Parameter Estimation Procedures for Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement

Marvin Tammen, *Student Member, IEEE*, and Simon Doclo, *Senior Member, IEEE*

***Abstract*—Aiming at exploiting temporal correlations across consecutive time frames in the short-time Fourier transform (STFT) domain, multi-frame algorithms for single-microphone speech enhancement have been proposed. Typically, the multi-frame filter coefficients are either estimated directly using deep neural networks or a certain filter structure is imposed, e.g., the multi-frame minimum variance distortionless response (MFMVDR) filter structure. Recently, it was shown that integrating the fully differentiable MFMVDR filter into an end-to-end supervised learning framework employing temporal convolutional networks (TCNs) allows for a high estimation accuracy of the required parameters, i.e., the speech inter-frame correlation vector and the interference covariance matrix. In this paper, we investigate different covariance matrix structures, namely Hermitian positive-definite, Hermitian positive-definite Toeplitz, and rank-1. The main differences between the considered matrix structures lie in the number of parameters that need to be estimated by the TCNs as well as the required linear algebra operations. For example, assuming a rank-1 matrix structure, we show that the MFMVDR filter can be written as a linear combination of the TCN outputs, significantly reducing computational complexity. In addition, we consider a covariance matrix estimation procedure based on recursive smoothing. Experimental results on the deep noise suppression challenge dataset show that the estimation procedure using the Hermitian positive-definite matrix structure yields the best performance, closely followed by the rank-1 matrix structure at a much lower complexity. Furthermore, imposing the MFMVDR filter structure instead of directly estimating the multi-frame filter coefficients slightly but consistently improves the speech enhancement performance.**

***Index Terms*—Matrix structures, multi-frame filtering, MVDR filter, speech enhancement, supervised learning.**

I. Introduction

In MANY speech communication systems such as hearing aids, mobile phones, and smart speakers, the microphones pick up ambient noise in addition to the desired speech signal, often leading to a degradation of speech quality and speech intelligibility. Hence, a large variety of single- and multi-microphone speech enhancement algorithms have been proposed [1], [2], [3], [4], [5]. Typically, single-microphone speech enhancement algorithms first apply a transform to the noisy time domain signal, either a signal-independent transform such as the short-time Fourier transform (STFT) or an auditory-inspired filterbank [1], [2], [6], a signal-dependent transform such as the Karhunen-Loève transform [7], [8], or a learned transform [9].

In the STFT domain, it is frequently assumed that adjacent STFT coefficients are uncorrelated over time and frequency, which is a suitable assumption when considering sufficiently long time frames. In this case, the clean speech STFT coefficients can be estimated by simply applying a gain (commonly called mask) to the noisy STFT coefficients. Assuming the phase of the speech STFT coefficients to be uniformly distributed, it has been shown that the minimum mean square error estimate of the clean speech phase is the noisy phase [10], resulting in a real-valued mask. The latter approaches mainly differ in the considered network architecture and in the target definition, e.g., mask approximation, signal approximation, or a-priori signal-to-noise-ratio (SNR) approximation. Supervised learning-based approaches employing a real-valued mask have also been proposed for other filterbanks, e.g., a multi-phase gammatone filterbank [6] or a learned filterbank [9], [11]. The popular Conv-TasNet algorithm [9] combines an end-to-end encoder-masking-decoder structure with a deep neural network (DNN) architecture based on temporal convolutional networks (TCNs) [12] and the scale-invariant signal-to-distortion-ratio (SI-SDR) loss function, which is defined in the time domain. The learned linear encoder first transforms the time domain noisy microphone signal into a hidden representation that is optimized for speech separation. The speakers are then separated by applying real-valued masks to the hidden representation. Finally, the estimated hidden representations of the individual speakers are transformed back to the time domain using a learned linear decoder. Instead of applying a real-valued mask to the noisy STFT coefficients, complex-valued masking algorithms have been proposed, which aim at not only improving the amplitude but also the phase. To estimate this complex-valued mask, both statistical model-based approaches [13] as well as supervised learning-based approaches [14], [15], [16], [17], [18], [19] have been proposed.

Aiming at exploiting temporal correlations, algorithms have been proposed which can be broadly separated into two

Manuscript received 28 November 2022; revised 22 May 2023 and 24 July 2023; accepted 4 August 2023. Date of publication 18 August 2023; date of current version 28 August 2023. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through Germany's Excellence Strategy - EXC 2177/1 - under Grant 390895286. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Keisuke Kinoshita. *(Corresponding author: Marvin Tammen.)*

The authors are with the Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4all, University of Oldenburg, 26129 Oldenburg, Germany (e-mail: marvin.tammen@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

Digital Object Identifier 10.1109/TASLP.2023.3306715

categories. First, implicitly exploiting temporal correlations across consecutive STFT frames, supervised learning-based complex spectral mapping algorithms have been proposed, which exploit spectro-temporal patterns in the noisy microphone signal to directly estimate the clean speech STFT coefficients [20]. Second, explicitly exploiting temporal correlations across consecutive STFT frames, multi-frame algorithms for single-microphone speech enhancement have been proposed, which apply a (complex-valued) filter instead of a gain to the STFT coefficients. In this article, we focus on the last category of algorithms. In [21], an estimate of the clean speech STFT coefficients was obtained by applying a spectro-temporal filter to a neighborhood of the noisy STFT coefficients. The filter coefficients were directly estimated using a DNN with bidirectional long short-term memory layers. Alternatively, it has been proposed to impose a certain structure on the spectro-temporal filter. Similarly to the minimum variance distortionless response (MVDR) beamformer [3], [22], [23] for multi-microphone scenarios, a frequently used structure for single-microphone scenarios is the multi-frame minimum variance distortionless response (MFMVDR) filter [1], [24]. The MFMVDR filter estimates the clean speech STFT coefficients by minimizing the output interference power while preserving the speech inter-frame correlation (IFC) [1], [24], thereby requiring estimates of the highly time-varying speech IFC vector and the interference covariance matrix. Although it has been shown that the MFMVDR filter yields a very good noise reduction with little speech distortion provided that accurate estimates of these quantities are available, its performance is rather sensitive to estimation errors, especially in the speech IFC vector [25]. To estimate the speech IFC vector and the interference covariance matrix from the noisy STFT coefficients, both statistical model-based as well as supervised learning-based approaches have been proposed. In [26], a maximum likelihood estimator for the speech IFC vector has been proposed, assuming a white Gaussian interference and requiring an estimate of the a-priori SNR. Alternatively, in [27] the speech IFC vector was estimated based on a low-rank model of the speech covariance matrix. Conventionally, the interference covariance matrix is estimated using a recursive smoothing procedure, requiring an estimate of the speech presence probability (SPP) [28].

In [29] we proposed a supervised learning-based approach to estimate all required quantities of the MFMVDR filter by integrating the fully differentiable MFMVDR filter into an end-to-end supervised learning framework using TCNs. Instead of using a loss defined on the quantities, the TCNs were trained by minimizing the SI-SDR loss function at the output of the MFMVDR filter. It was shown that this estimation approach yields an improved speech enhancement performance compared with statistical model-based estimators. Similarly, in [23] a supervised learning-based approach to estimate all required quantities of the multi-microphone multi-frame MVDR filter was proposed. In contrast to the single-microphone algorithm in [29], the multi-microphone algorithm in [23] additionally uses direction of arrival features. Estimates of the speech and noise components are first obtained by applying complex-valued multi-microphone multi-frame filters to the noisy STFT coefficients. These estimates are then used to compute estimates of the instantaneous spatio-temporal covariance matrices, which are mapped to the quantities required by the multi-microphone multi-frame MVDR filter using DNNs.

Whereas in [29] the covariance matrices were assumed to be Hermitian positive-definite, in this article we investigate different matrix structures for the covariance matrices, namely Hermitian positive-definite, Hermitian positive-definite Toeplitz, and rank-1. For the Hermitian positive-definite matrix structure, we consider an estimation procedure based on recursive smoothing, where only the smoothing factors are estimated using TCNs, as well as an estimation procedure based on the Cholesky decomposition. The main differences between the considered covariance matrix estimation procedures lie in the number of parameters that need to be estimated by the TCNs as well as the required linear algebra operations, yielding a different computational complexity. In the case of rank-1 covariance matrices, we show that the MFMVDR filter can be written as a linear combination of the TCN outputs, significantly reducing the computational complexity. Experimental results on the deep noise suppression (DNS) challenge dataset including stationary and non-stationary noise at SNRs ranging from 0 dB to 19 dB show that the covariance matrix estimation procedure using the Hermitian positive-definite matrix structure based on the Cholesky decomposition yields the best performance. Interestingly, the estimation procedure using the rank-1 matrix structure yields only a slightly lower performance, with the advantage of being computationally less demanding. Furthermore, it is shown for the best-performing MFMVDR filters that imposing the MFMVDR filter structure instead of directly estimating the multi-frame filter coefficients slightly but consistently improves the speech enhancement performance.

The remainder of the article is organized as follows. In Section II, we describe the signal model, define the MFMVDR filter, and introduce the considered matrix structures for the covariance matrices. In Section III, the conventional SPP-based supervised learning approach to estimate the quantities of the MFMVDR filter is reviewed. In Section IV, the proposed signal-based supervised learning-based approach to estimate the quantities of the MFMVDR filter is described, including multiple procedures to estimate the required covariance matrices. The simulation setup is discussed in Section V, and the corresponding simulation results are presented in Section VI.

II. Multi-Frame MVDR Filter for Single-Microphone Speech Enhancement

A. Signal Model

We consider an acoustic scenario in which a speech source and additive noise (e.g., traffic, keyboard typing, fan noise) are recorded by a single microphone. In the STFT domain, the noisy STFT coefficient $Y_{k,l}$ at the k-th frequency bin and l-th time frame is given by

$$Y_{k,l} = X_{k,l} + N_{k,l}, \tag{1}$$

where $X_{k,l}$ and $N_{k,l}$ denote the speech and noise STFT coefficient, respectively. Assuming independent frequency bins, each frequency bin will be processed separately, such that the index k will be omitted in the remainder of this article.

