

In a hybrid design, model-based and data-driven designs are fused aiming to combine the benefits of both, i.e., achieving the advantage of both the interpretable signal model and the

This research was carried out at the ESAT Laboratory of KU Leuven, Belgium, and at Eriksholm Research Centre, Denmark, in the frame of Research Council KU Leuven C14-21-0075 "A holistic approach to the design of integrated and distributed digital signal processing algorithms for audio and speech communication devices", Aspirant Grant 11PDH24N (for A. Roebben), and Grant For A Long Stay Abroad V438025N (for A. Roebben) from the Research Foundation - Flanders (FWO). The scientific responsibility is assumed by its authors. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the FWO and the Flemish Government.

Arnout Roebben, Jitske Van Peer, Toon van Waterschoot and Marc Moonen are with the Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, B-3001 Leuven, Belgium (e-mail: arnout.roebben@esat.kuleuven.be; jitske.vanpeer@student.kuleuven.be; toon.vanwaterschoot@esat.kuleuven.be; marc.moonen@esat.kuleuven.be).

Jesper Jensen is with the Eriksholm Research Centre (Part of Demant), 3070 Snekkersten, Denmark, and also with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: jesj@demant.com).

Jan Wouters is with the Department of Neurosciences, Research Group ExpORL, KU Leuven, B-3000 Leuven, Belgium (e-mail: jan.wouters@med.kuleuven.be).

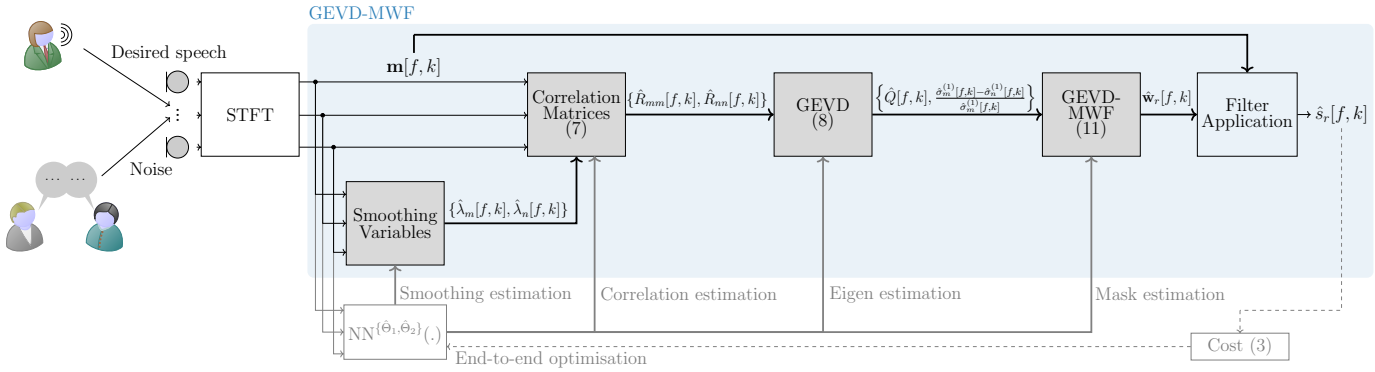

Fig. 1. Block diagram of the proposed hybrid GEVD-MWF algorithms. When the neural network directly estimates the filter, data-driven mask estimation is obtained. Without neural network-based parameter estimation, the model-based GEVD-MWF is obtained. The gray-shaded blocks accept parameters estimated by the neural network. The figures have been generated using [37].

neural network performance [2]. Generally, hybrid designs can be divided into two classes. A first class consists of methods that estimate the parameters of a model-based algorithm using a neural network, i.e., the parameters provided by the neural network are inserted into the model-based algorithm [2], [23]–[30]. This neural network is either a priori optimised for parameter estimation or optimised end-to-end. In the former case, this neural network is optimised, in isolation, for a parameter estimation task before run-time [23]–[25]. In the latter case, during training, the neural network outputs parameters that are inserted into the model-structure, and this hybrid model is subsequently applied for the noise reduction task. The cost associated with this noise reduction task is then fed back through the model to the neural network to optimise performance on the training set [26]–[30], resulting in a hybrid algorithm of which the model structure is optimal with respect to the signal model and cost function used to derive the underlying model-based algorithm, and of which the parameters are optimised for a (potentially different) cost function and training set. These end-to-end trained networks have shown to outperform their a priori parameter estimation counterparts [26]. A second class consists of methods in which model-based algorithms and data-driven algorithms are cascaded [31], [32]. A variant thereof consists in replacing the model-based algorithm with a hybrid algorithm [33]–[36]. An advantage of the first class is that the resulting hybrid algorithms truly interpolate between model-based algorithms and data-driven algorithms, as removing the data-driven parameter estimation results in the model-based algorithms, while directly estimating the filter coefficients results in the data-driven algorithms. Both classes of hybrid designs have been applied to single- [25], [26], [36] and multi-microphone [23], [24], [27]–[35] scenarios.

In this paper, a hybrid algorithm design is applied to multi-microphone noise reduction with the GEVD-MWF, wherein the end-to-end trained neural network estimates the generalised eigenvectors and eigenvalues, the correlation matrices, or the smoothing variables as used by the GEVD-MWF model-structure. The output activation functions and dimensions of the neural network are chosen such that the estimated parameters satisfy the properties of valid eigenvectors, eigenvalues, correlation matrices, or smoothing variables. A block diagram of the proposed algorithm is shown in Fig. 1. It will be shown that the hybrid GEVD-MWF algorithms can outperform model-based and even data-driven counterparts through leveraging both the signal model, originating from the model-based design, and neural network-based parameter estimation, originating from the data-driven design, while also producing (interpretable) parameters that show a degree of similarity with the spectrogram of the clean speech, despite these parameters being optimised end-to-end. As such, the hybrid algorithms seem to benefit from using the model structure.

The paper is structured as follows. The noise reduction problem is described in Section II. In Section III and Section IV, model-based design using the GEVD-MWF and data-driven design using mask estimation are reviewed, respectively. The proposed hybrid algorithm design using the GEVD-MWF is discussed in Section V. Using the experimental setup as described in Section VI, the proposed hybrid algorithms with the estimation methods for the generalised eigenvectors and eigenvalues, correlation matrices, or smoothing variables are analysed and compared to the model-based and data-driven algorithms in Section VII. Python code is available at [38]. Finally, Section VIII draws the conclusions.

II. Problem Statement

Considering an M-microphone setup, the noisy microphone signals $m_i[f,k]$, $i \in \{1,\cdots,M\}$, formulated in the short-time Fourier transform (STFT) domain with frequency bin index $f \in \{0,\cdots,F-1\}$ and frame index $k \in \{0,\cdots,K-1\}$, can be stacked in a microphone signal vector $\mathbf{m}[f,k] \in \mathbb{C}^M$ as

$$\mathbf{m}[f,k] = \begin{bmatrix} m_1[f,k] & \cdots & m_M[f,k] \end{bmatrix}^{\mathsf{T}}. \tag{1}$$

This microphone signal vector $\mathbf{m}[f,k]$ can be decomposed into a desired speech signal vector $\mathbf{s}[f,k] \in \mathbb{C}^M$ and a noise signal vector $\mathbf{n}[f,k] \in \mathbb{C}^M$ as

$$\mathbf{m}[f,k] = \mathbf{s}[f,k] + \mathbf{n}[f,k]. \tag{2}$$

Multi-microphone NR aims at reducing the noise while preserving the desired speech through filtering the microphone signals [1]. Throughout this paper, this filter $\mathbf{w}_r[f,k] \in \mathbb{C}^M$ is designed by minimising the MSE between the desired speech

$s_r[f,k]$ in a chosen reference microphone $r \in \{1,\cdots,M\}$, and the filtered microphone signal vector $\hat{s}_r[f,k]$, i.e., [1], [7]

$$\mathbf{w}_r[f,k] = \underset{\mathbf{w}[f,k]}{\operatorname{argmin}} \mathbb{E}\{|s_r[f,k] - \underbrace{\mathbf{w}[f,k]^H \mathbf{m}[f,k]}_{\triangleq \hat{s}_r[f,k]}|^2\}, \quad (3)$$

where $\mathbb{E}\{.\}$ denotes the expected value operator. This procedure can be repeated for multiple reference microphones r depending on potential post-processing [1], [32].

III. Model-based Design

In a model-based design, the cost function (3) is optimised using the signal model (2) with the assumption of uncorrelated desired speech and noise, i.e., $\mathbb{E}\{\mathbf{s}[f,k]\mathbf{n}[f,k]^H\} = 0_{M\times M}$ [1], [7]. Under these assumptions, the multi-channel Wiener filter (MWF) is obtained as [1], [7]

$$\mathbf{w}_r[f,k] = R_{mm}[f,k]^{-1} R_{ss}[f,k]\mathbf{t}_r, \quad (4)$$

with $\mathbf{t}_r \in \mathbb{C}^M$ representing a unit vector with 0-elements everywhere except for a 1 at position r to select reference microphone r, with $R_{mm}[f,k] \in \mathbb{C}^{M\times M}$ the microphone correlation matrix,

$$\begin{aligned} R_{mm}[f,k] &= \mathbb{E}\{\mathbf{m}[f,k]\mathbf{m}[f,k]^H\} && (5a)\\ &= \mathbb{E}\{\mathbf{s}[f,k]\mathbf{s}[f,k]^H\} + \mathbb{E}\{\mathbf{n}[f,k]\mathbf{n}[f,k]^H\} && (5b)\\ &\triangleq R_{ss}[f,k] + R_{nn}[f,k], && (5c)\end{aligned}$$

and with $R_{ss}[f,k]$ and $R_{nn}[f,k]$ the desired speech and noise correlation matrix, respectively. In practice, the desired speech is assumed to exhibit an on-off behaviour due to speech pauses, while the noise is assumed to be always-on [1], [7]. This on-off behaviour can be tracked using a voice activity detector (VAD), i.e., VAD$[f,k]$, resulting in the following regimes [7]

$$\begin{cases} \text{VAD}[f,k] = 1 : \mathbf{s}[f,k] + \mathbf{n}[f,k] \text{ recorded} \\ \text{VAD}[f,k] = 0 : \mathbf{n}[f,k] \text{ recorded.} \end{cases} \quad (6)$$

Consequently, the estimates of $R_{mm}[f,k]$ and $R_{nn}[f,k]$, can be computed using recursive averaging as [1], [7]

$$\begin{aligned} \hat{R}_{mm}[f,k] &= \hat{\lambda}_m[f,k]\hat{R}_{mm}[f,k-1] + \\ &\quad (1-\hat{\lambda}_m[f,k])(\mathbf{m}[f,k]\mathbf{m}[f,k]^H), && (7a)\\ \hat{R}_{nn}[f,k] &= \hat{\lambda}_n[f,k]\hat{R}_{nn}[f,k-1] + \\ &\quad (1-\hat{\lambda}_n[f,k])(\mathbf{m}[f,k]\mathbf{m}[f,k]^H), && (7b)\end{aligned}$$

wherein the microphone (resp. noise) smoothing variable $\hat{\lambda}_m[f,k]$ (resp. $\hat{\lambda}_n[f,k]$) $\in [0,1]$ is set close to 1 when VAD$[f,k] = 0$ (resp. VAD$[f,k] = 1$) to prevent updating of $\hat{R}_{mm}[f,k]$ (resp. $\hat{R}_{nn}[f,k]$) during periods of speech inactivity (resp. activity).

