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

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

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

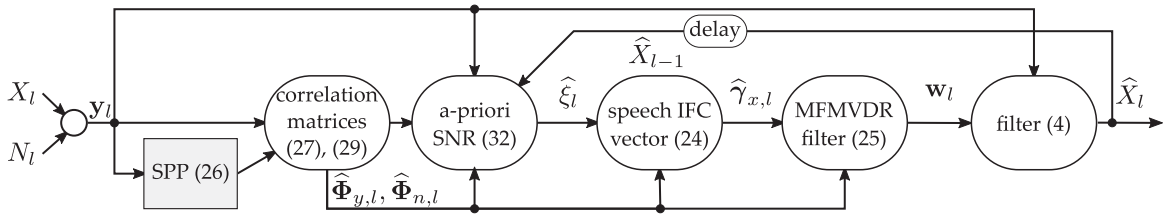
In this approach, a DNN is trained to map the logarithmic magnitude of the noisy STFT coefficients to an estimate of the SPP, i.e.,

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

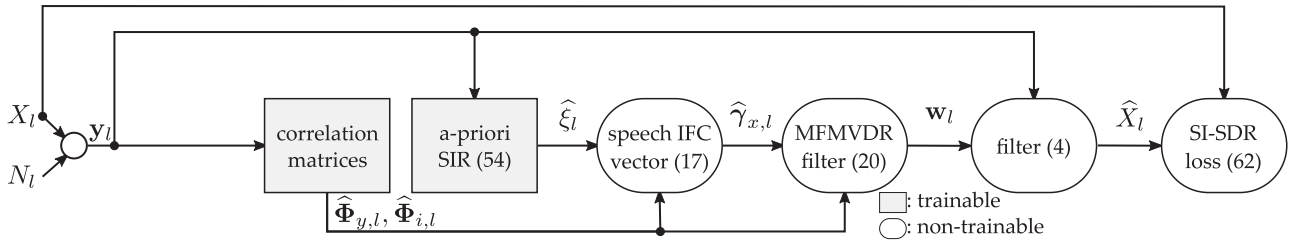
The training target is the SPP defined in [40] using oracle information about the noise component, i.e., no end-to-end training using a signal-based loss function is performed (for more details on the training procedure, we refer to Section V-C). The SPP estimate in (26) is then used to estimate the 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.,

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