In many single-frame speech enhancement algorithms, a mask M_l is applied to the noisy STFT coefficient to obtain an estimate of the speech STFT coefficient, i.e.,

$$\widehat{X}_l = M_l Y_l, \tag{2}$$

where the mask M_l can be either real-valued [2], [30], [31], [32], [33], [34], [35], [36] or complex-valued [13], [14], [15], [16], [17], [18], [19]. In contrast, multi-frame speech enhancement algorithms apply a (complex-valued) filter to the N-dimensional noisy vector $\mathbf{y}_l$, defined as

$$\mathbf{y}_l = \begin{bmatrix} Y_l & Y_{l-1} & \dots & Y_{l-N+1} \end{bmatrix}^{\mathsf{T}}, \tag{3}$$

where $\circ^{\mathsf{T}}$ denotes the transpose operator and where we define $Y_l = 0 \; \forall \; l < 0$. An estimate of the speech STFT coefficient is obtained as

$$\widehat{X}_l = \sum_{\mu=0}^{N-1} W_{l,\mu}^* Y_{l-\mu} = \mathbf{w}_l^{\mathsf{H}} \mathbf{y}_l, \tag{4}$$

where N denotes the number of filter coefficients, $\circ^*$ denotes the complex-conjugate operator, $\circ^{\mathsf{H}}$ denotes the Hermitian transpose operator, and the multi-frame filter is defined as

$$\mathbf{w}_l = \begin{bmatrix} W_{l,0} & W_{l,1} & \dots & W_{l,N-1} \end{bmatrix}^{\mathsf{T}}, \tag{5}$$

where $W_{l,\mu}$ denotes the μ-th filter coefficient of $\mathbf{w}_l$.

Using (1), the noisy vector in (3) can be written as

$$\mathbf{y}_l = \mathbf{x}_l + \mathbf{n}_l, \tag{6}$$

where the speech and noise vectors $\mathbf{x}_l$ and $\mathbf{n}_l$ are defined similarly as $\mathbf{y}_l$. Assuming the speech and noise STFT coefficients to be uncorrelated, the $N \times N$-dimensional noisy covariance matrix $\mathbf{\Phi}_{y,l} = \mathcal{E}\{\mathbf{y}_l \mathbf{y}_l^{\mathsf{H}}\}$, with $\mathcal{E}\{\circ\}$ the expectation operator, can be written as

$$\mathbf{\Phi}_{y,l} = \mathbf{\Phi}_{x,l} + \mathbf{\Phi}_{n,l}, \tag{7}$$

where $\mathbf{\Phi}_{x,l} = \mathcal{E}\{\mathbf{x}_l \mathbf{x}_l^{\mathsf{H}}\}$ and $\mathbf{\Phi}_{n,l} = \mathcal{E}\{\mathbf{n}_l \mathbf{n}_l^{\mathsf{H}}\}$ denote the speech and noise covariance matrix, respectively. To exploit the speech correlation across consecutive time frames, it was proposed in [1], [24] to decompose the speech vector $\mathbf{x}_l$ into temporally correlated and uncorrelated components, i.e.,

$$\mathbf{x}_l = \underbrace{\boldsymbol{\gamma}_{x,l} X_l}_{\text{correlated}} + \underbrace{\mathbf{x}'_l}_{\text{uncorrelated}}. \tag{8}$$

The normalized speech inter-frame correlation (IFC) vector $\boldsymbol{\gamma}_{x,l}$ contains the correlation between the speech STFT coefficient X_l at the l-th time frame and the N most recent speech STFT coefficients, i.e.,

$$\boldsymbol{\gamma}_{x,l} = \frac{\mathcal{E}\{\mathbf{x}_l X_l^*\}}{\mathcal{E}\{|X_l|^2\}} = \frac{\mathbf{\Phi}_{x,l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{x,l}\mathbf{e}}, \tag{9}$$

where $\mathbf{e} = [1 \; 0 \; \dots \; 0]^{\mathsf{T}}$ denotes an N-dimensional selection vector and the normalization factor $\mathcal{E}\{|X_l|^2\} = \mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{x,l}\mathbf{e} = \phi_{x,l}$ corresponds to the speech power spectral density. Due to this normalization, the first element of the speech IFC vector is equal to one, i.e.,

$$\boldsymbol{\gamma}_{x,l}^{\mathsf{H}} \mathbf{e} = 1, \tag{10}$$

and the first element of $\mathbf{x}'_l$ in (8) is equal to zero. It is important to note that the correlated speech component $\boldsymbol{\gamma}_{x,l} X_l$ in (8) contains both the desired speech component X_l as well as components that are correlated with X_l. The speech components that are uncorrelated with X_l, i.e., the elements of $\mathbf{x}'_l$, are considered to be undesired. On the one hand, for temporally correlated sounds, e.g., voiced sounds, the correlated component $\boldsymbol{\gamma}_{x,l} X_l$ is typically dominant compared to the uncorrelated component $\mathbf{x}'_l$. On the other hand, for temporally uncorrelated sounds, e.g., unvoiced sounds, the uncorrelated component $\mathbf{x}'_l$ may be quite large.

Substituting (8) into (6), we obtain the multi-frame signal model

$$\mathbf{y}_l = \boldsymbol{\gamma}_{x,l} X_l + \underbrace{\mathbf{x}'_l + \mathbf{n}_l}_{=:\mathbf{i}_l}, \tag{11}$$

where the interference vector $\mathbf{i}_l$ contains both the uncorrelated speech component as well as the noise component. Using (11), the noisy covariance matrix in (7) can be written as

$$\mathbf{\Phi}_{y,l} = \phi_{x,l} \boldsymbol{\gamma}_{x,l} \boldsymbol{\gamma}_{x,l}^{\mathsf{H}} + \underbrace{\mathbf{\Phi}_{x',l} + \mathbf{\Phi}_{n,l}}_{=:\mathbf{\Phi}_{i,l}}, \tag{12}$$

with the interference covariance matrix $\mathbf{\Phi}_{i,l} = \mathcal{E}\{\mathbf{i}_l \mathbf{i}_l^{\mathsf{H}}\}$ and $\mathbf{\Phi}_{x',l} = \mathcal{E}\{\mathbf{x}'_l \mathbf{x}'^{\mathsf{H}}_l\}$. Using (10), it can be shown that

$$\mathbf{\Phi}_{y,l}\mathbf{e} = \phi_{x,l}\boldsymbol{\gamma}_{x,l} + \mathbf{\Phi}_{i,l}\mathbf{e}, \tag{13}$$

such that the speech IFC vector can be written as

$$\boldsymbol{\gamma}_{x,l} = \frac{\mathbf{\Phi}_{y,l}\mathbf{e}}{\phi_{x,l}} - \frac{\mathbf{\Phi}_{i,l}\mathbf{e}}{\phi_{x,l}}. \tag{14}$$

By defining the a-priori signal-to-interference-ratio (SIR) as

$$\xi_l = \frac{\phi_{x,l}}{\phi_{i,l}}, \tag{15}$$

with $\phi_{i,l} = \mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{i,l}\mathbf{e}$ the interference power spectral density, and using

$$\phi_{y,l} = \mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{y,l}\mathbf{e} = \phi_{x,l} + \phi_{i,l}, \tag{16}$$

the speech IFC vector in (14) can be written in terms of the noisy covariance matrix $\mathbf{\Phi}_{y,l}$, the interference covariance matrix $\mathbf{\Phi}_{i,l}$, and the a-priori SIR ξ_l as

$$\boldsymbol{\gamma}_{x,l} = \frac{1+\xi_l}{\xi_l}\frac{\mathbf{\Phi}_{y,l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{y,l}\mathbf{e}} - \frac{1}{\xi_l}\frac{\mathbf{\Phi}_{i,l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{i,l}\mathbf{e}}. \tag{17}$$

Since the first element of $\mathbf{x}'_l$ is equal to zero, it can be easily shown that $\phi_{x',l} = \mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{x',l}\mathbf{e} = 0$, such that the a-priori SIR in (15) is equal to the a-priori SNR, i.e.,

$$\xi_l = \frac{\phi_{x,l}}{\phi_{i,l}} = \frac{\phi_{x,l}}{\phi_{n,l} + \phi_{x',l}} = \frac{\phi_{x,l}}{\phi_{n,l}}. \tag{18}$$

B. Multi-Frame MVDR Filter

In [1], [24] the MFMVDR filter was proposed, which aims at minimizing the output interference power spectral density while not distorting the correlated speech component, i.e.,

$$\mathbf{w}_l = \underset{\mathbf{w} \in \mathbb{C}^N}{\operatorname{argmin}} \quad \mathbf{w}^{\mathsf{H}} \mathbf{\Phi}_{i,l} \mathbf{w}, \quad \text{s.t. } \mathbf{w}^{\mathsf{H}} \boldsymbol{\gamma}_{x,l} = 1. \tag{19}$$

Solving (19) yields the MFMVDR filter vector

$$\mathbf{w}_l = \frac{\mathbf{\Phi}_{i,l}^{-1} \boldsymbol{\gamma}_{x,l}}{\boldsymbol{\gamma}_{x,l}^{\mathsf{H}} \mathbf{\Phi}_{i,l}^{-1} \boldsymbol{\gamma}_{x,l}}. \tag{20}$$

To implement the MFMVDR filter in (20), estimates of the interference covariance matrix $\mathbf{\Phi}_{i,l}$ and the speech IFC vector $\boldsymbol{\gamma}_{x,l}$ are required. Due to the highly time-varying speech correlation, accurately estimating these quantities is not straightforward, and estimation errors may result in reduced noise reduction and speech distortion in the output of the MFMVDR filter. In particular, estimation errors in the speech IFC vector may result in time-varying distortion perceivable as musical noise. Hence, in (20) the interference covariance matrix $\mathbf{\Phi}_{i,l}$ has often been replaced either by the noisy covariance matrix $\mathbf{\Phi}_{y,l}$, leading to the multi-frame minimum power distortionless response filter [24], [26], [27], [37], which is very sensitive to estimation errors in the speech IFC vector [25], or by the noise covariance matrix $\mathbf{\Phi}_{n,l}$ (see Section III) [28], thereby however neglecting the uncorrelated speech component $\mathbf{x}_l'$.