While from (5c), it is seen that the estimated desired speech correlation matrix $\hat{R}_{ss}[f,k]$ can be computed as $\hat{R}_{mm}[f,k] - \hat{R}_{nn}[f,k]$, this subtraction generally does not lead to a good estimate of $R_{ss}[f,k]$ [7]. Indeed, plain subtraction does not guarantee positive semi-definiteness, and generally leads to a full-rank estimate [7]. However, theoretically, $R_{ss}[f,k]$ is positive semi-definite as it is a Hermitian, square matrix, and, theoretically, it is assumed to be low-rank. For example, when there is only one desired speaker, the desired speech signal vector can be represented as $\mathbf{s}[f,k] = \mathbf{h}[f,k]s_0[f,k]$ with $s_0[f,k]$ the source signal and with $\mathbf{h}[f,k] \in \mathbb{C}^M$ containing the ATFs from the source to the microphones [1], [7]. In this scenario, $R_{ss}[f,k] = \mathbb{E}\{s_0[f,k]s_0[f,k]^*\}\mathbf{h}[f,k]\mathbf{h}[f,k]^H$ is indeed both positive semi-definite, as $\mathbb{E}\{s_0[f,k]s_0[f,k]^*\} = E\{|s_0[f,k]|^2\} \geq 0$, and rank-1.

Thus, $\hat{R}_{ss}[f,k]$ is ideally constrained to be positive semi-definite and low-rank. Both conditions can be incorporated through a generalised eigenvalue decomposition (GEVD) of the matrix pair $\{\hat{R}_{mm}[f,k], \hat{R}_{nn}[f,k]\}$ [7]. Indeed, with $\hat{Q}[f,k] \in \mathbb{C}^{M\times M}$ representing the generalised eigenvectors and with $\hat{\Sigma}_{mm}[f,k] \in \mathbb{C}^{M\times M}$ and $\hat{\Sigma}_{nn}[f,k] \in \mathbb{C}^{M\times M}$ representing the diagonal matrices containing $\{\hat{\sigma}_m^{(1)},\cdots,\hat{\sigma}_m^{(M)}\}$ and $\{\hat{\sigma}_n^{(1)},\cdots,\hat{\sigma}_n^{(M)}\}$, which define the generalised eigenvalues $\frac{\hat{\sigma}_m^{(i)}}{\hat{\sigma}_n^{(i)}}$, respectively, the GEVD jointly diagonalises $\hat{R}_{mm}[f,k]$ and $\hat{R}_{nn}[f,k]$, i.e.,

$$\begin{aligned} \hat{R}_{mm}[f,k] &= \hat{Q}[f,k]\hat{\Sigma}_{mm}[f,k]\hat{Q}[f,k]^H && (8a)\\ &= \hat{Q}[f,k]\text{diag}(\hat{\sigma}_m^{(1)}[f,k],\cdots,\hat{\sigma}_m^{(M)}[f,k])\hat{Q}[f,k]^H, && (8b)\\ \hat{R}_{nn}[f,k] &= \hat{Q}[f,k]\hat{\Sigma}_{nn}[f,k]\hat{Q}[f,k]^H && (8c)\\ &= \hat{Q}[f,k]\text{diag}(\hat{\sigma}_n^{(1)}[f,k],\cdots,\hat{\sigma}_n^{(M)}[f,k])\hat{Q}[f,k]^H. && (8d)\end{aligned}$$

Within this equation, diag$(.)$ represents the operator to create a diagonal matrix, and it is assumed, without loss of generality, that the generalised eigenvalues are sorted as

$$\frac{\hat{\sigma}_m^{(1)}}{\hat{\sigma}_n^{(1)}} \geq \frac{\hat{\sigma}_m^{(2)}}{\hat{\sigma}_n^{(2)}} \geq \cdots \geq \frac{\hat{\sigma}_m^{(M)}}{\hat{\sigma}_n^{(M)}}. \quad (9)$$

The estimated desired speech correlation matrix can then be enforced to be of rank-R, wherein R can be freely chosen, by computing $\hat{\sigma}_s^{(i)}[f,k] = \hat{\sigma}_m^{(i)}[f,k] - \hat{\sigma}_n^{(i)}[f,k]$, and by only retaining the first R eigenmodes [7]

$$\begin{aligned} \hat{R}_{ss}[f,k] = \hat{Q}[f,k]\text{diag}(&\hat{\sigma}_m^{(1)}[f,k] - \hat{\sigma}_n^{(1)}[f,k],\cdots,\\ &\hat{\sigma}_m^{(R)}[f,k] - \hat{\sigma}_n^{(R)}[f,k],0,\cdots,0)\hat{Q}[f,k]^H. \end{aligned} \quad (10)$$

Throughout this paper, one desired speaker is assumed, thereby setting $R = 1$ according to the above reasoning and according to [7]. Plugging the estimated microphone correlation matrix (8b) and desired speech correlation matrix (10) into the MWF formulation of (4) then yields the final closed-form solution of the GEVD-MWF as given in [7]

$$\hat{\mathbf{w}}_r[f,k] = \hat{Q}[f,k]^{-H}\text{diag}\Big(\frac{\hat{\sigma}_m^{(1)}[f,k] - \hat{\sigma}_n^{(1)}[f,k]}{\hat{\sigma}_m^{(1)}[f,k]},0,\cdots,0\Big)\hat{Q}[f,k]^H\mathbf{t}_r. \quad (11)$$

IV. Data-driven Design

In a data-driven design, no specific signal model is assumed [2]. The filter, termed mask in literature [2], is obtained from a neural network that takes microphone signals as input. More specifically, current and past microphone signals (frame k –

$L_K + 1$ to frame k) may be leveraged by a real-valued neural network to provide the filter $\hat{\mathbf{w}}_r[f,k]$, i.e.,[1]

$$\boxed{\begin{aligned}&\hat{\mathbf{w}}_r[f,k] = \Re(\hat{\mathbf{w}}_r[f,k]) + j\Im(\hat{\mathbf{w}}_r[f,k])\\ &\begin{bmatrix}\Re(\hat{\mathbf{w}}_r[k])\\ \Im(\hat{\mathbf{w}}_r[k])\end{bmatrix} = \mathrm{NN}^{\{\hat{\mathbf{w}}_r\}}\left(\begin{bmatrix}\Re(\underline{\mathbf{m}}[k])\\ \Im(\underline{\mathbf{m}}[k])\end{bmatrix}\right),\end{aligned}} \tag{12}$$

with

$$\begin{aligned}\underline{\mathbf{m}}[k] = [\mathbf{m}[0,k-L_K+1]^\mathsf{T} \cdots \mathbf{m}[F-1,k-L_K+1]^\mathsf{T} \\ \cdots \mathbf{m}[0,k]^\mathsf{T} \cdots \mathbf{m}[F-1,k]^\mathsf{T}]^\mathsf{T}, \end{aligned} \tag{13a}$$

$$\hat{\mathbf{w}}_r[k] = \left[\hat{\mathbf{w}}_r[0,k]^\mathsf{T} \quad \cdots \quad \hat{\mathbf{w}}_r[F-1,k]^\mathsf{T}\right]^\mathsf{T}, \tag{13b}$$

with $\mathrm{NN}^{\{\hat{\mathbf{w}}_r\}}(.)$ the neural network for mask estimation, and with $\Re(.)$ and $\Im(.)$ denoting the real and imaginary part of a complex number. The multi-microphone mask is defined in correspondence with [17]–[19]. In the alternative definition of [15], [16], a mask is computed using all microphone signals but applied only to the signal of reference microphone r. This alternative definition is encompassed in the definition of (12) when the filter coefficients of $\hat{\mathbf{w}}_r[f,k]$ corresponding to microphones other than r are set to zero. The parameters of the neural network are optimised on training data, according to the cost function (3), or another cost function. It is also noted that while the resulting mask is applied in a per-frequency bin manner, the mask itself is a function of (possibly) all input frequency bins as per [15]–[19]. This is opposed to the model-based design of Section III, where no information is exchanged between frequency bins when estimating $\hat{\mathbf{w}}_r[f,k]$.

V. Hybrid Design

In the proposed hybrid algorithm design, the GEVD-MWF structure is leveraged and complemented with methods to to estimate the eigenvectors and eigenvalues (Section V-A), the microphone and noise correlation matrices (Section V-B), or the microphone and noise smoothing variables (Section V-C). These various parameters may be denoted in general by the matrix pair $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$. The different choices of $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$ will be detailed in the following sections, but one example of the eigen estimation corresponds to $\hat{\Theta}_1[f,k] = \hat{Q}[f,k]$ and $\hat{\Theta}_2[f,k] = \hat{\sigma}_m^{(1)}[f,k]$, and the smoothing estimation corresponds to $\hat{\Theta}_1[f,k] = \hat{\lambda}_m[f,k]$ and $\hat{\Theta}_2[f,k] = \hat{\lambda}_n[f,k]$. Similarly to (12), the parameters in $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$ are obtained by providing current and past microphone signals (frame $k - L_K + 1$ to frame k) to a neural network $\mathrm{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)$, and then these parameters are used to compute the filter $\hat{\mathbf{w}}_r[f,k]$ as

$$\boxed{\begin{aligned}&\hat{\mathbf{w}}_r[f,k](\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]) \text{ of (11) with}\\ &\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} \leftarrow \begin{bmatrix}\hat{\boldsymbol{\theta}}_1[f,k]\\ \hat{\boldsymbol{\theta}}_2[f,k]\end{bmatrix}\\ &\begin{bmatrix}\hat{\boldsymbol{\theta}}_1[k]\\ \hat{\boldsymbol{\theta}}_2[k]\end{bmatrix} = g\left(\mathrm{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}\left(\begin{bmatrix}\Re(\underline{\mathbf{m}}[k])\\ \Im(\underline{\mathbf{m}}[k])\end{bmatrix}\right)\right) = g(\hat{\boldsymbol{\mu}}[k]),\end{aligned}} \tag{14}$$

with $g(.)$ the element-wise activation function, up to a permutation, of the final layer, with

$$\hat{\boldsymbol{\theta}}_1[k] = \left[\hat{\boldsymbol{\theta}}_1[0,k]^\mathsf{T} \quad \cdots \quad \hat{\boldsymbol{\theta}}_1[F-1,k]^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{C_{\theta_1}} \tag{15a}$$

$$\hat{\boldsymbol{\theta}}_2[k] = \left[\hat{\boldsymbol{\theta}}_2[0,k]^\mathsf{T} \quad \cdots \quad \hat{\boldsymbol{\theta}}_2[F-1,k]^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{C_{\theta_2}} \tag{15b}$$

$$\hat{\boldsymbol{\mu}}[k] = \left[\hat{\boldsymbol{\mu}}[0,k]^\mathsf{T} \quad \cdots \quad \hat{\boldsymbol{\mu}}[F-1,k]^\mathsf{T}\right]^\mathsf{T} \in \mathbb{R}^{C} \tag{15c}$$

and with $C = C_{\theta_1} + C_{\theta_2}$. The neural network outputs $\left[\hat{\boldsymbol{\theta}}_1[f,k]^\mathsf{T} \quad \hat{\boldsymbol{\theta}}_2[f,k]^\mathsf{T}\right]^\mathsf{T}$, which maps to $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$ through algebraic operations as will be detailed in the following sections. For example, for the eigen estimation with $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{Q}[f,k], \hat{\sigma}_m^{(1)}[f,k]\}$, $\hat{\boldsymbol{\theta}}_1[f,k]$ is a real-valued vector that contains the real and imaginary part of $\hat{Q}[f,k]$ with $C_{\theta_1} = F \cdot 2M^2$, and as $\hat{\sigma}_m^{(1)}[f,k]$ is a real-valued scalar, $C_{\theta_2} = F$ and $\hat{\Theta}_2[f,k] = \hat{\boldsymbol{\theta}}_2[f,k] = \hat{\sigma}_m^{(1)}[f,k]$. The output dimensions and activations of the neural network are chosen such that $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$ are valid parameters, e.g., such that the estimated smoothing variables are between zero and one. As with the data-driven design, the neural network may integrate information across multiple frequency bins to estimate $\left[\hat{\boldsymbol{\theta}}_1[f,k]^\mathsf{T} \quad \hat{\boldsymbol{\theta}}_2[f,k]^\mathsf{T}\right]^\mathsf{T}$.

