2022 International Workshop on Acoustic Signal Enhancement (IWAENC) | 978-1-6654-6867-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IWAENC53105.2022.9914742

DEEP MULTI-FRAME MVDR FILTERING FOR BINAURAL NOISE REDUCTION

Marvin Tammen, Simon Doclo

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all
University of Oldenburg, Germany
{marvin.tammen, simon.doclo}@uni-oldenburg.de

ABSTRACT

To improve speech intelligibility and speech quality in noisy environments, binaural noise reduction algorithms for head-mounted assistive listening devices are of crucial importance. Several binaural noise reduction algorithms such as the well-known binaural minimum variance distortionless response (MVDR) beamformer have been proposed, which exploit spatial correlations of both the target speech and the noise components. Furthermore, for single-microphone scenarios, multi-frame algorithms such as the multi-frame MVDR (MFMVDR) filter have been proposed, which exploit temporal instead of spatial correlations. In this contribution, we propose a binaural extension of the MFMVDR filter, which exploits *both* spatial and temporal correlations. The binaural MFMVDR filters are embedded in an end-to-end deep learning framework, where the required parameters, i.e., the speech spatio-temporal correlation vectors as well as the (inverse) noise spatio-temporal covariance matrix, are estimated by temporal convolutional networks (TCNs) that are trained by minimizing the mean spectral absolute error loss function. Simulation results comprising measured binaural room impulses and diverse noise sources at signal-to-noise ratios from -5 dB to 20 dB demonstrate the advantage of utilizing the binaural MFMVDR filter structure over directly estimating the binaural multi-frame filter coefficients with TCNs.

Index Terms— binaural noise reduction, multi-frame filtering, supervised learning

1. INTRODUCTION

In many speech communication scenarios, head-mounted assistive listening devices such as binaural hearing aids capture not only the target speaker, but also ambient noise, resulting in a degradation of speech quality and speech intelligibility. Hence, several binaural noise reduction algorithms have been proposed, which typically assume that adjacent short-time Fourier transform (STFT) coefficients are uncorrelated over time. This assumption is suitable when considering sufficiently long frames and a small frame overlap. In that case, the speech STFT coefficients at a left and right reference microphone can be estimated by applying (complex-valued) single-frame binaural filters to the available microphone signals. Several approaches have been proposed to estimate these single-frame binaural filters, which can be categorized into statistical model-based approaches (e.g., [1]–[4]) and supervised learning-based approaches (e.g., [5]–[11]). While the statistical model-based approaches can be mainly differentiated w.r.t. their underlying optimization problem and how the required parameters are estimated, the supervised learning-based approaches mainly differ in the used deep neural network (DNN) architecture and loss function.

With the goal of exploiting temporal correlations between neighboring STFT coefficients, multi-frame methods have been proposed for both single- and multi-microphone noise reduction, which apply (complex-valued) multi-frame filters to the most recent noisy STFT coefficients of each microphone. Similarly to the single-frame methods mentioned above, several approaches have been proposed to estimate these multi-frame filters, which can again be categorized into statistical model-based approaches (e.g., [12], [13]) and supervised learning-based approaches (e.g., [14]–[18]). In contrast to the single-frame approaches, however, there is a lack of studies that considered multi-frame approaches for *binaural* noise reduction.

Aiming at utilizing both spatial correlations as in the binaural minimum variance distortionless response (MVDR) beamformer [1], [3] and temporal correlations as in the multi-frame MVDR (MFMVDR) filter [12], [16], we propose to extend the MFMVDR filter to binaural listening scenarios. To implement the binaural MFMVDR filter, estimates of the speech spatio-temporal correlation vectors (STCVs) as well as the (inverse) noise spatio-temporal covariance matrix (STCM) are required. Similarly as in [16], the binaural MFMVDR filter is embedded in an end-to-end supervised learning framework as shown in Fig. 1, where all required parameters are estimated using temporal convolutional networks (TCNs) that are trained using the mean spectral absolute error (MSAE) loss function [19]. Simulation results using measured binaural room impulse responses from [20] as well as clean speech and noise from the third Deep Noise Suppression Challenge (DNS3) [21] at signal-to-noise-ratios (SNRs) from $-5\,\mathrm{dB}$ to $20\,\mathrm{dB}$ show that the proposed deep binaural MFMVDR filter outperforms directly estimating the single- or multi-frame binaural filter coefficients using TCNs, i.e., without exploiting the structure of the deep binaural MFMVDR filter.