In this article, we will consider the formulation in (17) for the speech IFC vector, such that the quantities to be estimated are the noisy covariance matrix $\mathbf{\Phi}_{y,l}$, the interference covariance matrix $\mathbf{\Phi}_{i,l}$, and the a-priori SIR ξ_l. We will consider different supervised learning-based approaches, which differ both in the training target as well as in the procedure to estimate the covariance matrices. In Section III, we review the approach presented in [28], where the training target is the SPP (used to estimate $\mathbf{\Phi}_{n,l}$ and ξ_l), and hence no end-to-end training using a signal-based loss function is performed. In Section IV, we propose several procedures to estimate $\mathbf{\Phi}_{y,l}$, $\mathbf{\Phi}_{i,l}$, and ξ_l by integrating the fully differentiable MFMVDR filter into a supervised learning framework and using a signal-based loss function for end-to-end training.

C. Covariance Matrix Structures

In this article, we will consider different matrix structures for the $N \times N$-dimensional noisy and interference covariance matrices, which differ in the number of parameters required to determine these matrices. First, by definition, covariance matrices are Hermitian. We assume that the considered covariance matrices are full-rank (rank-N), such that they are positive-definite, i.e., all eigenvalues are real-valued and larger than zero. Hence, the noisy and interference covariance matrices can be decomposed using the Cholesky decomposition [38] as

$$\mathbf{\Phi}_{\nu,l} = \mathbf{L}_{\nu,l} \mathbf{L}_{\nu,l}^{\mathsf{H}}, \quad \nu \in \{y, i\}, \tag{21}$$

where the Cholesky factor $\mathbf{L}_{\nu,l}$ is an $N \times N$-dimensional complex-valued lower-triangular matrix with real and positive diagonal elements, determined by N^2 real-valued parameters.

Assuming the signals to be stationary over N frames, the covariance matrices also exhibit a Toeplitz structure, i.e., the elements on all diagonals are equal. It has been shown in [39] that Hermitian positive-definite Toeplitz (PDT) matrices can be decomposed using their so-called balanced Vandermonde factorization as

$$\mathbf{\Phi}_{\nu,l} = \mathbf{V}_{\nu,l} \mathbf{D}_{\nu,l} \mathbf{V}_{\nu,l}^{\mathsf{H}}, \quad \nu \in \{y, i\}, \tag{22}$$

with $\mathbf{D}_{\nu,l}$ an $N \times N$-dimensional diagonal matrix with real and positive elements and $\mathbf{V}_{\nu,l}$ an $N \times N$-dimensional balanced Vandermonde matrix, defined as

$$\mathbf{V}_{\nu,l} = \begin{bmatrix} 1 & \zeta_{\nu,l,0}^{1} & \zeta_{\nu,l,0}^{2} & \cdots & \zeta_{\nu,l,0}^{N-1} \\ 1 & \zeta_{\nu,l,1}^{1} & \zeta_{\nu,l,1}^{2} & \cdots & \zeta_{\nu,l,1}^{N-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta_{\nu,l,N-1}^{1} & \zeta_{\nu,l,N-1}^{2} & \cdots & \zeta_{\nu,l,N-1}^{N-1} \end{bmatrix}, \tag{23}$$

with $\zeta_{\nu,l,\mu}$ a complex number on the unit circle, i.e., $\zeta_{\nu,l,\mu} = \exp(j\theta_{\nu,l,\mu})$ $\forall \mu \in \{0, \ldots, N-1\}$. Hence, since a balanced Vandermonde matrix can be fully described by the angles $\theta_{\nu,l,\mu}$, the matrices $\mathbf{V}_{\nu,l}$ and $\mathbf{D}_{\nu,l}$ are described by N real-valued parameters each. It should be noted that this assumption presumably holds better for the noise component than for the speech and interference components, which tend to be quite non-stationary.

III. SPP-Based Deep MFMVDR Filter

In this section, we briefly review the SPP-based deep MFMVDR filter approach presented in [28] (depicted in Fig. 1(a)), which will be used as one of the baseline approaches in the simulations. This approach neglects the uncorrelated speech component $\mathbf{x}_l'$ in (8), such that the interference covariance matrix $\mathbf{\Phi}_{i,l}$ in (12) reduces to the noise covariance matrix $\mathbf{\Phi}_{n,l}$. The speech IFC vector $\boldsymbol{\gamma}_{x,l}$ in (17) and the MFMVDR filter vector in (20) are then equal to

$$\boldsymbol{\gamma}_{x,l} = \frac{1+\xi_l}{\xi_l} \frac{\mathbf{\Phi}_{y,l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{y,l}\mathbf{e}} - \frac{1}{\xi_l} \frac{\mathbf{\Phi}_{n,l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\mathbf{\Phi}_{n,l}\mathbf{e}}, \tag{24}$$

$$\mathbf{w}_l = \frac{\mathbf{\Phi}_{n,l}^{-1} \boldsymbol{\gamma}_{x,l}}{\boldsymbol{\gamma}_{x,l}^{\mathsf{H}} \mathbf{\Phi}_{n,l}^{-1} \boldsymbol{\gamma}_{x,l}}. \tag{25}$$

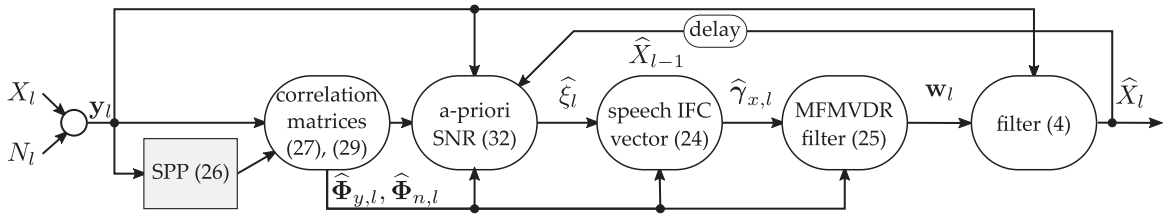
In this approach, a DNN is trained to map the logarithmic magnitude of the noisy STFT coefficients to an estimate of the SPP, i.e.,

$$\widehat{\text{SPP}}_l = \text{DNN}_{\text{SPP}} \left\{ \log_{10} |Y_l| \right\}. \tag{26}$$

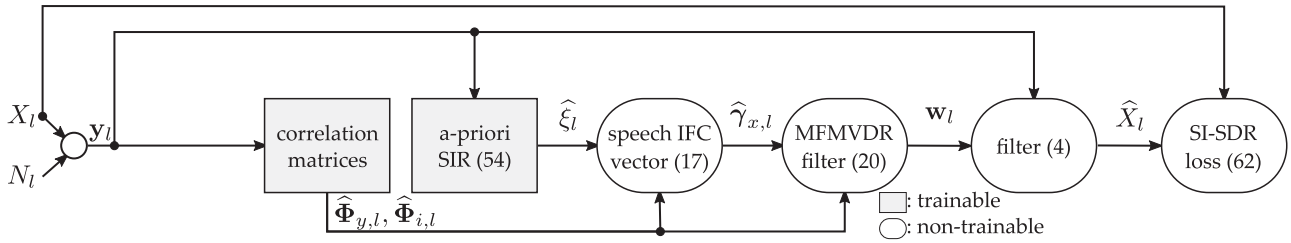
The training target is the SPP defined in [40] using oracle information about the noise component, i.e., no end-to-end training using a signal-based loss function is performed (for more details on the training procedure, we refer to Section V-C). The SPP estimate in (26) is then used to estimate the noise covariance matrix $\mathbf{\Phi}_{n,l}$ based on recursive smoothing using a time-varying SPP-based smoothing factor $\lambda_{n,l}^{\text{SPP}}$ [41], i.e.,

$$\widehat{\mathbf{\Phi}}_{n,l}^{\text{SPP}} = \lambda_{n,l}^{\text{SPP}} \widehat{\mathbf{\Phi}}_{n,l-1}^{\text{SPP}} + (1 - \lambda_{n,l}^{\text{SPP}}) \mathbf{y}_l \mathbf{y}_l^{\mathsf{H}} \tag{27}$$

$$\lambda_{n,l}^{\text{SPP}} = \alpha_n^{\text{SPP}} + (1 - \alpha_n^{\text{SPP}}) \widehat{\text{SPP}}_l, \tag{28}$$



(a) Baseline SPP-based deep MFMVDR filter [28]. The SPP estimator is trained independently from the MFMVDR filter. The speech estimate $\widehat{X}_{l-1}$ from the previous frame is used to estimate the a-priori SNR using the decision-directed approach (DDA).



(b) Proposed signal-based deep MFMVDR filter. The covariance matrix estimators and the a-priori SIR estimator are jointly trained taking into account the MFMVDR filter using the signal-based SI-SDR loss function.