While the cost function for this parameter optimisation may be chosen differently from the cost function for the model-based design, in this paper, the MSE cost function (3) is selected for both purposes. The filter constructed following the model-based design (11), the data-driven design (12), and the hybrid design (14) is then considered to be optimised for the same cost function, thereby allowing for a fair performance comparison.

These resulting hybrid algorithms also truly interpolate between model-based and data-driven algorithms. Indeed, when there are no neural networks involved in the estimation of $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$, the hybrid design (14) reverts back to a model-based design (11). When directly estimating the filter coefficients, i.e., $\hat{\Theta}_1[f,k] = \Re(\hat{\mathbf{w}}_r[f,k])$ and $\hat{\Theta}_2[f,k] = \Im(\hat{\mathbf{w}}_r[f,k])$, the hybrid design (14) transforms into the data-driven design (12). In Fig. 1, a block diagram of the hybrid GEVD-MWF illustrates how (11), (12), and (14) are estimated in practice, and how these solutions link to one another. Additionally, Fig. 1 illustrates the estimation of the generalised eigenvectors and eigenvalues, estimation of the correlation matrices, and estimation of the smoothing coefficients.

A. Eigen Estimation

In the proposed eigen estimation, the goal is to estimate a matrix pair $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$, related to the generalised eigenvectors and eigenvalues as used in (11). As apparent from the definition of the GEVD in (8), scalars can be interchanged between the columns of the generalised eigenvectors and the corresponding diagonal elements. Four estimation methods are presented here that differ in the way this scaling ambiguity is resolved, and depending on the specific parameters estimated, as will be detailed in the following.

1) $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{V}[f,k], \hat{\sigma}_m^{(1)}[f,k]\}$:
Typically, GEVD algorithms do not directly provide $\hat{Q}[f,k]$, but rather $\hat{V}[f,k]$ [39], [40], which relates to $\hat{Q}[f,k]$ as

$$\hat{Q}[f,k] = \hat{V}[f,k]^{-H} \tag{16}$$

[1]For convolutional neural networks (CNNs) L_K has a fixed value determined by the input and network parameters. For recurrent neural networks (RNNs), L_K does theoretically correspond to $k+1$, yielding an infinite context. In Section VII, the role of L_K is further detailed.

In these GEVD algorithms [39], [40], the scaling ambiguity is typically resolved by setting

$$\hat{\sigma}_n^{(1)}[f,k] = \cdots = \hat{\sigma}_n^{(M)}[f,k] = 1. \tag{17}$$

The output arguments of these GEVD algorithms can be mimicked by defining $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{V}[f,k], \hat{\sigma}_m^{(1)}[f,k]\}$. Consequently, the output dimensions and activations of the neural network are defined to produce valid parameters $\hat{V}[f,k]$ and $\hat{\sigma}_m^{(1)}[f,k]$. As generalised eigenvectors do not exhibit orthogonality properties, $2M^2$ parameters need to be estimated for $\hat{V}[f,k]$, i.e., M^2 parameters for $\Re(\hat{V}[f,k])$ and M^2 parameters for $\Im(\hat{V}[f,k])$. In addition, one parameter needs to be estimated for the positive and real-valued $\hat{\sigma}_m^{(1)}[f,k]$. Nevertheless mimicking the fact that in the model-based algorithm, M diagonal elements are estimated, after which the largest generalised eigenvalue (9) is selected, in the proposed eigen estimation M values $\hat{\sigma}_m^{(i)}[f,k]$ are computed, to which max-pooling is applied. To this end, with $\hat{\boldsymbol{\mu}}[f,k] \in \mathbb{R}^{(2M^2+M)}$, as defined in (15c), $g\left(\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)\right)$, as defined in (14), outputs the following parameters

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\theta}}_1[f,k]\\ \hat{\boldsymbol{\theta}}_2[f,k]\end{bmatrix} &= \begin{bmatrix}\begin{bmatrix}\Re(\hat{\mathbf{v}}[f,k])\\ \Im(\hat{\mathbf{v}}[f,k])\end{bmatrix}\\ \hat{\sigma}_m^{(1)}[f,k]\end{bmatrix}\\ &= \begin{bmatrix}\text{ELU}\left(\begin{bmatrix}\mathbb{I}_{2M^2} & 0_{2M^2\times M}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\\ \max\left(\text{SP}\left(\begin{bmatrix}0_{M\times 2M^2} & \mathbb{I}_M\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)+1\right)\end{bmatrix},\end{aligned} \tag{18}$$

which map to $\hat{V}[f,k]$ and $\hat{\sigma}_m^{(1)}[f,k]$ as

$$\begin{aligned}\hat{\Theta}_1[f,k] &= \hat{V}[f,k] \qquad (19a)\\ &= \text{v2m}(\Re(\hat{\mathbf{v}}[f,k])) + j\cdot\text{v2m}(\Im(\hat{\mathbf{v}}[f,k])),\\ &= \text{v2m}\left(\begin{bmatrix}\mathbb{I}_{M^2} & 0_{M^2\times M^2}\end{bmatrix}\hat{\boldsymbol{\theta}}_1[f,k]\right)\\ &\quad + j\cdot\text{v2m}\left(\begin{bmatrix}0_{M^2\times M^2} & \mathbb{I}_{M^2}\end{bmatrix}\hat{\boldsymbol{\theta}}_1[f,k]\right),\\ \hat{\Theta}_2[f,k] &= \hat{\boldsymbol{\theta}}_2[f,k] = \hat{\sigma}_m^{(1)}[f,k]. \qquad (19b)\end{aligned}$$

In this equation, $\mathbb{I}_M$ represents an $M\times M$ identity matrix, and the output activations are defined to produce valid parameters. To this end, the exponential linear unit activation function (ELU(.)) allows for both positive and negative values to define the coefficients of the generalised eigenvectors, while the softplus activation function (SP(.)) only allows non-negative values to define the diagonal elements of $\hat{\Sigma}_{mm}[f,k]$. Additionally, as $\hat{\sigma}_m^{(1)}[f,k] = \hat{\sigma}_s^{(1)}[f,k] + \hat{\sigma}_n^{(1)}[f,k]$ and as $\hat{\sigma}_n^{(1)}[f,k] = 1$, given (17), $\hat{\sigma}_m^{(1)}[f,k]$ should be larger than one. This property is enforced by addition with one before max-pooling in (18). Finally, the mapping from an M^2-dimensional vector to an $M\times M$ matrix is represented using the vector-to-matrix (v2m(.)) operation.

2) $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{Q}[f,k], \hat{\sigma}_m^{(1)}[f,k]\}$*:* Instead of $\hat{V}[f,k]$, $\hat{Q}[f,k]$ could be directly estimated. Indeed, as it is not clear in advance whether estimating $\hat{V}[f,k]$ or $\hat{Q}[f,k]$ is beneficial performance-wise, both estimation procedures are explored. This matrix $\hat{Q}[f,k]$, like $\hat{V}[f,k]$, is an $M\times M$ matrix with complex elements. Consequently, with the assumption (17), and with $\hat{\boldsymbol{\mu}}[f,k] \in \mathbb{R}^{(2M^2+M)}$, $g\left(\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)\right)$ outputs the following parameters

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\theta}}_1[f,k]\\ \hat{\boldsymbol{\theta}}_2[f,k]\end{bmatrix} &= \begin{bmatrix}\begin{bmatrix}\Re(\hat{\mathbf{q}}[f,k])\\ \Im(\hat{\mathbf{q}}[f,k])\end{bmatrix}\\ \hat{\sigma}_m^{(1)}[f,k]\end{bmatrix}\\ &= \begin{bmatrix}\text{ELU}\left(\begin{bmatrix}\mathbb{I}_{2M^2} & 0_{2M^2\times M}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\\ \max\left(\text{SP}\left(\begin{bmatrix}0_{M\times 2M^2} & \mathbb{I}_M\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)+1\right)\end{bmatrix},\end{aligned} \tag{20}$$

which map to $\hat{Q}[f,k]$, rather than to $\hat{V}[f,k]$, and $\hat{\sigma}_m^{(1)}[f,k]$ as

$$\begin{aligned}\hat{\Theta}_1[f,k] &= \hat{Q}[f,k] \qquad (21a)\\ &= \text{v2m}(\Re(\hat{\mathbf{q}}[f,k])) + j\cdot\text{v2m}(\Im(\hat{\mathbf{q}}[f,k])),\\ \hat{\Theta}_2[f,k] &= \hat{\boldsymbol{\theta}}_2[f,k] = \hat{\sigma}_m^{(1)}[f,k]. \qquad (21b)\end{aligned}$$

3) $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{Q}[f,k], \frac{\hat{\sigma}_m^{(1)}[f,k]-\hat{\sigma}_n^{(1)}[f,k]}{\hat{\sigma}_m^{(1)}[f,k]}\}$*:* In the expression of the GEVD-MWF in (11), the elements $\hat{\sigma}_m^{(1)}[f,k]$ and $\hat{\sigma}_n^{(1)}[f,k]$ are only used to define the ratio $\frac{\hat{\sigma}_m^{(1)}[f,k]-\hat{\sigma}_n^{(1)}[f,k]}{\hat{\sigma}_m^{(1)}[f,k]}$, hence this ratio could also be estimated directly. This effectively limits the range of valid $\hat{\Theta}_2[f,k]$ parameters from non-negative to ranging between zero and one. To this end, $\hat{\Theta}_1 = \hat{Q}[f,k]$ is again estimated as per (20) and (21a), while, with $\hat{\boldsymbol{\mu}}[f,k] \in \mathbb{R}^{(2M^2+M)}$, $g\left(\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)\right)$ outputs the following $\hat{\boldsymbol{\theta}}_2[f,k] = \hat{\Theta}_2[f,k]$ parameter

$$\begin{aligned}\hat{\Theta}_2[f,k] &= \hat{\boldsymbol{\theta}}_2[f,k]\\ &= \max\left(\text{SIGM}\left(\begin{bmatrix}0_{M\times 2M^2} & \mathbb{I}_M\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\right).\end{aligned} \tag{22}$$

Indeed, the sigmoid (SIGM(.)) activation function enforces the estimated parameter to range from zero to one. As with (18) and (20), max-pooling is applied to enforce the rank-1 constraint.