2. SIGNAL MODEL

We consider an acoustic scenario with a single speech source and a single noise source, both located in a reverberant room, recorded by binaural hearing aids with M microphones. In the STFT domain, the noisy microphone signals $y_{m,f,t}$ are given by

$$y_{m,f,t} = x_{m,f,t} + n_{m,f,t}, \tag{1}$$

where $x_{m,f,t}$ and $n_{m,f,t}$ denote the speech and noise components, respectively, at the m-th microphone, the f-th frequency bin, and the t-th time frame. Since all frequency bins are processed independently, the index f will be omitted in the remainder of this paper.

In *single*-microphone multi-frame noise reduction algorithms [12],

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2177/1 - Project ID 390895286.

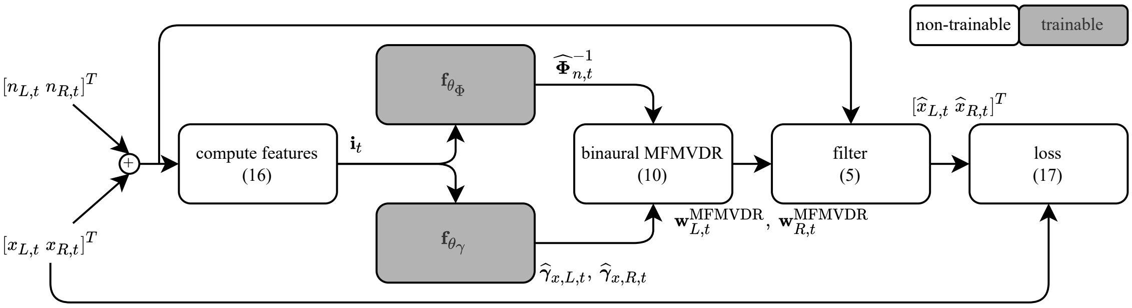
non-trainable
trainable
$[n_{L,t}\ n_{R,t}]^T$
$[x_{L,t}\ x_{R,t}]^T$
compute features
(16)
$\mathbf{i}_t$
$\mathbf{f}_{\theta_\Phi}$
$\mathbf{f}_{\theta_\gamma}$
$\widehat{\mathbf{\Phi}}_{n,t}^{-1}$
binaural MFMVDR
(10)
$\mathbf{w}_{L,t}^{\text{MFMVDR}}$, $\mathbf{w}_{R,t}^{\text{MFMVDR}}$
$\widehat{\gamma}_{x,L,t}$, $\widehat{\gamma}_{x,R,t}$
filter
(5)
$[\widehat{x}_{L,t}\ \widehat{x}_{R,t}]^T$
loss
(17)

Fig. 1. Block diagram of the proposed deep binaural MFMVDR filter.

[16], the noisy multi-frame vector $\bar{\mathbf{y}}_{m,t} \in \mathbb{C}^N$ is defined as

$$\bar{\mathbf{y}}_{m,t} = \begin{bmatrix} y_{m,t} & \cdots & y_{m,t-N+1} \end{bmatrix}^{\mathsf{T}}, \tag{2}$$

with $\circ^{\mathsf{T}}$ denoting the transpose operator, such that (1) can be written as $\bar{\mathbf{y}}_{m,t} = \bar{\mathbf{x}}_{m,t} + \bar{\mathbf{n}}_{m,t}$. In this case, using a complex-valued multi-frame filter $\bar{\mathbf{w}}_{m,t} \in \mathbb{C}^N$, the speech component $x_{m,t}$ is estimated as

$$\widehat{x}_{m,t} = \bar{\mathbf{w}}_{m,t}^{\mathsf{H}} \bar{\mathbf{y}}_{m,t}, \tag{3}$$

where $\circ^{\mathsf{H}}$ denotes the conjugate transpose operator.