Fig. 1. Block diagrams of the SPP-based and signal-based deep MFMVDR filters. The shaded rectangular boxes represent trainable modules, i.e., TCNs, whereas the rounded boxes represent non-trainable modules.

with α_n^{SPP} a constant. The noisy covariance matrix $\mathbf{\Phi}_{y,l}$ is estimated based on recursive smoothing using a fixed smoothing factor λ_y^{SPP}, i.e.,

$$\widehat{\mathbf{\Phi}}_{y,l}^{\text{SPP}} = \lambda_y^{\text{SPP}} \widehat{\mathbf{\Phi}}_{y,l-1}^{\text{SPP}} + (1-\lambda_y^{\text{SPP}}) \mathbf{y}_l \mathbf{y}_l^{\mathsf{H}}. \tag{29}$$

To ensure invertibility of the estimated noise covariance matrix before using it in (25), diagonal loading is applied, i.e.,

$$\tilde{\widehat{\mathbf{\Phi}}}_{n,l}^{\text{SPP}} = \widehat{\mathbf{\Phi}}_{n,l}^{\text{SPP}} + \rho_{n,l}^{\text{SPP}} \mathbf{I}_N, \tag{30}$$

with $\tilde{\circ}$ denoting regularization, $\mathbf{I}_N$ the $N \times N$-dimensional identity matrix, and $\rho_{n,l}^{\text{SPP}}$ a regularization factor, which is defined as [26], [42]

$$\rho_{n,l}^{\text{SPP}} = \frac{\rho}{N} \text{trace}\left\{ \widehat{\mathbf{\Phi}}_{n,l}^{\text{SPP}} \right\}, \tag{31}$$

with ρ a small constant. The a-priori SNR ξ_l in the l-th time frame is estimated using the decision-directed approach (DDA) [30], i.e.,

$$\widehat{\xi}_l^{\text{SPP}} = \lambda_{\text{DDA}} \frac{\left|\widehat{X}_{l-1}\right|^2}{\widehat{\phi}_{n,l-1}^{\text{SPP}}} + (1-\lambda_{\text{DDA}}) \max\left\{ \frac{|Y_l|^2}{\widehat{\phi}_{n,l}^{\text{SPP}}} - 1, 0 \right\}, \tag{32}$$

with λ_{DDA} a smoothing factor, $\widehat{\phi}_{n,l}^{\text{SPP}} = \mathbf{e}^{\mathsf{T}} \widehat{\mathbf{\Phi}}_{n,l}^{\text{SPP}} \mathbf{e}$ an estimate of the noise power spectral density based on the estimated noise covariance matrix in (27), and $\widehat{X}_{l-1}$ the estimated speech component in the previous frame.

IV. Signal-Based Deep MFMVDR Filter

Contrary to the SPP-based approach described in the previous section, in this section we propose a signal-based deep MFMVDR filter approach (depicted in Fig. 1(b)), where all quantities are estimated with DNNs that are jointly trained using a signal-based loss function at the output of the MFMVDR filter. In other words, the training of the DNNs is guided by the speech estimate obtained at the output of the deep MFMVDR filter, i.e., no ground-truth quantities are required. As already mentioned in Section II, the quantities required to compute the speech IFC vector $\boldsymbol{\gamma}_{x,l}$ in (17) and the MFMVDR filter vector in (20) are the noisy covariance matrix $\mathbf{\Phi}_{y,l}$, the interference covariance matrix $\mathbf{\Phi}_{i,l}$, and the a-priori SIR ξ_l. A separate DNN is used per quantity, with different input features for the DNNs estimating the covariance matrices (Section IV-A) and the a-priori SIR (Section IV-B).

A. Covariance Matrices

In this section we propose different estimation procedures for the noisy and interference covariance matrices $\mathbf{\Phi}_{y,l}$ and $\mathbf{\Phi}_{i,l}$. All estimation procedures have in common that the DNN estimating $\mathbf{\Phi}_{y,l}$ and the DNN estimating $\mathbf{\Phi}_{i,l}$ are jointly trained using a signal-based loss function (see Fig. 1(b)), i.e., without the need for defining target covariance matrices. In the following, we will consider Hermitian positive-definite, Hermitian PDT, and rank-1 matrix structures, where the main difference lies in the number of parameters that need to be estimated as well as in the required linear algebra operations. It should be noted that similarly to (30), diagonal loading is applied to the estimated interference covariance matrix before using it in (20). Since the covariance matrices contain complex-valued elements, we propose to use a concatenation of the logarithmic magnitude as well as the cosine and sine of the phase of the noisy STFT coefficients as input features $\mathbf{f}_l$ for both DNNs, i.e.,

$$\mathbf{f}_l = \begin{bmatrix} \log_{10}(|Y_l| + \epsilon) & \cos \angle Y_l & \sin \angle Y_l \end{bmatrix}^{\mathsf{T}}, \tag{33}$$

where ϵ denotes a small positive constant to avoid numerical issues, and $\angle\circ$ denotes the phase of $\circ$. Note that both the cosine and sine of the phase are used in order to avoid an ambiguous phase encoding [43].

1) Hermitian Positive-Definite: Based on the assumed Hermitian positive-definite structure for covariance matrices, we propose two different estimation procedures. The first procedure is based on recursive smoothing and only requires one parameter to be estimated by each DNN, while the second procedure is based on the Cholesky decomposition and requires N^2 parameters to be estimated by each DNN.

a) Recursive Smoothing (RS): Similarly to (27), the noisy and interference covariance matrices are estimated as

$$\widehat{\mathbf{\Phi}}^{\text{RS}}_{\nu,l} = \lambda^{\text{RS}}_{\nu,l}\widehat{\mathbf{\Phi}}^{\text{RS}}_{\nu,l-1} + (1-\lambda^{\text{RS}}_{\nu,l})\mathbf{y}_l\mathbf{y}_l^{\mathsf{H}}, \tag{34}$$

where the recursive smoothing factors are estimated using DNNs, i.e.,

$$\lambda^{\text{RS}}_{\nu,l} = \text{sigmoid}\{\text{DNN}^{\text{RS}}_{\nu}\{\mathbf{f}_l\}\}, \tag{35}$$

where a sigmoid activation function is used to ensure that the recursive smoothing factors are bounded to $[0,1]$. The proposed recursive smoothing procedure differs from the conventional recursive smoothing procedure (see Section III) by allowing a time-varying smoothing factor for both covariance matrices and by jointly training the DNNs with a signal-based loss function instead of an SPP-based loss function.

b) Cholesky Decomposition (CD): As already mentioned in Section II-C, the $N \times N$-dimensional lower-triangular Cholesky factors $\mathbf{L}_{y,l}$ and $\mathbf{L}_{i,l}$ of the noisy and interference covariance matrices are fully determined by N^2 real-valued parameters each, which can be stacked in the vectors $\boldsymbol{h}^{\text{CD}}_{y,l}, \boldsymbol{h}^{\text{CD}}_{i,l} \in \mathbb{R}^{N^2}$. We propose to estimate these vectors using separate DNNs, i.e.,

$$\widehat{\boldsymbol{h}}^{\text{CD}}_{y,l} = \text{DNN}^{\text{CD}}_{y}\{\mathbf{f}_l\}, \quad \widehat{\boldsymbol{h}}^{\text{CD}}_{i,l} = \text{DNN}^{\text{CD}}_{i}\{\mathbf{f}_l\}. \tag{36}$$

The estimated Cholesky factors $\widehat{\mathbf{L}}_{y,l}$ and $\widehat{\mathbf{L}}_{i,l}$ are assembled from $\widehat{\boldsymbol{h}}^{\text{CD}}_{y,l}$ and $\widehat{\boldsymbol{h}}^{\text{CD}}_{i,l}$ as described in Algorithm 1, where separate subsets of the real-valued vector elements are used for the real strictly lower triangular part, the imaginary strictly lower triangular part, and the real positive diagonal part. Positivity of the diagonal part is ensured by using a softplus activation function. Finally, estimates of the covariance matrices are obtained as $\widehat{\mathbf{\Phi}}^{\text{CD}}_{y,l} = \widehat{\mathbf{L}}_{y,l}\widehat{\mathbf{L}}^{\mathsf{H}}_{y,l}$ and $\widehat{\mathbf{\Phi}}^{\text{CD}}_{i,l} = \widehat{\mathbf{L}}_{i,l}\widehat{\mathbf{L}}^{\mathsf{H}}_{i,l}$.

2) Hermitian Positive-Definite Toeplitz (PDT): When assuming stationarity of the noisy and interference components over N frames, the respective covariance matrices $\mathbf{\Phi}_{y,l}$ and $\mathbf{\Phi}_{i,l}$ exhibit a Hermitian PDT structure. As mentioned in Section II-C, PDT matrices can be decomposed as $\mathbf{\Phi}_{\nu,l} = \mathbf{V}_{\nu,l}\mathbf{D}_{\nu,l}\mathbf{V}^{\mathsf{H}}_{\nu,l}$, where the balanced Vandermonde matrix $\mathbf{V}_{\nu,l}$ in (23) is fully determined by N angles $\theta_{\nu,l,\mu}$, and the (positive-definite) diagonal matrix $\mathbf{D}_{\nu,l}$ is fully determined by N real-valued parameters. These angles and parameters can be stacked in the vectors $\boldsymbol{h}^{\text{PDT}}_{y,l}, \boldsymbol{h}^{\text{PDT}}_{i,l} \in \mathbb{R}^{2N}$. We propose to estimate these vectors using separate DNNs, i.e.,

$$\widehat{\boldsymbol{h}}^{\text{PDT}}_{y,l} = \text{DNN}^{\text{PDT}}_{y}\{\mathbf{f}_l\}, \quad \widehat{\boldsymbol{h}}^{\text{PDT}}_{i,l} = \text{DNN}^{\text{PDT}}_{i}\{\mathbf{f}_l\}. \tag{37}$$

Algorithm 1: Construct a Hermitian Positive-Definite Matrix $\widehat{\mathbf{\Phi}}$ From a Vector of Real-Valued Parameters $\widehat{\boldsymbol{h}}$ Using its Cholesky Decomposition. stril$\{\circ\}$ Assembles a Strictly Lower Triangular Matrix from $\circ$.