4) $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{\overline{Q}}[f,k], \frac{\hat{\sigma}_m^{(1)}[f,k]-\hat{\sigma}_n^{(1)}[f,k]}{\hat{\sigma}_m^{(1)}[f,k]}\}$*:* The scaling ambiguity can also be resolved by normalising the columns in $\hat{Q}[f,k]$ such that each element in the first row corresponds to one, i.e., with $\mathbf{1}_M \in \mathbb{R}^M$ a vector of all ones,

$$\hat{Q}[f,k] = \begin{bmatrix}\mathbf{1}_M & \hat{\overline{Q}}[f,k]^{\mathsf{T}}\end{bmatrix}^{\mathsf{T}}, \tag{23}$$

This provides two benefits. First, a reduced number of parameters needs to be estimated for the generalised eigenvectors, i.e., only $2(M-1)M$, rather than $2M^2$, parameters are required to reconstruct $\hat{\overline{Q}}[f,k] \in \mathbb{C}^{(M-1)\times M}$. Second, referring to (8), multiplying $\hat{R}_{mm}[f,k]$ and $\hat{R}_{nn}[f,k]$ with a real-valued scalar then no longer affects $\hat{Q}[f,k]$, e.g., due to an increase in signal power of the noise source. To this end, $\hat{\boldsymbol{\theta}}_2[f,k] = \hat{\Theta}_2[f,k] = \frac{\hat{\sigma}_m^{(1)}[f,k]-\hat{\sigma}_n^{(1)}[f,k]}{\hat{\sigma}_m^{(1)}[f,k]}$ is estimated according to (22), while with $\hat{\boldsymbol{\mu}}[f,k] \in \mathbb{R}^{(2M^2-M)}$, $g\left(\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)\right)$ outputs the following $\hat{\boldsymbol{\theta}}_1[f,k]$ parameter

$$\begin{aligned}\hat{\boldsymbol{\theta}}_1[f,k] &= \begin{bmatrix}\Re(\hat{\overline{\mathbf{q}}}[f,k])\\ \Im(\hat{\overline{\mathbf{q}}}[f,k])\end{bmatrix}\\ &= \text{ELU}\left(\begin{bmatrix}\mathbb{I}_{2M^2-2M} & 0_{(2M^2-2M)\times M}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right),\end{aligned} \tag{24}$$

which maps to $\hat{\overline{Q}}[f,k]$ as

$$\hat{\Theta}_1[f,k] = \hat{\overline{Q}}[f,k] \tag{25a}$$

$$= \text{v2m}(\Re(\hat{\overline{\mathbf{q}}}[f,k])) + j \cdot \text{v2m}(\Im(\hat{\overline{\mathbf{q}}}[f,k])). \quad (25b)$$

B. Correlation Estimation

In the proposed correlation estimation, the goal is to estimate a matrix pair $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$, related to the microphone and noise correlation matrices as used in (8). Three estimation methods are presented here, i.e., the estimation of the Cholesky factors, similar to [26], [30], [36], and two novel approaches by estimating Hermitian matrices, which are constrained to be positive semi-definite. Indeed, it is not clear a priori which decomposition is beneficial performance-wise.

1) $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{\Gamma}_{mm}[f,k], \hat{\Gamma}_{nn}[f,k]\}$*:*
In [26], [30], [36], it has been shown that a symmetric, positive semi-definite matrix can be estimated through estimating a Cholesky decomposition using a neural network. Specifically, this method has been applied for single-microphone correlation matrices in a multi-frame MVDR [26], [36], for single-microphone correlation matrices in a multi-frame regular MWF [36], and for multi-microphone correlation matrices in a multi-frame MWF that is decomposed as an MVDR followed by a post-filter [30]. Here, a similar approach for multi-microphone correlation matrices in a GEVD-MWF model structure is proposed. Indeed, $\hat{R}_{mm}[f,k]$ and $\hat{R}_{nn}[f,k]$ can be decomposed as [26], [30], [36]

$$\hat{R}_{mm}[f,k] = \hat{\Gamma}_{mm}[f,k]\hat{\Gamma}_{mm}[f,k]^H, \quad (26a)$$
$$\hat{R}_{nn}[f,k] = \hat{\Gamma}_{nn}[f,k]\hat{\Gamma}_{nn}[f,k]^H, \quad (26b)$$

with $\hat{\Gamma}_{mm}[f,k] \in \mathbb{C}^{M\times M}$ and $\hat{\Gamma}_{nn}[f,k] \in \mathbb{C}^{M\times M}$ upper-triangular matrices (Cholesky factors). Each Cholesky factor consists of M real-valued diagonal elements, and $S = \frac{M(M-1)}{2}$ complex-valued off-diagonal elements. To this end, with $\hat{\boldsymbol{\mu}}[f,k] \in \mathbb{R}^{2M^2}$, $g\left(\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)\right)$ outputs the following parameters

$$\begin{bmatrix}\hat{\boldsymbol{\theta}}_1[f,k]\\ \hat{\boldsymbol{\theta}}_2[f,k]\end{bmatrix} = \begin{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\delta}}_m[f,k]\\ \Re(\hat{\boldsymbol{\gamma}}_m[f,k])\\ \Im(\hat{\boldsymbol{\gamma}}_m[f,k])\end{bmatrix}\\ \begin{bmatrix}\hat{\boldsymbol{\delta}}_n[f,k]\\ \Re(\hat{\boldsymbol{\gamma}}_n[f,k])\\ \Im(\hat{\boldsymbol{\gamma}}_n[f,k])\end{bmatrix}\end{bmatrix} \quad (27)$$
$$= \begin{bmatrix}\begin{bmatrix}\text{SP}\left(\begin{bmatrix}\mathbb{I}_M & 0_{M\times 2S}\end{bmatrix}\begin{bmatrix}\mathbb{I}_{M^2} & 0_{M^2\times M^2}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\\ \text{PRELU}\left(\begin{bmatrix}0_{S\times M} & \mathbb{I}_S & 0_{S\times S}\end{bmatrix}\begin{bmatrix}\mathbb{I}_{M^2} & 0_{M^2\times M^2}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\\ \text{PRELU}\left(\begin{bmatrix}0_{S\times (M+S)} & \mathbb{I}_S\end{bmatrix}\begin{bmatrix}\mathbb{I}_{M^2} & 0_{M^2\times M^2}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\end{bmatrix}\\ \begin{bmatrix}\text{SP}\left(\begin{bmatrix}\mathbb{I}_M & 0_{M\times 2S}\end{bmatrix}\begin{bmatrix}0_{M^2\times M^2} & \mathbb{I}_{M^2}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\\ \text{PRELU}\left(\begin{bmatrix}0_{S\times M} & \mathbb{I}_S & 0_{S\times S}\end{bmatrix}\begin{bmatrix}0_{M^2\times M^2} & \mathbb{I}_{M^2}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\\ \text{PRELU}\left(\begin{bmatrix}0_{S\times (M+S)} & \mathbb{I}_S\end{bmatrix}\begin{bmatrix}0_{M^2\times M^2} & \mathbb{I}_{M^2}\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\end{bmatrix}\end{bmatrix},$$

which map to $\hat{\Gamma}_{mm}[f,k]$ and $\hat{\Gamma}_{nn}[f,k]$ as

$$\hat{\Theta}_1[f,k] = \hat{\Gamma}_{mm}[f,k] \quad (28a)$$
$$= \text{striu}(\Re(\hat{\boldsymbol{\gamma}}_m[f,k])) + j \cdot \text{striu}(\Im(\hat{\boldsymbol{\gamma}}_m[f,k])) + \text{diag}(\hat{\boldsymbol{\delta}}_m[f,k]), \quad (28b)$$
$$\hat{\Theta}_2[f,k] = \hat{\Gamma}_{nn}[f,k] \quad (28c)$$
$$= \text{striu}(\Re(\hat{\boldsymbol{\gamma}}_n[f,k])) + j \cdot \text{striu}(\Im(\hat{\boldsymbol{\gamma}}_n[f,k])) + \text{diag}(\hat{\boldsymbol{\delta}}_n[f,k]). \quad (28d)$$

The softplus activation function (SP(.)) bounds the diagonal elements to be non-negative, while the parametric rectified linear unit activation function (PRELU(.)), as per [26], allows for both positive and negative values. The mapping from an $\frac{M(M-1)}{2}$-dimensional vector to a strictly upper-triangular $M \times M$ matrix is represented using the striu(.) operation.

2) $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{R}^{(H)}_{mm}[f,k], \hat{R}^{(H)}_{nn}[f,k]\}$ *With Regularisation:*
Rather than using $\hat{\Gamma}_{mm}[f,k]$ and $\hat{\Gamma}_{nn}[f,k]$ as Cholesky factors in (28), these upper-triangular matrices, constructed from (27), can be used to create Hermitian, not necessarily positive semi-definite, matrices $\hat{R}^{(H)}_{mm}[f,k] \in \mathbb{C}^{M\times M}$ and $\hat{R}^{(H)}_{nn}[f,k] \in \mathbb{C}^{M\times M}$, respectively, as

$$\hat{\Theta}_1[f,k] = \hat{R}^{(H)}_{mm}[f,k] = \hat{\Gamma}_{mm}[f,k] + \hat{\Gamma}_{mm}[f,k]^H, \quad (29a)$$
$$\hat{\Theta}_2[f,k] = \hat{R}^{(H)}_{nn}[f,k] = \hat{\Gamma}_{nn}[f,k] + \hat{\Gamma}_{nn}[f,k]^H. \quad (29b)$$

The Hermitian matrices $\hat{R}^{(H)}_{mm}[f,k]$ and $\hat{R}^{(H)}_{nn}[f,k]$ cannot be directly utilised in (8) as they are not necessarily positive semi-definite, and hence not necessarily valid correlation matrices. This property can, however, be imposed by regularising both matrices

$$\hat{R}_{mm}[f,k] = \hat{R}^{(H)}_{mm}[f,k] + |\hat{\alpha}_m[f,k]|\,\mathbb{I}_M, \quad (30a)$$
$$\hat{R}_{nn}[f,k] = \hat{R}^{(H)}_{nn}[f,k] + |\hat{\alpha}_n[f,k]|\,\mathbb{I}_M, \quad (30b)$$

where $\hat{\alpha}_m[f,k]$ (resp. $\hat{\alpha}_n[f,k]$) corresponds to the negative eigenvalue of $\hat{R}^{(H)}_{mm}[f,k]$ (resp. $\hat{R}^{(H)}_{nn}[f,k]$) that is largest in magnitude. This regularisation does indeed ensure positive semi-definiteness, as any Hermitian matrix, e.g., $\hat{R}^{(H)}_{mm}[f,k]$, can be decomposed using an eigenvalue decomposition with eigenvectors $\hat{T}[f,k] \in \mathbb{C}^{M\times M}$ and eigenvalues $\hat{\alpha}^{(i)}_m[f,k] \in \mathbb{R}$ as [40]

$$\hat{R}^{(H)}_{mm}[f,k] = \hat{T}[f,k]\text{diag}(\hat{\alpha}^{(1)}_m[f,k], \cdots, \alpha^{(M)}_m[f,k])\hat{T}[f,k]^H, \quad (31)$$

with $\hat{T}[f,k]\hat{T}[f,k]^H = \mathbb{I}_{M\times M}$, such that (30) can be written as

$$\hat{R}_{mm}[f,k] = \hat{T}[f,k]\big(\text{diag}(\hat{\alpha}^{(1)}_m[f,k], \cdots, \hat{\alpha}^{(M)}_m[f,k]) + \max_{i=1,\cdots,M}\{|\hat{\alpha}^{(i)}_m[f,k]| : \hat{\alpha}^{(i)}_m[f,k] < 0\}\mathbb{I}_{M\times M}\big)T[f,k]^H, \quad (32)$$

thereby enforcing positive semi-definiteness. This correlation estimation with regularisation thus yields an alternative to the correlation estimation using the Cholesky factors by using the $\hat{\Theta}_1[f,k]$ and $\hat{\Theta}_2[f,k]$ in a different manner.