In *multi*-microphone multi-frame noise reduction algorithms [13], [17], [18], the noisy multi-microphone multi-frame vector $\mathbf{y}_{m,t} \in \mathbb{C}^{NM}$ is defined as

$$\mathbf{y}_t = \begin{bmatrix} \bar{\mathbf{y}}_{1,t}^{\mathsf{T}} & \cdots & \bar{\mathbf{y}}_{M,t}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}, \tag{4}$$

such that (1) can be written as $\mathbf{y}_t = \mathbf{x}_t + \mathbf{n}_t$. Without loss of generality, in this paper we consider the case $M = 2$, with one hearing aid per side and one microphone per hearing aid, i.e., $m \in \{L, R\}$, where L and R denote the left and right side, respectively. In this case, using (complex-valued) binaural multi-frame filters $\mathbf{w}_{m,t} \in \mathbb{C}^{2N}$ with $2N$ taps each, the binaural speech components are estimated as

$$\widehat{x}_{m,t} = \mathbf{w}_{m,t}^{\mathsf{H}} \mathbf{y}_t. \tag{5}$$

Assuming that the speech and noise components are spatio-temporally uncorrelated, the noisy spatio-temporal covariance matrix (STCM) $\mathbf{\Phi}_{y,t} = \mathcal{E}\{\mathbf{y}_t \mathbf{y}_t^{\mathsf{H}}\} \in \mathbb{C}^{2N \times 2N}$, with $\mathcal{E}\{\circ\}$ the expectation operator, can be written as

$$\mathbf{\Phi}_{y,t} = \mathbf{\Phi}_{x,t} + \mathbf{\Phi}_{n,t}, \tag{6}$$

where $\mathbf{\Phi}_{x,t}$ and $\mathbf{\Phi}_{n,t}$ are defined similarly as $\mathbf{\Phi}_{y,t}$.

In order to exploit speech correlations across successive time frames, it has been proposed in [12] to decompose the (single-microphone) multi-frame speech vector into a temporally correlated and a temporally uncorrelated component. Similarly, the binaural multi-frame speech vector $\mathbf{x}_t$ can be decomposed into a spatio-temporally correlated and a spatio-temporally uncorrelated component w.r.t. the current left or the right speech STFT coefficient $x_{m,t}$:

$$\mathbf{x}_t = \underbrace{\boldsymbol{\gamma}_{x,m,t} x_{m,t}}_{\text{correlated}} + \underbrace{\mathbf{x}_{m,t}'}_{\text{uncorrelated}} \tag{7}$$

The highly time-varying left or right speech spatio-temporal correlation vector (STCV) $\boldsymbol{\gamma}_{x,m,t} \in \mathbb{C}^{2N}$ describes the correlation between the N most recent left and right speech STFT coefficients and the current left or the right speech STFT coefficient $x_{m,t}$, and it is defined as

$$\boldsymbol{\gamma}_{x,m,t} = \frac{\mathcal{E}\{\mathbf{x}_t x_{m,t}^*\}}{\mathcal{E}\{|x_{m,t}|^2\}}, \tag{8}$$

where $\circ^*$ denotes the conjugate operator and with $\mathbf{e}_L^{\mathsf{T}} \boldsymbol{\gamma}_{x,L,t} = \mathbf{e}_R^{\mathsf{T}} \boldsymbol{\gamma}_{x,R,t} = 1$. Here, $\mathbf{e}_L$ and $\mathbf{e}_R$ denote selection vectors with their first or $N+1$-th element equal to 1, respectively, and the other elements equal to 0.

3. DEEP BINAURAL MULTI-FRAME MVDR FILTER

Aiming at minimizing the output noise power spectral density while leaving the correlated speech component undistorted, in [12] the MFMVDR filter for single-microphone noise reduction has been proposed. In this paper, we propose to extend the single-microphone MFMVDR filter to binaural scenarios by considering the spatio-temporal correlations of the speech and noise components for the left and right side, i.e.,

$$\underset{\mathbf{w}_{m,t}}{\operatorname{argmin}} \quad \mathbf{w}_{m,t}^{\mathsf{H}} \mathbf{\Phi}_{n,t} \mathbf{w}_{m,t} \quad \text{s.t.} \quad \mathbf{w}_{m,t}^{\mathsf{H}} \boldsymbol{\gamma}_{x,m,t} = 1. \tag{9}$$