procedure CHOLESKY($\widehat{\boldsymbol{h}} \in \mathbb{R}^{N^2}$)
  construct strictly lower triangular matrices:
  $\widehat{\mathbf{A}}^{\Re} = \text{stril}\{[\widehat{\boldsymbol{h}}]_{0:\frac{1}{2}(N^2-N)-1}\}$
  $\widehat{\mathbf{A}}^{\Im} = \text{stril}\{[\widehat{\boldsymbol{h}}]_{\frac{1}{2}(N^2-N):N^2-N-1}\}$
  construct real-valued positive diagonal matrix:
  $\widehat{\mathbf{B}} = \text{diag}\{\text{softplus}\{[\widehat{\boldsymbol{h}}]_{N^2-N:N^2-1}\}\}$
  construct Cholesky factor:
  $\widehat{\mathbf{L}} = \widehat{\mathbf{A}}^{\Re} + j\widehat{\mathbf{A}}^{\Im} + \widehat{\mathbf{B}}, \quad j^2 = -1$
  return $\widehat{\mathbf{\Phi}} = \widehat{\mathbf{L}}\widehat{\mathbf{L}}^{\mathsf{H}}$

Algorithm 2: Construct a Hermitian PDT Matrix $\widehat{\mathbf{\Phi}}$ From a Vector of Real-Valued Parameters $\widehat{\mathbf{h}}$ Using its Vandermonde Factorization.

procedure VANDERMONDE($\widehat{\mathbf{h}} \in \mathbb{R}^{2N}$)
  $\widehat{\boldsymbol{\zeta}} = \begin{bmatrix}\widehat{\zeta}_0 & \dots & \widehat{\zeta}_{N-1}\end{bmatrix}^{\mathsf{T}} = \exp(j\pi\tanh\{[\widehat{\mathbf{h}}]_{0:N-1}\})$
  using $\widehat{\boldsymbol{\zeta}}$, assemble $\widehat{\mathbf{V}}$ as in (23)
  $\widehat{\mathbf{D}} = \text{diag}\{\text{softplus}\{[\widehat{\mathbf{h}}]_{N:2N-1}\}\}$
  return $\widehat{\mathbf{\Phi}} = \widehat{\mathbf{V}}\widehat{\mathbf{D}}\widehat{\mathbf{V}}^{\mathsf{H}}$

The estimated Hermitian PDT matrices $\widehat{\mathbf{\Phi}}^{\text{PDT}}_{y,l}$ and $\widehat{\mathbf{\Phi}}^{\text{PDT}}_{i,l}$ are assembled from $\widehat{\boldsymbol{h}}^{\text{PDT}}_{y,l}$ and $\widehat{\boldsymbol{h}}^{\text{PDT}}_{i,l}$ as described in Algorithm 2. The angles are computed from the first N elements of $\widehat{\boldsymbol{h}}^{\text{PDT}}_{\nu,l}$, bounded to $[-\pi,\pi]$ using a scaled hyperbolic tangent activation function, i.e., $\widehat{\theta}_{\nu,l,\mu} = \pi\tanh\{[\widehat{\boldsymbol{h}}^{\text{PDT}}_{\nu,l}]_{\mu}\}$. The diagonal matrix $\widehat{\mathbf{D}}_{\nu,l}$ is computed from the second N elements of $\widehat{\boldsymbol{h}}^{\text{PDT}}_{\nu,l}$, where positivity of the diagonal elements is ensured by using a softplus activation function. Finally, estimates of the covariance matrices are obtained as $\widehat{\mathbf{\Phi}}^{\text{PDT}}_{y,l} = \widehat{\mathbf{V}}_{y,l}\widehat{\mathbf{D}}_{y,l}\widehat{\mathbf{V}}^{\mathsf{H}}_{y,l}$ and $\widehat{\mathbf{\Phi}}^{\text{PDT}}_{i,l} = \widehat{\mathbf{V}}_{i,l}\widehat{\mathbf{D}}_{i,l}\widehat{\mathbf{V}}^{\mathsf{H}}_{i,l}$.

3) Rank-1 (R1): Although the noisy and interference covariance matrices are full-rank, here we assume that these matrices can be approximated using a rank-1 structure,[1] i.e.,

$$\mathbf{\Phi}^{\text{R1}}_{y,l} = \boldsymbol{h}^{\text{R1},\mathbb{C}}_{y,l}\left(\boldsymbol{h}^{\text{R1},\mathbb{C}}_{y,l}\right)^{\mathsf{H}}, \tag{38}$$

$$\mathbf{\Phi}^{\text{R1}}_{i,l} = \boldsymbol{h}^{\text{R1},\mathbb{C}}_{i,l}\left(\boldsymbol{h}^{\text{R1},\mathbb{C}}_{i,l}\right)^{\mathsf{H}}, \tag{39}$$

where $\boldsymbol{h}^{\text{R1},\mathbb{C}}_{y,l}$ and $\boldsymbol{h}^{\text{R1},\mathbb{C}}_{i,l}$ denote N-dimensional complex-valued vectors fully determined by $2N$ real-valued parameters each, which can be stacked in the vectors $\boldsymbol{h}^{\text{R1}}_{y,l}, \boldsymbol{h}^{\text{R1}}_{i,l} \in \mathbb{R}^{2N}$. We propose to estimate these vectors using separate DNNs, i.e.,

$$\widehat{\boldsymbol{h}}^{\text{R1}}_{y,l} = \text{DNN}^{\text{R1}}_{y}\{\mathbf{f}_l\}, \quad \widehat{\boldsymbol{h}}^{\text{R1}}_{i,l} = \text{DNN}^{\text{PDT}}_{i}\{\mathbf{f}_l\}. \tag{40}$$

The vectors $\widehat{\boldsymbol{h}}^{\text{R1},\mathbb{C}}_{y,l}$ and $\widehat{\boldsymbol{h}}^{\text{R1},\mathbb{C}}_{i,l}$ can then be obtained as

$$\widehat{\boldsymbol{h}}^{\text{R1},\mathbb{C}}_{\nu,l} = [\widehat{\boldsymbol{h}}^{\text{R1}}_{\nu,l}]_{0:N-1} + j\,[\widehat{\boldsymbol{h}}^{\text{R1}}_{\nu,l}]_{N:2N-1}. \tag{41}$$

[1] Although we realize that this approximation lacks a theoretical motivation, especially for the noisy covariance matrix, it will result in a lower-complexity estimation procedure with a very good performance (see Section VI).

Since the rank-1 matrix $\widehat{\mathbf{\Phi}}_{i,l}^{\mathrm{R1}} = \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}})^{\mathsf{H}}$ is not invertible and hence can not be directly used in (20), diagonal loading is applied, i.e.,

$$\tilde{\widehat{\mathbf{\Phi}}}_{i,l}^{\mathrm{R1}} = \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}} \left(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right)^{\mathsf{H}} + \rho_{i,l}^{\mathrm{R1}} \mathbf{I}_N, \tag{42}$$

where the regularization factor $\rho_{i,l}^{\mathrm{R1}}$ is defined as in (31), i.e.,

$$\rho_{i,l}^{\mathrm{R1}} = \frac{\rho}{N} \mathrm{trace}\left\{\widehat{\mathbf{\Phi}}_{i,l}^{R1}\right\} = \frac{\rho}{N} \left\|\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right\|_2^2, \tag{43}$$

with ρ a small constant and $\|\circ\|_2$ denoting the ℓ^2-norm of $\circ$. Using the (regularized) rank-1 covariance matrices in (38) and (42), it can be shown that the MFMVDR filter in (20) can be directly computed as a linear combination of the DNN outputs $\widehat{\boldsymbol{h}}_{y,l}^{\mathrm{R1},\mathbb{C}}$ and $\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}$ without requiring computationally complex matrix inversions. First, using (38) and (42) in (17), the speech IFC vector can be written as a linear combination of the DNN outputs, i.e.,

$$\widehat{\gamma}_{x,l}^{\mathrm{R1}} = \alpha_{y,l} \widehat{\boldsymbol{h}}_{y,l}^{\mathrm{R1},\mathbb{C}} + \alpha_{i,l} \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}} + \alpha_{e,l} \mathbf{e}, \tag{44}$$

with

$$\alpha_{y,l} = \frac{1+\widehat{\xi}_l}{\widehat{\xi}_l} \frac{1}{[\widehat{\boldsymbol{h}}_{y,l}^{\mathrm{R1},\mathbb{C}}]_0} \tag{45}$$

$$\alpha_{i,l} = -\frac{1}{\widehat{\xi}_l} \frac{[\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}]_0^*}{|[\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}]_0|^2 + \rho_{i,l}^{\mathrm{R1}}} \tag{46}$$

$$\alpha_{e,l} = \alpha_{i,l} \frac{\rho_{i,l}^{\mathrm{R1}}}{[\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}]_0^*}. \tag{47}$$

By using the matrix inversion lemma, it can be easily shown that the inverse of the interference covariance matrix in (42) is equal to

$$\left(\tilde{\widehat{\mathbf{\Phi}}}_{i,l}^{\mathrm{R1}}\right)^{-1} = \frac{1}{\rho_{i,l}^{\mathrm{R1}}} \left(\mathbf{I}_N - \eta_l \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}} \left(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right)^{\mathsf{H}}\right), \tag{48}$$

with

$$\eta_l = \frac{1}{\rho_{i,l}^{\mathrm{R1}} + \left\|\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right\|_2^2} \in \mathbb{R}. \tag{49}$$

By plugging (48) into (20), the MFMVDR filter vector can be written as

$$\mathbf{w}_l^{\mathrm{R1}} = \frac{\left(\mathbf{I}_N - \eta_l \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}} \left(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right)^{\mathsf{H}}\right)}{\left(\widehat{\boldsymbol{\gamma}}_{x,l}^{\mathrm{R1}}\right)^{\mathsf{H}} \left(\mathbf{I}_N - \eta_l \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}} \left(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right)^{\mathsf{H}}\right) \widehat{\boldsymbol{\gamma}}_{x,l}^{\mathrm{R1}}} \widehat{\boldsymbol{\gamma}}_{x,l}^{\mathrm{R1}} \tag{50}$$

$$= \frac{1}{\kappa_l} \left(\mathbf{I}_N - \eta_l \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}} \left(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right)^{\mathsf{H}}\right) \widehat{\boldsymbol{\gamma}}_{x,l}^{\mathrm{R1}}, \tag{51}$$

with

$$\kappa_l = \left\|\widehat{\boldsymbol{\gamma}}_{x,l}^{\mathrm{R1}}\right\|_2^2 - \eta_l \left|\left(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right)^{\mathsf{H}} \widehat{\boldsymbol{\gamma}}_{x,l}^{\mathrm{R1}}\right|^2 \in \mathbb{R}. \tag{52}$$