3) $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{R}^{(H)}_{mm}[f,k], \hat{R}^{(H)}_{nn}[f,k]\}$ *With Manifold Projection:*
Instead of regularising $\hat{R}^{(H)}_{mm}[f,k]$ (resp. $\hat{R}^{(H)}_{nn}[f,k]$), the matrix can also be rendered positive semi-definite by projecting it on the positive semi-definite manifold as [41]

$$\hat{R}_{mm}[f,k] = \Pi_{\text{SPD}}(\hat{R}^{(H)}_{mm}[f,k]), \quad (33a)$$
$$\hat{R}_{nn}[f,k] = \Pi_{\text{SPD}}(\hat{R}^{(H)}_{nn}[f,k]), \quad (33b)$$

where $\Pi_{\text{SPD}}(.)$ denotes the projection. This projection corresponds to computing an eigenvalue decomposition of $\hat{R}^{(H)}_{mm}[f,k]$ (resp. $\hat{R}^{(H)}_{nn}[f,k]$) and replacing the negative eigenvalues with zero (or a small positive number). While with the regularisation (30) a factor $|\hat{\alpha}_m[f,k]|$ (resp. $|\hat{\alpha}_n[f,k]|$) is

added to each eigenvalue of $\hat{R}_{mm}^{(H)}[f,k]$ (resp. $\hat{R}_{nn}^{(H)}[f,k]$), with the projection onto the positive semi-definite manifold (33) the positive eigenvalues remain unchanged while the negative eigenvalues are replaced with zero (or a small positive number).

C. Smoothing Estimation

In the proposed correlation estimation, the goal is to estimate a matrix pair $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$, related to the microphone and noise smoothing variables as used in (7). The approach has been shown to be effective in hybrid designs [26], [33], but has not been applied to the GEVD-MWF before.

$\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\} = \{\hat{\lambda}_m[f,k], \hat{\lambda}_n[f,k]\}$:
The smoothing variables $\hat{\lambda}_m[f,k]$ and $\hat{\lambda}_n[f,k]$ can be output by $g\left(\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)\right)$, with $\hat{\boldsymbol{\mu}}[f,k] \in \mathbb{R}^2$, as [26]

$$\begin{bmatrix}\hat{\Theta}_1[f,k]\\ \hat{\Theta}_2[f,k]\end{bmatrix} = \begin{bmatrix}\hat{\theta}_1[f,k]\\ \hat{\theta}_2[f,k]\end{bmatrix} = \begin{bmatrix}\hat{\lambda}_m[f,k]\\ \hat{\lambda}_n[f,k]\end{bmatrix} = \begin{bmatrix}\text{SIGM}\left(\begin{bmatrix}1 & 0\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\\ \text{SIGM}\left(\begin{bmatrix}0 & 1\end{bmatrix}\hat{\boldsymbol{\mu}}[f,k]\right)\end{bmatrix}. \quad (34)$$

The sigmoid (SIGM(.)) activation function ensures the estimated smoothing variables are ranging between zero and one. As these smoothing variables are dependent on the frequency bin index f, they may vary between different bins for the same frame index k.

VI. Experiment Design

A. Acoustic Scenarios

Two datasets are employed to simulate the source-to-microphone propagation in a hearing instrument setting, namely the Hearpiece dataset [42], which was recorded with in-the-ear (ITE) hearing instruments with four microphones at each side in an anechoic room, and the Multi-Array Room Acoustic Database (MYRiAD) [43], which was recorded with behind-the-ear (BTE) hearing instruments in a room with $T_{60} = 0.5\,\text{s}$.

The training set generated from the Hearpiece dataset is created from the impulse responses from twenty-five loudspeakers to both the ipsilateral in-ear - outervent and ipsilateral entrance - concha microphone pairs at each side of twenty subjects wearing the ITE hearing instrument, to which the algorithms are applied monaurally. One desired speech source and one noise source, each placed randomly in one of the twenty-five available source positions, are considered. The validation set generated from the Hearpiece dataset is created from the same available source locations, but for three unseen subjects, and also considers one desired speech and one noise source. Two test sets are generated from this Hearpiece dataset. One test set again containing the same available source positions, one desired speech source, and one noise source, but for three unseen subjects. Another test set is created from the impulse responses for the same three unseen subjects but using fifteen unseen available source positions, as well as two, rather than one, noise sources, such that there are more noise sources than degrees of freedom as defined by the $M = 2$ ipsilateral microphones. The first test set enables the study of the generalisation across unseen source signals, while its combination with the second test set enables the study of the effect of, and generalisation across, unseen positions of the sources and an unseen number of sources.

	Training	**Validation**	**Testing**
Spatial component	Hearpiece [42]	Hearpiece [42]	Hearpiece [42]
			Hearpiece [42]
	MYRiAD [43]	MYRiAD [43]	MYRiAD [43]
Number subjects	20/26	3/26 unseen	3/26 unseen
			3/26 unseen
	1/1	1/1	1/1
Number source positions	25	25 seen	25 seen
			15 unseen
	18	18 seen	6 unseen
Number noise sources	1	1	1
			2
			2

TABLE I
SUMMARY OF THE SPATIAL COMPONENTS USED TO CONSTRUCT THE DATASET ORIGINATING FROM THE HEARPIECE DATASET WITH SAME POSSIBLE SOURCE POSITIONS IN TRAINING AND TEST SET (TOP), THE DATASET ORIGINATING FROM THE HEARPIECE DATASET WITH DISTINCT SOURCE POSITIONS IN TRAINING AND TEST SET (MIDDLE), THE DATASET ORIGINATING FROM THE MYRIAD DATASET (BOTTOM). THE TERMINOLOGY 'SEEN' AND 'UNSEEN' ARE USED WITH RESPECT TO THE TRAINING SET.

The training and validation set generated from the MYRiAD dataset are created from impulse responses from eighteen loudspeakers to the two BTE hearing instrument microphones at each side of the dummy head in the P1 position, as detailed in [43]. Again, one desired speech and one noise source are used. The test set generated from the MYRiAD dataset uses six unseen loudspeakers, one desired speech source and two noise sources. Table I summarises the spatial components of the three generated datasets.

The speech and noise files are selected from the 28-speaker dataset of [44], but remixed for more challenging signal-to-noise ratios (SNRs) and to incorporate the source-to-microphone propagation. Within the dataset of [44], the desired speech source signal is represented by the centre for speech technology voice cloning toolkit (CSTR VCTK) sentences [45], and the noise is represented by audio sampled from the diverse environments multichannel acoustic noise database (DEMAND) [46]. The root mean square (RMS) value of the desired speech source signal is set to one. After convolving the source audio with the impulse responses, the noise is rescaled to achieve an SNR at reference microphone r as sampled from a uniform distribution $U[-5\,\text{dB}, 5\,\text{dB}]$ for the training and validation sets, and to achieve an SNR of $0\,\text{dB}$ for the test sets. The speech and noise files were distinct for the training, validation, and test set.

B. Algorithm Settings

A DFT size of $20\,\text{ms}$, i.e., $F = 320$ is employed as per [9], together with a hop size of $10\,\text{ms}$, a square root Hamming window, and a sampling frequency of $16 \cdot 10^3\text{Hz}$. The network architecture corresponds to a pruned version of the causal convolutional recurrent network (CRN) of [9] with convolutional neural network (CNN), long short-term

memory (LSTM), transposed CNN (CNN-Trans), and batch normalisation (BN) layers as shown in Fig. 2. This pruning consists of reducing the number of CNN and CNN-Trans layers and using smaller channel sizes to reduce the overall number of parameters. The real and imaginary part of the STFT-transformed microphone signals are stacked one after the other, and used as input to the neural network, where only the first 161 frequency bins are retained due to the conjugate symmetry of the STFT. The number of output channels C, as well as the output activation functions, are varied based on the estimated variable as discussed in detail in Section V.

Either one neural network $\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.)$ is used to jointly estimate $\{\hat{\Theta}_1[f,k],\hat{\Theta}_2[f,k]\}$, such that both parameters arise from a common feature space, or separate networks are used to estimate $\hat{\Theta}_1[f,k]$ and $\hat{\Theta}_2[f,k]$, i.e.,

$$\text{NN}^{\{\hat{\Theta}_1,\hat{\Theta}_2\}}(.) = \begin{bmatrix} \text{NN}^{\{\hat{\Theta}_1\}}(.) \\ \text{NN}^{\{\hat{\Theta}_2\}}(.) \end{bmatrix}. \tag{35}$$

In both cases the architecture of Fig. 2 is used, such that the case with one neural network has approximately half the number of parameters of the two neural network case, i.e., approximately 750k parameters compared to approximately 1.5M parameters. Subscripts $._1$ and $._2$ are used to distinguish between both cases. The number of neural network parameters does not differ substantially between the different parameter estimation methods. For example, with one neural network, the smoothing estimation has the smallest number of parameters with $749,254$ parameters, while the eigen estimation involving the estimation of $\hat{V}[f,k]$ or $\hat{Q}[f,k]$ has the largest number of parameters with $750,046$ parameters.

The Adam optimiser is used with an initial learning rate of $2 \cdot 10^{-4}$, a decay factor of 0.98 every 2 epochs, and a maximum of 50 epochs. Early stopping is used with a 5-epoch patience. The batch size is set to 8, and gradient norms exceeding 1 are clipped. For the model-based GEVD-MWF, either an oracle time-domain VAD is assumed available, as denoted by $.\text{VAD}_s$, or a power-based time-domain VAD is calculated on the microphone signal of microphone r, as denoted by $.\text{VAD}_m$. For the model-based GEVD-MWF baseline, $\hat{\lambda}_m[f,k] = 0.995, \forall f$ when $\text{VAD}[k] = 1$ and $\hat{\lambda}_m[f,k] = 1, \forall f$ when $\text{VAD}[k] = 0$, as per [47]. Similarly, $\hat{\lambda}_n[f,k] = 0.995, \forall f$ when $\text{VAD}[k] = 0$ and $\hat{\lambda}_n[f,k] = 1, \forall f$ when $\text{VAD}[k] = 1$, as per [47]. To ensure well-flowing gradients, $\hat{R}_{mm}[f,k]$ and $\hat{R}_{nn}[f,k]$ are regularised with $\text{diag}(\begin{bmatrix}1.2 & 1.4\end{bmatrix}) \cdot 10^{-6}$ and $\text{diag}(\begin{bmatrix}1.1 & 0.9\end{bmatrix}) \cdot 10^{-6}$, containing distinct numbers, respectively, to promote distinct and positive eigenvalues. The first microphone is selected as reference microphone ($r = 1$). To enable a fair comparison, all filters and neural networks are optimised with respect to the MSE cost (3).