Solving this optimization problem, the binaural MFMVDR filters are given by

$$\boxed{\mathbf{w}_{m,t}^{\text{MFMVDR}} = \frac{\mathbf{\Phi}_{n,t}^{-1} \boldsymbol{\gamma}_{x,m,t}}{\boldsymbol{\gamma}_{x,m,t}^{\mathsf{H}} \mathbf{\Phi}_{n,t}^{-1} \boldsymbol{\gamma}_{x,m,t}}} \tag{10}$$

As has been shown for the *single-microphone* MFMVDR filter [22], the performance of the (binaural) MFMVDR filter depends on how well the required parameters, i.e., the inverse noise STCM as well as the speech STCVs, are estimated from the noisy STFT coefficients. In contrast to using statistical model-based estimators similar to [23], we embed the binaural MFMVDR filter in an end-to-end supervised learning framework similar to [16], with the parameters estimated by TCNs (see Fig. 1). The TCNs are trained by minimizing the MSAE loss function [19] computed at the output of the deep binaural MFMVDR filter instead of providing explicit parameter labels. A-priori knowledge about the properties of the estimated parameters is exploited as described in the following two sections.

3.1. Speech Spatio-Temporal Correlation Vector

The left and right speech STCVs each are two $2N$-dimensional complex-valued vectors (cf. (8)), hence consisting of $8N$ *real*-valued coefficients $\mathbf{h}_{\gamma,t}^{\mathbb{R}} \in \mathbb{R}^{8N}$ ($4N$ for the real part and $4N$ for the imaginary part). To estimate these real-valued coefficients, we propose to use a TCN $\mathbf{f}_{\theta_\gamma}$ with parameters $\boldsymbol{\theta}_\gamma$, which is fed input features $\mathbf{i}_t$ derived from the noisy STFT coefficients, i.e.,

$$\widehat{\mathbf{h}}_{\gamma,t}^{\mathbb{R}} = \mathbf{f}_{\theta_\gamma}\{\mathbf{i}_t\}, \tag{11}$$

with the features $\mathbf{i}_t$ defined in (16). To construct a $4N$-dimensional complex-valued vector $\widehat{\mathbf{h}}^{\mathbb{C}}_{\gamma,t}$ from the $8N$-dimensional real-valued vector $\widehat{\mathbf{h}}^{\mathbb{R}}_{\gamma,t}$, the first $4N$ elements of $\widehat{\mathbf{h}}^{\mathbb{R}}_{\gamma,t}$ are used for the real components and the second $4N$ elements are used for the imaginary components, i.e.,

$$\widehat{\mathbf{h}}^{\mathbb{C}}_{\gamma,t} = [\widehat{\mathbf{h}}^{\mathbb{R}}_{\gamma,t}]_{0:4N-1} + j\,[\widehat{\mathbf{h}}^{\mathbb{R}}_{\gamma,t}]_{4N:8N-1}, \tag{12}$$

where $j^2 = -1$. To ensure that the first or $N+1$-th element of the speech STCVs is equal to 1 (cf. (8)), the speech STCVs are finally obtained as

$$\widehat{\gamma}_{x,L,t} = \frac{[\widehat{\mathbf{h}}^{\mathbb{C}}_{\gamma,t}]_{0:2N-1}}{\mathbf{e}^{\mathsf{T}}_{L}[\widehat{\mathbf{h}}^{\mathbb{C}}_{\gamma,t}]_{0:2N-1}}, \quad \widehat{\gamma}_{x,R,t} = \frac{[\widehat{\mathbf{h}}^{\mathbb{C}}_{\gamma,t}]_{2N:4N-1}}{\mathbf{e}^{\mathsf{T}}_{R}[\widehat{\mathbf{h}}^{\mathbb{C}}_{\gamma,t}]_{2N:4N-1}}. \tag{13}$$