Finally, by plugging (44) into (51), the MFMVDR filter vector can be written as a linear combination of the DNN outputs, i.e.,

$$\mathbf{w}_l^{\mathrm{R1}} = \frac{1}{\kappa_l} \left[\alpha_{y,l} \widehat{\boldsymbol{h}}_{y,l}^{\mathrm{R1},\mathbb{C}} + \left(\alpha_{i,l} - \eta_l \left(\widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}}\right)^{\mathsf{H}} \widehat{\boldsymbol{\gamma}}_{x,l}^{\mathrm{R1}}\right) \widehat{\boldsymbol{h}}_{i,l}^{\mathrm{R1},\mathbb{C}} + \alpha_{e,l} \mathbf{e}\right]. \tag{53}$$

B. A-Priori SIR

To estimate the a-priori SIR ξ_l, a similar approach is used as for the covariance matrices. Instead of training a DNN to map input features to an a-priori SIR target, a DNN is trained by minimizing a signal-based loss function at the output of the MFMVDR filter (see Fig. 1(b)), where it should be noted that the DNN estimating the a-priori SIR is trained jointly with the DNNs estimating the covariance matrices (see Section IV-A). Since the a-priori SIR is a quantity relating signal powers, magnitude information is assumed to be sufficient for its estimation. Hence, similarly as in (26), we propose to use the logarithmic magnitude of the noisy STFT coefficients as input feature, i.e.,

$$\widehat{\xi}_l = \mathrm{softplus}\{\mathrm{DNN}_\xi\{\log_{10}|Y_l|\}\}, \tag{54}$$

where positivity of $\widehat{\xi}_l$ is ensured by using a softplus activation function.

V. Simulation Setup

In this section, we present our simulation setup, more in particular the used datasets (Section V-A), the baseline speech enhancement algorithms (Section V-B), and the settings of the considered algorithms (Section V-C).

A. Datasets

The data used to train, validate, and test the DNNs are based on the first DNS challenge [44].

1) Training & Validation: The training and validation datasets were generated using the official dataset generator of the DNS challenge [44], i.e., 500 h of clean speech from 2150 speakers from the LibriSpeech dataset [45], 60000 noise clips from the Audioset dataset [46], and 10000 noise clips from the Freesound and DEMAND datasets [47], [48]. Noisy utterances were generated by randomly choosing clean speech and noise and mixing them at broadband SNRs ranging from 0 dB to 19 dB with a length of 4 s per utterance.

2) Testing: To test the considered speech enhancement algorithms, we used the official DNS challenge test dataset, comprising English clean speech from the Graz University dataset [49] and noise from the Audioset and Freesound datasets, totaling 150 utterances at SNRs ranging from 0 dB to 19 dB. All test utterances have a length of 10 s, and the training, validation, and test datasets are disjoint.

B. Baseline Algorithms

In addition to comparing the different proposed estimation procedures for the signal-based deep MFMVDR filter (Section IV), we compare the signal-based deep MFMVDR filter to several baseline algorithms:

1) SPP-based deep MFMVDR filter (Section III), aiming at investigating the difference between using the SPP as a training target and using a signal-based loss function at the output of the MFMVDR filter.
2) *Masking:* Aiming at investigating the benefit of multi-frame filtering, i.e., $N > 1$, we also consider single-frame masking algorithms as described in (2), i.e., $\widehat{X}_l = M_l Y_l$,

either using a real-valued mask $M_l \leftarrow M_l^{\mathbb{R}}$ or a complex-valued mask $M_l \leftarrow M_l^{\mathbb{C}}$. The real-valued mask is estimated using a DNN as

$$M_l^{\mathbb{R}} = \text{sigmoid}\{\text{DNN}^{\mathbb{R}}\{\mathbf{f}_l\}\}, \tag{55}$$

where a sigmoid activation function is used to ensure that the mask is bounded to $[0, 1]$. The complex-valued mask is estimated using a DNN as

$$\begin{bmatrix} \Re\{M_l^{\mathbb{C}}\} \\ \Im\{M_l^{\mathbb{C}}\} \end{bmatrix} = \tanh\{\text{DNN}^{\mathbb{C}}\{\mathbf{f}_l\}\}, \tag{56}$$

where $\Re\{\circ\}$ and $\Im\{\circ\}$ denote the real and imaginary part of $\circ$, respectively, and a hyperbolic tangent activation function is used to ensure that both parts are bounded to $[-1, 1]$.

3) *Direct multi-frame filtering (DMFF):* Aiming at investigating the benefit of imposing the MFMVDR filter structure in (20), we also consider directly estimating the complex-valued elements of the N-dimensional multi-frame filter in (5) using a DNN, similarly to [21]. The real and imaginary parts of the multi-frame filter $\mathbf{w}_l^{\text{DMFF}}$ are estimated as

$$\begin{bmatrix} \Re\{\mathbf{w}_l^{\text{DMFF}}\} \\ \Im\{\mathbf{w}_l^{\text{DMFF}}\} \end{bmatrix} = \tanh\{\text{DNN}^{\text{DMFF}}\{\mathbf{f}_l\}\}, \tag{57}$$

where a hyperbolic tangent activation function is used to ensure that the real and imaginary parts of all filter coefficients are bounded to $[-1, 1]$ as in [21].

C. Algorithmic Settings

For all considered algorithms, the same STFT framework was used. To increase speech correlation across consecutive STFT frames, a high temporal resolution was utilized, i.e., a frame length of 8 ms and a frame overlap of 75% , similarly as in [24], [29]. A square root Hann window was used both as analysis and synthesis window. All algorithms using multi-frame filters, i.e., all deep MFMVDR filters and the DMFF algorithm, used a filter length of $N = 5$, enabling the algorithms to exploit temporal correlations within 16 ms. Similarly as in [28], for the SPP-based deep MFMVDR filter we used a smoothing constant $\alpha_n^{\text{SPP}} = 0.9694$ (corresponding to about 50 ms) for the noise covariance matrix in (28), a smoothing factor $\lambda_y^{\text{SPP}} = 0.8464$ (corresponding to about 12 ms) for the noisy covariance matrix in (29), and a smoothing factor $\lambda_{\text{DDA}} = 0.9408$ (corresponding to about 33 ms) for the decision-directed approach (DDA) in (32). The target SPP was defined as [40]

$$\text{SPP}_l = \left(1 + (1 + \xi_{\mathcal{H}_1}) e^{-\frac{|Y_l|^2}{\widehat{\phi}_{n,l}} \frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)^{-1} \tag{58}$$

$$\widehat{\phi}_{n,l} = \lambda_{n,l}^{\text{SPP}} \widehat{\phi}_{n,l-1} + (1 - \lambda_{n,l}^{\text{SPP}})|N_l|^2, \tag{59}$$

with $\lambda_{n,l}^{\text{SPP}}$ defined in (28). For the deep MFMVDR filter, diagonal loading with a fixed regularization constant $\rho = 10^{-3}$ was applied to all estimated interference or noise covariance matrices. In (33), we used $\epsilon = 1 \times 10^{-8}$ to avoid numerical issues. No recursive smoothing was applied in the case of the Cholesky decomposition (CD), PDT, or rank-1 (R1) deep MFMVDR filters.

To limit speech distortion, a smooth minimum gain was applied to the output $\widehat{X}_l$ of all considered algorithms, i.e., the final speech estimate $\widehat{X}_l^{\text{fin}}$ used in the evaluation was obtained as

$$\widehat{X}_l^{\text{fin}} = \beta_l \widehat{X}_l + (1 - \beta_l) G_{\min} Y_l, \tag{60}$$

where the factor β_l interpolates between the estimated speech component $\widehat{X}_l$ and a hard minimum gain $G_{\min}$ applied to the noisy STFT coefficients. We propose to use a smooth (time-varying) interpolation factor

$$\beta_l = \frac{1}{1 + e^{-2s(|\widehat{X}_l| - |G_{\min} Y_l|)}}, \tag{61}$$

where s controls how accurately the hard minimum gain $G_{\min}$ is approximated. The advantage of using the minimum gain approximation in (60) over a hard minimum gain (corresponding to $s = \infty$) is the fact that the former is smoothly differentiable, allowing for it to be used during back-propagation. For all algorithms, we used $G_{\min} = -17$ dB and $s = 10$.