C. Performance Measures

Noise reduction performance is measured by means of the signal-to-noise ratio (SNR) improvement ΔSNR and speech distortion SD, which are calculated as

$$\Delta\text{SNR} = \text{SNR}^{(\text{out})} - \text{SNR}^{(\text{in})} \tag{36a}$$

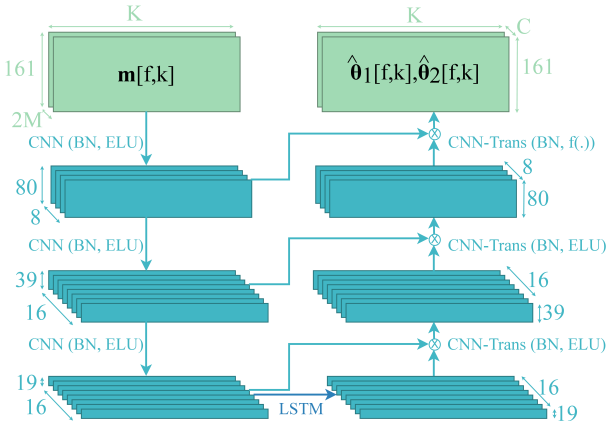

Fig. 2. The neural network architecture is a pruned version of the convolutional recurrent network (CRN) of [9].

$$= 10\log_{10}\left(\frac{P_s^{(\text{out})}}{P_n^{(\text{out})}}\right) - 10\log_{10}\left(\frac{P_s^{(\text{in})}}{P_n^{(\text{in})}}\right), \tag{36b}$$

$$\text{SD} = 10\log_{10}\left(\frac{P_s^{(\text{out})}}{P_s^{(\text{in})}}\right), \tag{36c}$$

with $P_s^{(\text{out})}$ and $P_n^{(\text{out})}$ (resp. $P_s^{(\text{in})}$ and $P_n^{(\text{in})}$) referring to the output (resp. input) power of the desired speech and noise in reference microphone r after (resp. before) applying the noise reduction filter of (11), (12), or (14). The perceptual evaluation of speech quality (PESQ) improvement ΔPESQ [48] is also computed as

$$\begin{aligned}\Delta\text{PESQ} = &\text{PESQ}(\hat{\mathbf{w}}_r[f,k]^{\mathsf{T}}\mathbf{m}[f,k], s_r[f,k]) \\ &- \text{PESQ}(\mathbf{t}_r^{\mathsf{T}}\mathbf{m}[f,k], s_r[f,k]).\end{aligned} \tag{37}$$

To focus on steady-state performance, the metrics are calculated on the second half of the signals.

VII. Results And Discussion

A. Performance

Figs. 3, 4, and 5 show the performance of the proposed hybrid GEVD-MWF variants (Section V), the model-based baselines (Section III), represented by the GEVD-MWF with oracle VAD knowledge and the GEVD-MWF with a VAD based on the microphone signals, and the data-driven baseline (Section IV), represented by the mask estimation. To this end, for Fig. 3 the Hearpiece dataset with same desired speech and noise source positions as during training and with one noise source is used, while for Fig. 4 and Fig. 5 distinct desired speech and noise source positions and two noise sources are used for the Hearpiece dataset and MYRiAD dataset, respectively. For the eigen estimation, the estimation of $\{\hat{Q}[f,k], \hat{\sigma}_m^{(1)}\}$ (Section V-A2) is shown, and for the correlation estimation, the estimation of the Cholesky factors $\{\hat{\Gamma}_{mm}[f,k], \hat{\Gamma}_{nn}[f,k]\}$ (Section V-B1) is shown.

Using the smoothing estimation (Section V-C) generally attains the best performance across the datasets, illustrated by the highest ΔSNR and ΔPESQ in both Hearpiece datasets, and the highest ΔSNR in the MYRiAD dataset. Only regarding the ΔPESQ in the MYRiAD dataset, mask estimation outperforms

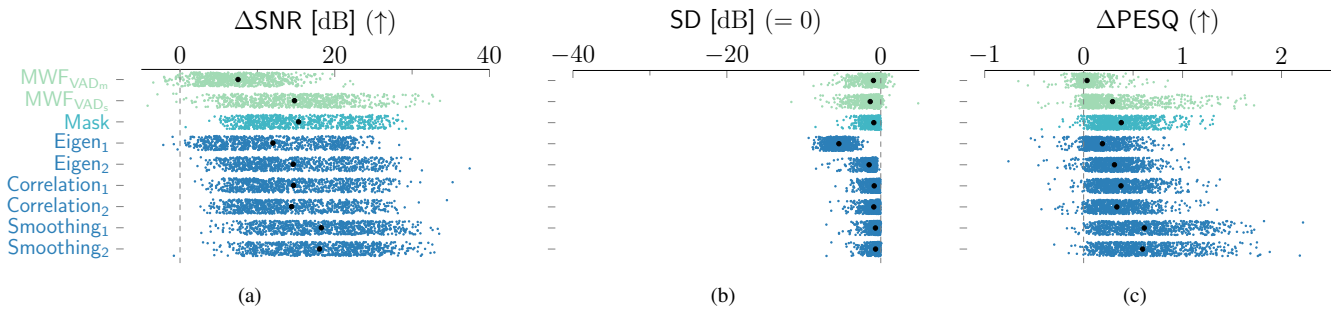

Fig. 3. Performance comparison on the Hearpiece dataset with the same possible desired speech and noise source positions in the test set as in the training set, and with one noise source. The mean performance is represented with a black dot, and the individual datapoints are represented with coloured dots.

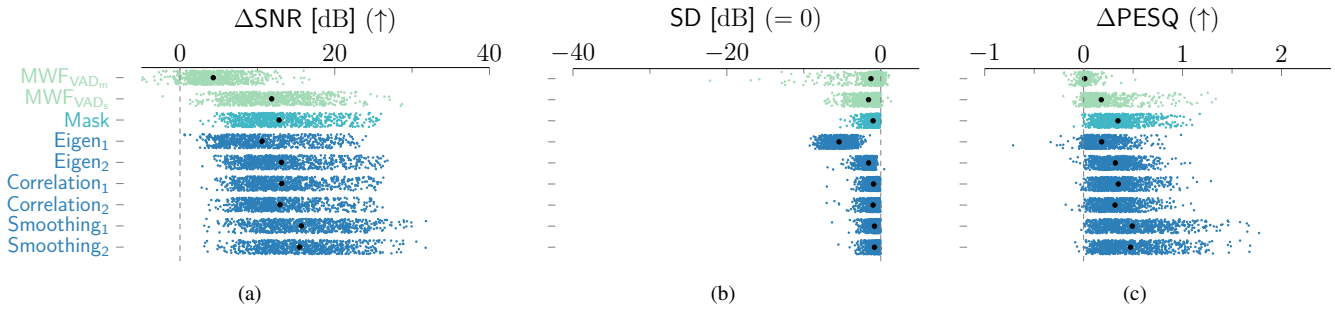

Fig. 4. Performance comparison on the Hearpiece dataset with different possible desired speech and noise source positions in the test set than in the training set, and with two noise sources. The mean performance is represented with a black dot, and the individual datapoints are represented with coloured dots.

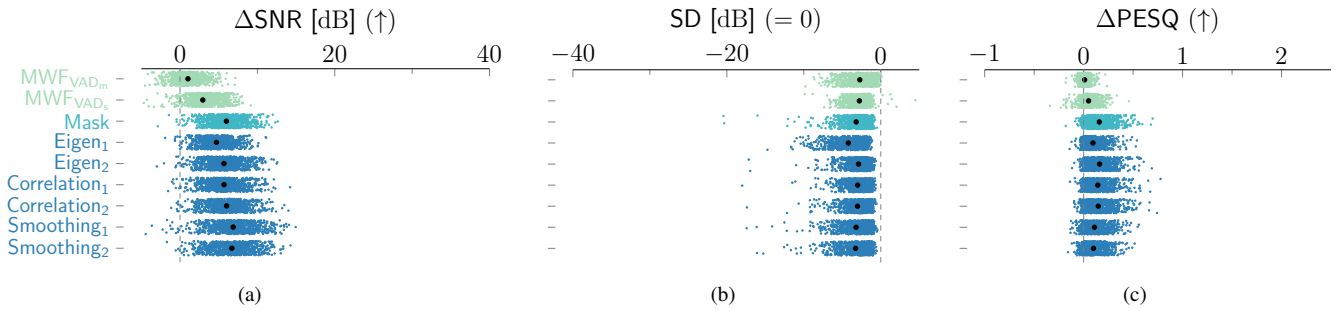

Fig. 5. Performance comparison on the MYRiAD dataset with different possible desired speech and noise source positions in the test set than in the training set, and with two noise sources. The mean performance is represented with a black dot, and the individual datapoints are represented with coloured dots.

the smoothing estimation due to a slight increase in speech distortion for the smoothing estimation. Despite the mask estimation having the most degrees of freedom as no model structure is imposed, the model structure, imposed in the hybrid algorithm variants, seems to aid the training, thereby obtaining improved performance of the hybrid GEVD-MWF with respect to its data-driven counterpart. This observation is consistent with [26], where for the multi-frame MVDR for single-microphone noise reduction, the hybrid algorithms also showed improved performance compared to the mask estimation. As the smoothing estimation generally outperforms the correlation estimation and the eigen estimation, an increased level of model structure seems beneficial. This is in contrast with the results of [26], where the correlation estimation was found to achieve higher performance than the smoothing estimation, although a different model structure was used in [26], i.e., a multi-frame MVDR for single-microphone noise reduction rather than a single-frame GEVD-MWF for multi-microphone noise reduction. Both hybrid and mask estimation algorithms also generally outperform the model-based GEVD-MWF with both the oracle VAD and the VAD computed on the microphone signals, certainly when the scenarios become more challenging due to the use of unseen source positions, two noise sources, and increased reverberation. Indeed, the parameters of the hybrid (and mask estimation) algorithms are only constrained to be valid parameters, while they have been optimised on the training set. The GEVD-MWF requires a user-defined smoothing variable, which when governed by a time-domain VAD also does not vary across frequency, while, e.g., the smoothing estimation does not require user-defined settings and can vary its parameters across frequency and time. Consequently, the smoothing estimation might outperform the model-based GEVD-MWF, in particular when using a VAD computed on the microphone signals, and the other hybrid

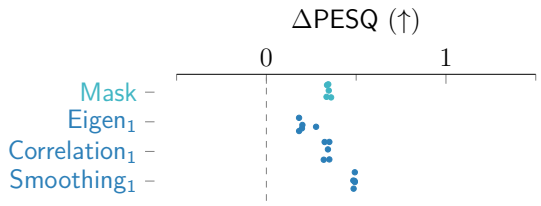

Fig. 6. Mean PESQ improvement on the Hearpiece dataset with distinct source positions in the training and test set, and two noise sources, across different neural network initialisations.

methods, which do not explicitly use a VAD, due to the use of accurate smoothing parameters optimised on the training set.