The DNNs were trained using the AdamW optimizer [50] with an initial learning rate of 3×10^{-4}. The learning rate was halved after the validation loss did not decrease for 3 consecutive epochs, and training was stopped if either 50 epochs were completed or the validation loss did not decrease for 10 consecutive epochs. The gradient ℓ^2-norms were clipped to 5, and a batch size of 4 was used. As loss function for the SPP-based deep MFMVDR filter we used the mean square error between the estimated SPP in (26) and the target SPP defined in (58). For all other algorithms, we used the signal-based SI-SDR as loss function, i.e.,

$$\text{SI-SDR} = 10 \log_{10} \left(\frac{|\alpha \bar{\mathbf{x}}|^2}{|\alpha \bar{\mathbf{x}} - \widehat{\bar{\mathbf{x}}}|^2} \right), \quad \alpha = \frac{\widehat{\bar{\mathbf{x}}}^{\mathsf{T}} \bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|_2^2}, \tag{62}$$

where $\bar{\mathbf{x}}$ and $\widehat{\bar{\mathbf{x}}}$ denote the clean speech signal and the estimated speech signal after inverse STFT processing in the time domain. All algorithms were implemented in PyTorch 1.10.1, and training and evaluation were performed on NVIDIA GeForce RTX 2080 Ti graphics cards. As the DNN architecture for all estimators, we used temporal convolutional network (TCN), which exhibit strong temporal and spectral modeling capabilities [12] and have been shown to be quite effective in the context of speech enhancement [9]. We used the official Conv-TasNet TCN implementation,[2] excluding the encoder and decoder modules, since the proposed algorithms operate in the STFT domain. We fixed the hyperparameters of all TCN modules to 2 stacks of 4 layers each, with a kernel size of 3, resulting in a temporal receptive field size of 61 frames (corresponding to 128 ms). Aiming at a fair comparison, the number of channels in the convolutional block and the number of channels in the bottleneck were varied to obtain a similar total number of trainable weights for all considered algorithms (see Table II). The ratio between these numbers was kept fixed at 4 as proposed in [9]. The

[2][Online]. Available: https://github.com/naplab/Conv-TasNet

TABLE I
SPEECH ENHANCEMENT PERFORMANCE ON THE DNS TEST DATASET, PRESENTED IN TERMS OF SI-SDR, NARROWBAND PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ) (PESQ-NB), WIDEBAND PESQ (PESQ-WB), SHORT-TIME OBJECTIVE INTELLIGIBILITY (STOI), AND DNSMOS FOR DEEP MFMVDR FILTERS USING DIFFERENT MATRIX ESTIMATION PROCEDURES BASED ON SPEECH PRESENCE PROBABILITY (SPP), RECURSIVE SMOOTHING (RS), CHOLESKY DECOMPOSITION (CD), POSITIVE-DEFINITE TOEPLITZ (PDT), AND RANK-1 (R1), THE REAL- AND COMPLEX-VALUED MASKING ALGORITHMS, THE DIRECT MULTI-FRAME FILTERING (DMFF) ALGORITHM, AS WELL AS THE DCCRN-MC AND DCUNET-MC ALGORITHMS

algorithm	SI-SDR / dB	PESQ-NB	PESQ-WB	STOI	DNSMOS
noisy	9.23	2.45	1.58	0.915	3.15
deep MFMVDR (SPP)	10.87	2.57	1.71	0.907	3.24
deep MFMVDR (RS)	15.03	2.97	2.24	0.942	3.45
deep MFMVDR (CD)	**17.60**	3.28	**2.71**	**0.961**	**3.75**
deep MFMVDR (PDT)	15.20	2.98	2.29	0.945	3.37
deep MFMVDR (R1)	17.31	3.25	2.67	0.959	3.74
masking (real)	16.24	3.11	2.44	0.949	3.57
masking (complex)	17.25	3.20	2.59	0.956	3.74
DMFF	17.37	3.20	2.61	0.955	3.73
DCCRN-MC [51]	16.50	3.21	—	0.951	—
DCUNET-MC [51]	17.46	**3.30**	—	**0.961**	—

motivation behind varying these numbers as opposed to the number of stacks/layers or the kernel size is to keep the temporal receptive field size fixed, which might otherwise result in an unfair comparison. Frequency bins were treated as channels in the context of the TCN architecture, resulting in the following input-to-output mapping: $\mathbb{R}^{B\times 3\,\mathrm{K}\times \mathrm{L}} \rightarrow \mathbb{R}^{B\times (n_{\text{param}}\,\mathrm{K})\times \mathrm{L}}$, with B the batch size, K the number of frequency bins, L the number of time frames, and n_{param} the number of parameters per frequency bin required by the specific covariance matrix estimation procedure.

VI. SIMULATION RESULTS

In this section, we investigate the speech enhancement performance and the computational complexity of the signal-based deep MFMVDR filters using the different proposed matrix estimation procedures proposed in Section IV. In Section VI-A, the SPP-based deep MFMVDR filter and the signal-based deep MFMVDR filters using the different proposed matrix estimation procedures are compared. In Section VI-B, the best-performing deep MFMVDR filters are compared with several baseline algorithms.

A. Comparison of Deep MFMVDR Filters

Table I shows the speech enhancement performance on the DNS test dataset for the SPP-based deep MFMVDR filter and the signal-based deep MFMVDR filters using the different matrix estimation procedures (RS, CD, PDT, R1). In this section, we focus on all considered deep MFMVDR filters. The speech enhancement performance is presented in terms of SI-SDR [52], the narrowband and wideband PESQ metrics [53], the STOI metric [54], as well as the Deep Noise Suppression Mean Opinion Score (DNSMOS) metric [55], using the clean speech signal as the reference signal. As can be observed, all considered signal-based MFMVDR filters yield a significant improvement in terms of all performance metrics. Comparing the SPP-based deep MFMVDR filter and the signal-based deep MFMVDR filter using the recursive smoothing (RS) matrix estimation procedure, which both utilize a variant of recursive smoothing, a clear benefit of defining the loss function on the signal level can be observed. A possible explanation is the fact that the SPP-based deep MFMVDR filter requires a number of smoothing parameters to be chosen by hand, i.e., α_n^{SPP} for the noise covariance matrix in (28) (also affecting the target used in the loss function), λ_y^{SPP} for the noisy covariance matrix in (29), and λ_{DDA} for the DDA in (32). The combination of these smoothing parameters critically affects the resulting speech enhancement performance, and different parameter choices may be more suitable for different acoustic conditions. In contrast, in the signal-based deep MFMVDR filter approach the required smoothing parameters, i.e., $\lambda_{y,l}^{\text{RS}}$ for the noisy covariance matrix and $\lambda_{i,l}^{\text{RS}}$ for the interference covariance matrix in (34) are determined by the DNNs.

Comparing the signal-based deep MFMVDR filters, it can be observed that the CD matrix estimation procedure consistently yields the highest improvement in terms of all performance metrics, closely followed by the R1 matrix estimation procedure. The RS and PDT matrix estimation procedures yield the lowest improvements.

As discussed in Section II-C, the CD matrix estimation procedure merely imposes a Hermitian positive-definite structure on the estimated covariance matrices, i.e., it actually does not restrict the estimated covariance matrices more than mathematically required. While the RS, PDT, and R1 matrix estimation procedures also yield Hermitian positive-definite covariance matrices, they impose further structure. More specifically, from (34) it can be seen that the RS matrix estimation procedure imposes rank-1 updates using the noisy vector. By imposing a Toeplitz structure, the PDT matrix estimation procedure imposes stationarity over N frames on the noisy and interference components, apparently leading to a significant reduction in speech enhancement performance compared with the CD matrix estimation procedure. Hence, a Toeplitz structure does not seem to be a viable choice for modeling the noisy and interference covariance matrices. Interestingly, the R1 matrix estimation procedure, which imposes a dominant principal subspace on the noisy and interference covariance matrices, yields quite a high speech enhancement performance.

For the deep MFMVDR filters, Table II shows the network size in terms of trainable weights, bottleneck dimension size

TABLE II
NETWORK SIZE, PRESENTED IN TERMS OF TRAINABLE WEIGHTS, BOTTLENECK DIMENSION SIZE AND MEMORY, AND COMPUTATIONAL COMPLEXITY, PRESENTED IN TERMS OF THE REAL-TIME FACTOR (RTF), THE CONTRIBUTION OF THE MFMVDR LINEAR ALGEBRA OPERATIONS TO THE RTF, THE NUMBER OF FLOPS, AND THE NUMBER OF ESTIMATED PARAMETERS

algorithm	trainable weights / M	bottleneck dimension	memory / MB	RTF	RTF contribution, MFMVDR / %	FLOPS / M	estimated parameters
deep MFMVDR (SPP)	5.3	231	195.6	0.176	54.9	22.9	$K = 65$
deep MFMVDR (RS)	4.9	128	137.3	0.167	47.9	26.9	$2K = 130$
deep MFMVDR (CD)	5.3	128	110.9	0.139	39.0	46.2	$2N^2K = 3250$
deep MFMVDR (PDT)	5.1	128	93.3	0.170	43.4	65.4	$4NK = 1300$
deep MFMVDR (R1)	5.1	128	85.2	0.100	7.5	32.5	$4NK = 1300$
masking (real)	5.0	226	30.2	0.075	0.0	16.3	$K = 65$
masking (complex)	5.0	226	28.5	0.077	0.0	16.4	$2K = 130$
DMFF	5.2	226	29.5	0.079	0.0	17.1	$2NK = 650$
DCCRN-MC [51]	20.0	—					
DCUNET-MC [51]	3.5	—					

These metrics were computed using the PyTorch profiler, averaged across the DNS test dataset.

and memory, and the computational complexity in terms of the real-time factor (RTF), defined as the ratio between processing duration and signal duration, as well as the contribution of the MFMVDR linear algebra operations to the RTF, the number of floating point operations per second (FLOPS), and the number of estimated parameters. As linear algebra operations any operation is counted between obtaining the DNN outputs and applying the multi-frame filter to the noisy vector. All metrics were computed using the PyTorch profiler for the DNS test dataset, i.e., 100 signals of length 10 s, using a single core of an AMD EPYC 7443P CPU clocked at 3.8 GHz. First, it can be observed that all deep MFMVDR filters exhibit an RTF that is significantly smaller than 1. Second, the signal-based deep MFMVDR filter using the R1 estimation procedure exhibits a significantly smaller RTF than the other deep MFMVDR filters. This can be explained by the fact that the SPP-based, RS, CD and PDT estimation procedures require relatively complex operations to construct the covariance matrix estimates from the DNN outputs (see, e.g., Algorithms 1 and 2) and require a matrix inversion to compute the MFMVDR filter in (20). Hence, as can be observed in Table II, the RTF is primarily determined by the MFMVDR linear algebra operations and not by the number of parameters that need to be estimated by the TCNs.

Relating the speech enhancement performance in Table I and the computational complexity in Table II, a benefit of the proposed R1 matrix estimation procedure can be identified, since it yields a speech enhancement performance that is only slightly lower than the CD matrix estimation procedure, while reducing the RTF, the number of FLOPS, and the memory consumption by approximately 30%.

B. Comparison With Baseline Algorithms

In this section, we focus on the best-performing deep MFMVDR filters, i.e., the signal-based deep MFMVDR filter using the CD and R1 matrix estimation procedures, and compare their speech enhancement performance (Table I) and computational complexity (Table II) with several baseline algorithms. As the first set of baseline algorithms, we consider the real- and complex-valued masking as well as the DMFF algorithms discussed in Section V-B, which are based on the same TCN architecture and trained using the same procedure as the proposed deep MFMVDR filters, thus allowing to investigate the effect of imposing the deep MFMVDR structure. As additional state-of-the-art baseline algorithms, we consider the deep complex convolutional recurrent network (DCCRN-MC) and the deep complex U-Net (DCUNET-MC) algorithms proposed in [51]. These algorithms integrate a complex convolutional recurrent-based or a complex U-Net-based encoder-decoder structure with complex convolutional block attention modules to estimate single-frame complex-valued masks, which are applied to the noisy STFT coefficients. The DCCRN-MC and DCUNET-MC were trained using a mixed loss function, comprising an SI-SDR term as well as a term consisting of the squared complex-valued mask error.

Note that the DCCRN-MC and DCUNET-MC algorithms were not retrained in the context of these simulations, and instead the official published results on the DNS challenge test dataset are presented, which is the reason for missing values in Tables I and II.

First, comparing the speech enhancement performance of the real- and complex-valued masking algorithms, it can be confirmed that adding the potential of phase enhancement yields an increased performance in terms of all considered metrics. Second, comparing the complex-valued masking and DMFF algorithms, it can be observed that the speech enhancement performance is hardly improved by increasing the number of filter coefficients from $N = 1$ to $N = 5$ (only improvement in terms of SI-SDR and wideband PESQ). Third, comparing the complex-valued masking algorithms employing either the TCN architecture, the DCCRN architecture, or the DCUNET architecture, it can be observed that the TCN-based complex-valued masking algorithm performs slightly better than the DCCRN-MC algorithm, while both algorithms are consistently outperformed by the DCUNET-MC algorithm. Fourth, we compare the best-performing proposed deep MFMVDR filters, i.e., using the R1 and the CD estimation procedure, with the best-performing baseline algorithms, i.e., the DMFF algorithm and the DCUNET-MC algorithm. The proposed R1 deep MFMVDR filter outperforms the DMFF algorithm (except in terms of SI-SDR) while being outperformed by the DCUNET-MC algorithm. In contrast, the proposed CD deep MFMVDR filter consistently outperforms the DMFF algorithm and outperforms

the DCUNET-MC algorithm in terms of SI-SDR while yielding a similar performance in terms of narrowband PESQ and STOI. Exemplary audio examples are available online.[3]

In terms of computational complexity, it can be observed in Table II that the RTF of the real- and complex-valued masking algorithms as well as the DMFF algorithm is lower than the RTF of the R1 and CD deep MFMVDR filters, due to the absence of MFMVDR linear algebra operations. Hence, for the considered algorithms employing a TCN architecture, a trade-off exists between speech enhancement performance and computational complexity.

VII. Conclusion

In this article, we proposed to integrate the MFMVDR filter for single-microphone speech enhancement into a deep learning framework. We proposed to estimate the required quantities of the MFMVDR filter, i.e., the noisy and interference covariance matrices as well as the a-priori SIR, using a signal-based loss function defined at the output of the MFMVDR filter. For the covariance matrices, we investigated different matrix structures, namely Hermitian positive-definite, Hermitian positive-definite Toeplitz (assuming stationarity of the noisy and interference components over N frames) and rank-1 (assuming a dominant principal subspace). For the Hermitian positive-definite matrix structure, we proposed an estimation procedure based on recursive smoothing, requiring one real-valued parameter for each covariance matrix, and an estimation procedure based on the Cholesky decomposition, requiring N^2 real-valued parameters for each covariance matrix. For the Hermitian positive-definite Toeplitz matrix structure, we proposed an estimation procedure involving balanced Vandermonde matrices, requiring $2N$ real-valued parameters for each covariance matrix. For the rank-1 matrix structure, we showed that the MFMVDR filter can be written as a linear combination of the DNN outputs, circumventing computationally complex matrix inversions, and proposed an estimation procedure requiring $2N$ real-valued parameters for each covariance matrix.

Experimental results on the DNS challenge dataset at SNRs ranging from 0 dB to 19 dB show that the matrix estimation procedure based on the Cholesky decomposition yields the best speech enhancement performance, closely followed by the computationally less complex rank-1 matrix estimation procedure. The matrix estimation procedures based on recursive smoothing and the positive-definite Toeplitz matrix structure yield the lowest speech enhancement performance, hinting that updating the covariance matrices based on recursive smoothing or assuming stationarity are not as appropriate. In addition, the simulation results show that the best-performing signal-based deep MFMVDR filter outperforms real- and complex-valued masking as well as direct multi-frame filtering, demonstrating the benefit of imposing structure on the multi-frame filters, and a competitive performance compared with state-of-the-art algorithms is demonstrated.

[3][Online]. Available: https://uol.de/en/sigproc/research/audio-demos/multi-frame-speech-enhancement/deep-mfmvdr-journal

References

[1] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer, 2011.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. San Rafael, CA, USA: Morgan & Claypool, 2013.

[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] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Hoboken, NJ, USA: Wiley, 2018.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[6] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via tasnet," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 36–40.

[7] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[8] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Process. Lett.*, vol. 9, no. 7, pp. 204–206, Jul. 2002.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[10] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Signal Process. Lett.*, vol. 15, pp. 213–216, 2008.

[11] S. Venkataramani, J. Casebeer, and P. Smaragdis, "End-to-End source separation with adaptive front-ends," in *Proc. IEEE 52nd Asilomar Conf. Signals, Syst., Comput.*, 2018, pp. 684–688.

[12] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," Apr. 2018, *arXiv:1803.01271*.

[13] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[14] D. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[15] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 370–382, May 2019.

[16] J. Lee and H. G. Kang, "A joint learning algorithm for complex-valued T-f masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1098–1109, Jun. 2019.

[17] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2019.

[18] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *Proc. Innov. Appl. Artif. Intell.*, 2020, pp. 9458–9465.

[19] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2472–2476.

[20] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

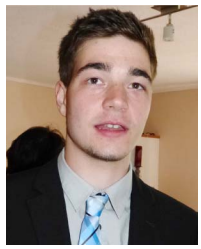
[21] W. Mack and E. A. P. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Process. Lett.*, vol. 27, pp. 61–65, 2020.

[22] B. D. Van Veen and K. M. Buckley, "Beamforming: A. versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[23] Z. Zhang et al., "Multi-channel multi-frame ADL-MVDR for target speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3526–3540, 2021.

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

[25] D. Fischer and S. Doclo, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *Proc. 25th Eur. Signal Process. Conf.*, 2017, pp. 603–607.
[26] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1355–1365, Sep. 2014.
[27] D. Fischer and S. Doclo, "Subspace-based speech correlation vector estimation for single-microphone multi-frame MVDR filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 856–860.
[28] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 191–195.
[29] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 8443–8447.
[30] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[31] 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, Jan. 2015.
[32] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[33] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2014, pp. 577–581.
[34] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 708–712.
[35] Z. Xu, S. Elshamy, Z. Zhao, and T. Fingscheidt, "Components loss for neural networks in mask-based speech enhancement," *EURASIP J. Audio, Speech, Music Process.*, vol. 2021, no. 24, pp. 1–20, 2021.
[36] Q. Zhang, A. M. Nicolson, M. Wang, K. Paliwal, and C.-X. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.
[37] D. Fischer, K. Brümann, and S. Doclo, "Comparison of parameter estimation methods for single-microphone multi-frame wiener filtering," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1809–1813.
[38] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1996.
[39] T. Bäckström, "Vandermonde factorization of toeplitz matrices and applications in filtering and warping," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6257–6263, Dec. 2013.
[40] T. Gerkmann and R. Hendriks, "Unbiased MMSE-Based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[41] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.
[42] F. Vincent and O. Besson, "Steering vector uncertainties and diagonal loading," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, 2004, pp. 327–331.
[43] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1335–1345, Aug. 2019.
[44] C. K. A. Reddy et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 2492–2496.
[45] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
[46] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 776–780.
[47] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 411–412.
[48] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 3591–3591, May 2013.
[49] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1509–1512.
[50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–18.
[51] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6648–6652.
[52] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.
[53] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) — A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.
[54] 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, Sep. 2011.
[55] C. K. A. ReddyV. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6493–6497.



Marvin Tammen (Student Member, IEEE) received the B.Eng. degree and the M.Sc. degree from University of Oldenburg, Oldenburg, Germany, in 2015 and 2018, respectively. Since September 2018, he has been working toward the Dr.-Ing. degree with the Signal Processing Group, University of Oldenburg. His research interests include signal processing for single and multi-microphone speech enhancement aided by deep neural networks.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from KU Leuven, Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with KU Leuven and McMaster University, Hamilton, ON, Canada. From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors in Leuven, Belgium. Since 2009, he has been a Full Professor with the University of Oldenburg, Oldenburg, Germany, and scientific Advisor for the Branch Hearing, Speech and Audio Technology HSA of the Fraunhofer Institute for Digital Media Technology IDMT. His research interests include center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. Prof. Doclo was the recipient of the several best paper awards including International Workshop on Acoustic Echo and Noise Control 2001, EURASIP Signal Processing 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019. He was a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing and a Member of the EAA Technical Committee on Audio Signal Processing. Since 2021, he has been the Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.