When comparing the performance on both Hearpiece datasets, the mean performance drops slightly, as expected, due to the signals arriving from unseen spatial positions and due to the use of an increased number of noise sources. Nevertheless, results show a good generalisation across spatial positions and number of sources, and relative performance between the algorithms generally remains the same.

Further, although relative performance differences might be small for some algorithms, this difference is reproducible across different network initialisations as shown for the ΔPESQ in Fig. 6.

Finally, for the correlation estimation and smoothing estimation, the performance does not change significantly whether one neural network is used to ouput both $\hat{\Theta}_1[f,k]$ and $\hat{\Theta}_2[f,k]$, or whether separate networks are used for each individual parameter. It appears a common feature space can be found from which both parameters can be estimated effectively. However, for the eigen estimation, this does not seem to hold true, as there is a larger drop in performance when the same network is used to estimate both generalised eigenvectors and eigenvalues as compared to using separate networks to estimate those parameters.

B. Context

As the neural network architecture employs an LSTM layer, theoretically, the context used to estimate the parameters for frame k is infinite, i.e., $L_K = k$ in (13a). In order to study how much context is effectively used, the parameter L_K can be fixed, such that only the frames in range between frame $k - L_K + 1$ and frame k are provided to the network to estimate the parameters for frame k. The states of the LSTM layer are subsequently reset before moving to the next frame. The performance on the Hearpiece dataset with distinct source positions and two noise sources is shown in Fig. 7.

The performance generally reaches a plateau around $L_K = 60$ frames. The relative performance difference between the different algorithms also remains the same as in Fig. 4 for the L_K-range under study. The eigen estimation using one neural network, which in Fig. 4 achieves lower performance than the other hybrid algorithms, indeed also achieves lower performance than its counterparts for $L_K > 60$. However, it retains this performance across the entire L_K-range, thereby reaching higher performance than the other algorithms for $L_K < 60$. Possibly, although learning a common feature space for both generalised eigenvectors and eigenvalues appears difficult, the network appears to have focused on a limited set of features which prove to be robust to variations in L_K.

C. Eigen And Correlation Estimation

Fig. 8 and Fig. 9 show the performance of the same experiment as in Section VII-B, yet for the different eigen estimation (Section V-A) and different correlation estimation (Section V-B) methods.

For both the eigen estimation and the correlation estimation, the relative difference between the different estimation methods is small, such that the specific estimation method used seems less critical than selecting an appropriate pair of parameters $\{\hat{\Theta}_1[f,k], \hat{\Theta}_2[f,k]\}$ to estimate, as detailed in Section VII-A. The correlation estimation with regularisation generally, slightly outperforms the correlation estimation with Cholesky estimation and manifold projection in terms of ΔSNR and SD when $L_K < 60$ both when using one network to estimate the two correlation matrices and when using one network for each correlation matrix. When $L_K \geq 60$, the relative performance difference between the different correlation estimation methods diminishes. The performance of the eigen estimation with one neural network remains constant across the entire L_K-range regardless of the estimation method used. As opposed to the different correlation estimation methods, no eigen estimation method consistently outperforms the other ones across the number of neural networks used for the estimation and a particular L_K-range.

D. Interpretability

The estimated parameters in the hybrid algorithms are solely constrained to be valid parameters, and are optimised for performance on the training set when optimising the neural network in an end-to-end fashion. The estimated parameters, nevertheless, show high similarity with the clean speech spectrogram. To this end, Fig. 10 shows an example spectrogram of the ground-truth desired speech and microphone signals at microphone r, the estimated desired speech generalised eigenvalue using the eigen estimation of $\{\hat{Q}[f,k], \hat{\sigma}_m^{(1)}\}$, the estimated desired speech power in microphone r with the correlation estimation using the estimation of the Cholesky factors, and the noise smoothing variable. For this Fig. 10, the Hearpiece dataset with distinct source positions and two noise sources has been used. Indeed, these parameters exhibit a strong correspondence with the spectrogram of the desired speech. Possibly, to optimally make use of the provided GEVD-MWF model structure, the parameters that yield optimal performance when trained end-to-end are in agreement with this structure. For example, referring to (7), in a model-based approach, $\hat{R}_{nn}[f,k]$ is not expected to be updated when there is activity of the desired speech. In Fig. 10, even though $\hat{\lambda}_n[f,k]$ has been trained in an end-to-end fashion, it shows the same behaviour. The eigen estimation using a single neural network shows less correspondence with the spectrogram of the desired speech than the eigen estimation with two neural networks, which is in accordance with the difference in performance as discussed in Section VII-A.

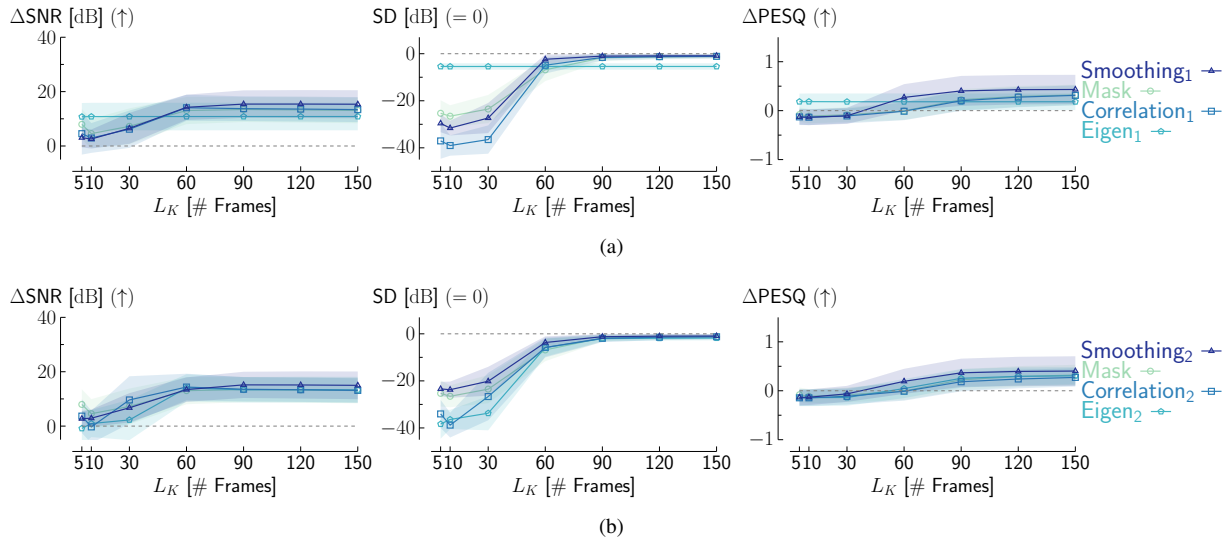

Fig. 7. Performance as a function of the context ranging from frame $k - L_K + 1$ to frame k to estimate the parameters at frame k. Mean performance is shown as a solid line, and the standard deviation as shading.

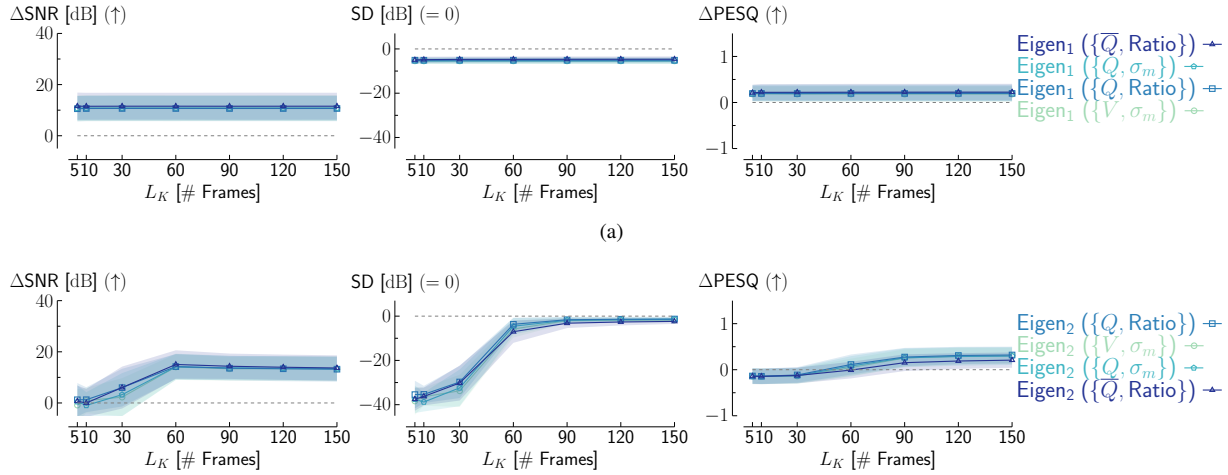

Fig. 8. Performance as a function of the context ranging from frame $k - L_K + 1$ to frame k to estimate the parameters at frame k for different eigen estimation methods. Mean performance is shown as a solid line, and the standard deviation as shading.

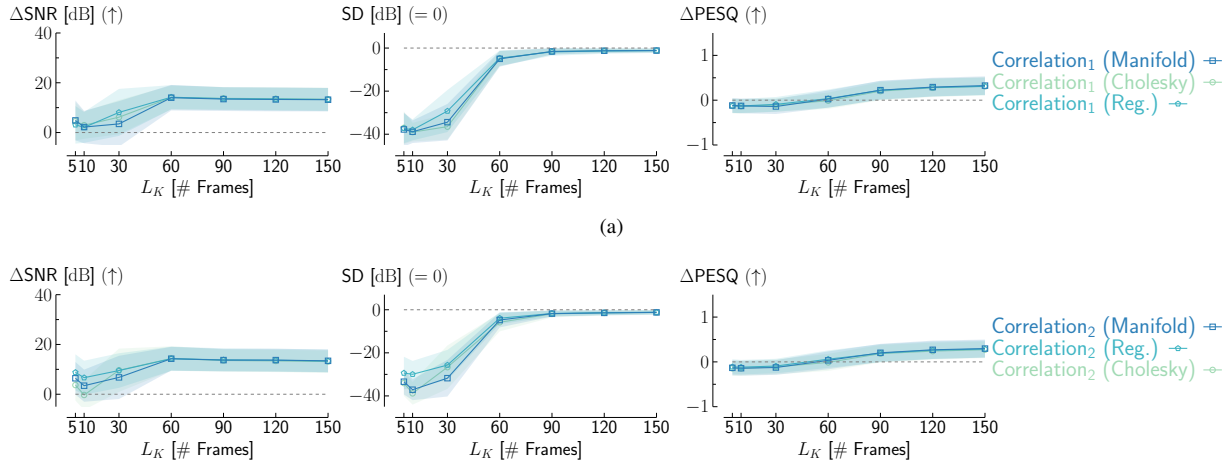

Fig. 9. Performance as a function of the context ranging from frame $k - L_K + 1$ to frame k to estimate the parameters at frame k for different correlation estimation methods. Mean performance is shown as a solid line, and the standard deviation as shading.

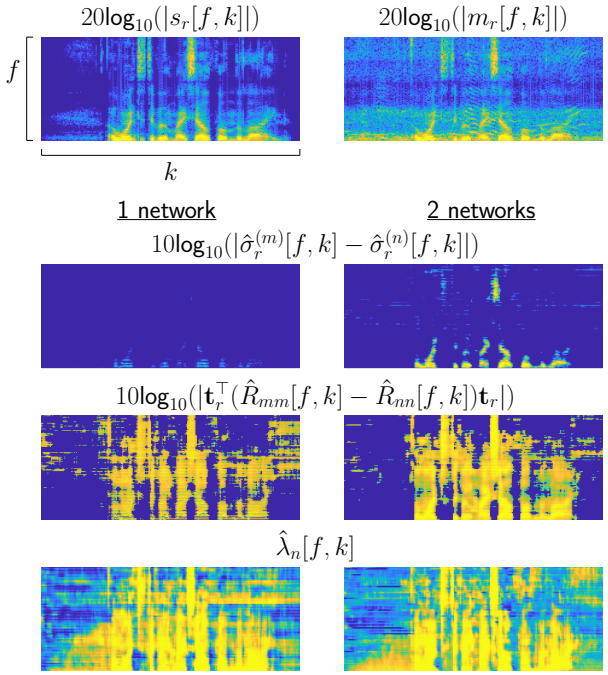

Fig. 10. The parameters, as estimated by the neural network in the hybrid design, show remarkable correspondence with the spectrogram of the clean speech signal of length 2.35 s and Nyquist frequency $\frac{f_s}{2}$ = 8 kHz.

VIII. Conclusion

In this paper, hybrid algorithm design has been applied to multi-microphone noise reduction using the generalised eigenvalue decomposition multi-channel Wiener filter (GEVD-MWF), where the generalised eigenvectors and eigenvalues, correlation matrices, or smoothing variables are estimated using a neural network. This neural network is optimised in an end-to-end fashion, i.e., the neural network is optimised for performance on the noise reduction task. To this end, the parameters as provided by the neural network are inserted in the model-structure of the GEVD-MWF and applied to the microphone signals to achieve noise reduction. The cost associated with this noise reduction task is then utilised to optimise the neural network parameters for training set performance. Correspondingly, while the model-parameters are enforced to be valid parameters by properly selecting output activation functions and dimensions of the neural network, the goal is not to provide optimal estimates for these parameters, but rather to achieve optimal noise reduction. The resulting hybrid algorithms interpolate between model-based estimation of the GEVD-MWF, where no neural networks are used, and data-driven mask estimation, where the filter coefficients are directly estimated.

It has been shown that the hybrid GEVD-MWF algorithms can outperform both the model-based GEVD-MWF and even the data-driven mask estimation, in particular using the smoothing estimation, i.e., the hybrid algorithms seem to benefit from using the model structure. Although the neural network is optimised end-to-end for performance on the training set, the estimated model-parameters show a degree of similarity with the spectrogram of the clean speech.

References

[1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*, 1st ed., ser. Signals and Communication Technology. Berlin, Heidelberg: Springer Berlin, Heidelberg, Mar. 2006.

[2] R. Haeb-Umbach, T. Nakatani, M. Delcroix, C. Boeddeker, and T. Ochiai, "Microphone array signal processing and deep learning for speech enhancement: Combining model-based and data-driven approaches to parameter estimation and filtering," *IEEE Signal Process. Mag.*, vol. 41, no. 6, pp. 12–23, Jan. 2024.

[3] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[4] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1256–1269, Oct. 2012.

[5] R. Lorenz and S. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1684–1696, May 2005.

[6] S. Gannot, "Speech processing utilizing the Kalman filter," *IEEE Instrum. Meas. Mag.*, vol. 15, no. 3, pp. 10–14, Jun. 2012.

[7] 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, Apr. 2014.

[8] W. Jiang, C. Sun, F. Chen, Y. Leng, and Q. Guo, "A novel skip connection mechanism based on channel-wise cross transformer for speech enhancement," *Multimedia Tools Appl.*, vol. 83, no. 12, pp. 34 849–34 866, Sep. 2023.

[9] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech 2018*, Hyderabad, India, Sep. 2018, pp. 3229–3233.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, Oct. 2015.

[11] C.-L. Liu, S.-W. Fu, Y.-J. Li, J.-W. Huang, H.-M. Wang, and Y. Tsao, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1888–1900, 2020.

[12] X. Li and R. Horaud, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *Proc. 2019 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2019, pp. 298–302.

[13] R. Miccini, C. Laroche, T. Piechowiak, and L. Pezzarossa, "Scalable speech enhancement with dynamic channel pruning," in *Proc. 2025 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Hyderabad, India, Apr. 2025, pp. 1–5.

[14] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 8, pp. 1296–1305, Jun. 2014.

[15] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense U-net for multichannel speech enhancement," in *Proc. 2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 836–840.

[16] X. Zeng, X. Zhang, and M. Wang, "A feature integration network for multi-channel speech enhancement," *Sensors*, vol. 24, no. 22, p. 7344, 2024.

[17] M. M. Halimeh and W. Kellermann, "Complex-valued spatial autoencoders for multichannel speech enhancement," in *Proc. 2022 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Singapore, Singapore, May 2022, pp. 261–265.

[18] X. Ren, X. Zhang, L. Chen, X. Zheng, C. Zhang, L. Guo, and B. Yu, "A causal U-net based neural beamforming network for real-time multi-channel speech enhancement," in *Proc. Interspeech 2021*, Brno, Czechia, Aug. 2021, pp. 1832–1836.

[19] W. Dai, X. Li, A. Politis, and T. Virtanen, "Reference channel selection by multi-channel masking for end-to-end multi-channel speech enhancement," in *Proc. 2024 32nd Eur. Signal Process. Conf. (EUSIPCO 2024)*, Lyon, France, Aug. 2024, pp. 241–245.

[20] Z. Hou, Q. Hu, T. Sun, Y. Hu, C. Zhu, and K. Chen, "Convolutional recurrent metriCGAN with spectral dimension compression for full-band speech enhancement," in *Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–2.

[21] A. Nurali, J. Azcarreta, A. Pandey, S. Lee, A. Aroudi, and C. Bilen, "A three-stage deep learning framework for real-time speech enhancement," in *Proc. Interspeech 2025*, Rotterdam, The Netherlands, Aug. 2025, pp. 1193–1197.

[22] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends Hear.*, vol. 27, p. 23312165231209913, 2023.

[23] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. 2016 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016.

[24] S. Tao, P. Mowlaee, J. R. Jensen, and M. Græsbøll Christensen, "Learning-based multi-channel speech presence probability estimation using a low-parameter model and integration with MVDR beamforming for multi-channel speech enhancement," in *Proc. 2024 Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Aalborg, Denmark, Sep. 2024, pp. 100–104.

[25] M. Tammen, D. Fischer, B. T. Meyer, and S. Doclo, "DNN-based speech presence probability estimation for multi-frame single-microphone speech enhancement," in *Proc. 2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 191–195.

[26] M. Tammen and S. Doclo, "Parameter estimation procedures for deep multi-frame MVDR filtering for single-microphone speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3237–3248, Aug. 2023.

[27] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *Proc. 2021 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6089–6093.

[28] C. He, C. Zhou, Z. Shi, and J. Chen, "Deep INCM reconstruction for adaptive beamforming," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Seoul, Republic of Korea, Apr. 2024, pp. 13 491–13 495.

[29] Y. Luo, "Streaming multi-channel speech separation with online time-domain generalized Wiener filter," in *Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.

[30] M. Tammen and S. Doclo, "Imposing correlation structures for deep binaural spatio-temporal Wiener filtering," *IEEE Trans. Audio Speech Lang. Process.*, pp. 1–15, Mar. 2025.

[31] Z. Cao and D. Li, "Beamforming and lightweight GRU neural network combination model for multi-channel speech enhancement," *Signal Image Video Process.*, vol. 18, no. 8, pp. 5677–5683, Sep. 2024.

[32] Z. Xu, A. Aroudi, K. Tan, A. Pandey, J.-S. Lee, B. Xu, and F. Nesta, "FoVNet: Configurable field-of-view speech enhancement with low computation and distortion for smart glasses," in *Proc. Interspeech 2024*, Kos, Greece, Sep. 2024, pp. 3350–3354.

[33] E. Grinstein, A. Pandey, C. Li, S. Srinivas, J. Azcarreta, J. Donley, S. Lee, A. Aroudi, and C. Bilen, "Controlling the parameterized multi-channel Wiener filter using a tiny neural network," in *Proc. 2025 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2025.

[34] T.-A. Hsieh, J. Donley, D. Wong, B. Xu, and A. Pandey, "On the importance of neural Wiener filter for resource efficient multichannel speech enhancement," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Seoul, Republic of Korea, Apr. 2024, pp. 12 181–12 185.

[35] C.-H. Lee, K. Patel, C. Yang, Y. Shen, and H. Jin, "An MVDR-embedded U-net beamformer for effective and robust multichannel speech enhancement," in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Seoul, Republic of Korea, Apr. 2024, pp. 8541–8545.

[36] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "Deep multi-frame filtering for hearing aids," in *Proc. Interspeech 2023*, Dublin, Ireland, Aug. 2023, pp. 3869–3873.

[37] N. Fleischhacker, "The tikzpeople package," 2017. [Online]. Available: https://ctan.org/pkg/tikzpeople

[38] A. Roebben, "A hybrid design for noise reduction with the generalised eigenvalue decomposition multi-channel Wiener filter," GitHub repository, 2025. [Online]. Available: https://cloud.esat.kuleuven.be/index.php/s/6iWPDcbBmSSnBM7

[39] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the solution of algebraic eigenvalue problems*, 1st ed., ser. Software, Environments, and Tools. Philadelphia, PA, U.S.A.: SIAM, Oct. 2000.

[40] G. H. Golub and C. F. Van Loan, *Matrix computations*, 4th ed., ser. Johns Hopkins studies in the mathematical sciences. Baltimore, Maryland, USA: The Johns Hopkins University Press, Feb. 2013.

[41] N. J. Higham, "Computing the nearest correlation matrix—a problem from finance," *IMA J. Numer. Anal.*, vol. 22, no. 3, pp. 329–343, Jul. 2002.

[42] F. Denk and B. Kollmeier, "The Hearpiece database of individual transfer functions of an openly available in-the-ear earpiece for hearing device research," *Acta Acust.*, vol. 5, no. 2, p. 2, Dec. 2021.

[43] T. Dietzen, R. Ali, M. Taseska, and T. van Waterschoot, "MYRiAD: A multi-array room acoustic database," *EURASIP J. Audio Speech Music Process.*, vol. 2023, no. 17, p. 17, Apr. 2023.

[44] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," Dataset available at: https://doi.org/10.7488/ds/2117, Aug. 2017.

[45] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Dataset available at: https://doi.org/10.7488/ds/2645, Nov. 2019.

[46] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meet. Acoust.*, vol. 19, no. 1, Jun. 2013, p. 035081.

[47] R. Ali, T. van Waterschoot, and M. Moonen, "A noise reduction strategy for hearing devices using an external microphone," in *Proc. 2017 25th Eur. Signal Process. Conf. (EUSIPCO 2017)*, Kos, Greece, Aug. 2017, pp. 1–5.

[48] T. S. S. I.-T. International Telecommunication Union, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," P.862.2, Geneva, Switzerland, 2007.