3.2. Spatio-Temporal Covariance Matrices

Since the $2N \times 2N$-dimensional STCM $\mathbf{\Phi}_{n,t}$ can be assumed to be Hermitian positive-definite, also its inverse $\mathbf{\Phi}^{-1}_{n,t}$ as required in (10) can be assumed to be Hermitian positive-definite. Hence, $\mathbf{\Phi}^{-1}_{n,t}$ has a unique Cholesky decomposition [24]:

$$\mathbf{\Phi}^{-1}_{n,t} = \mathbf{L}_t \mathbf{L}^{\mathsf{H}}_t, \tag{14}$$

with $\mathbf{L}_t \in \mathbb{C}^{2N\times 2N}$ a lower triangular matrix with positive real-valued diagonal. Due to its structure, $\mathbf{L}$ is determined by $(2N)^2$ real-valued coefficients. Similarly to the procedure for estimating the speech STCVs, we use a TCN $\mathbf{f}_{\theta_\Phi}$ with parameters $\boldsymbol{\theta}_\Phi$, which is fed input features $\mathbf{i}_t$, to estimate these real-valued coefficients $\widehat{\mathbf{h}}^{\mathbb{R}}_{\Phi,t} \in \mathbb{R}^{(2N)^2}$, i.e.,

$$\widehat{\mathbf{h}}^{\mathbb{R}}_{\Phi,t} = \mathbf{f}_{\theta_\Phi}\{\mathbf{i}_t\}. \tag{15}$$

Using $\widehat{\mathbf{h}}^{\mathbb{R}}_{\Phi,t}$, the lower triangular matrix with positive real-valued diagonal $\widehat{\mathbf{L}}_t$ is assembled. Finally, an estimate of $\mathbf{\Phi}^{-1}_{n,t}$ is obtained using (14) by replacing $\mathbf{L}_t$ with its estimate $\widehat{\mathbf{L}}_t$.

4. SIMULATIONS

In this section, the binaural noise reduction performance of the proposed deep binaural MFMVDR filter is compared with a number of baseline algorithms, which are described in Section 4.1. Sections 4.2 and 4.3 deal with the used datasets and the simulation settings, respectively. In Section 4.4, the simulation results are presented in terms of the perceptual evaluation of speech quality (PESQ) [25] and frequency-weighted segmental SNR (FWSSNR) [26] improvement.

4.1. Baseline Algorithms

The following baseline algorithms have been considered to allow investigating the effect of not using vs. using the proposed deep binaural MFMVDR structure for binaural multi-frame filtering. To achieve this goal, for the baseline algorithms the binaural multi-frame filters in (5) are not obtained using the binaural MFMVDR structure. Instead, the real and imaginary components of the baseline binaural multi-frame filters are directly estimated by a TCN, i.e., without the intermediate steps of speech STCVs and inverse noise STCM estimation and computation of (10). In addition, we investigate the effect of binaural single-frame vs. binaural multi-frame filtering. More specifically, we use the following end-to-end supervised learning-based baseline algorithms:

direct binaural single-frame filtering With $N = 1$ and $\mathbf{w}^{\mathrm{B1}}_{m,t} \in \mathbb{C}^2$, only spatial filtering is performed. The filter coefficients are estimated using a TCN $\mathbf{f}_{\mathrm{B1}}$ with parameters $\boldsymbol{\theta}_{\mathrm{B1}}$, i.e., $\mathbf{w}^{\mathrm{B1}}_{m,t} = \mathbf{f}_{\mathrm{B1}}\{\mathbf{i}_t\}$. The real and imaginary parts of the filter coefficients $\mathbf{w}^{\mathrm{B1}}_{m,t}$ are bounded to $[-1, 1]$ using a hyperbolic tangent activation function.

direct binaural multi-frame filtering With $N = 3$ and $\mathbf{w}^{\mathrm{B2}}_{m,t} \in \mathbb{C}^{2N}$, both spatial and temporal filtering are performed. The filter coefficients are estimated using a TCN $\mathbf{f}_{\mathrm{B2}}$ with parameters $\boldsymbol{\theta}_{\mathrm{B2}}$, i.e., $\mathbf{w}^{\mathrm{B2}}_{m,t} = \mathbf{f}_{\mathrm{B2}}\{\mathbf{i}_t\}$. The real and imaginary parts of the filter coefficients $\mathbf{w}^{\mathrm{B2}}_{m,t}$ are bounded to $[-1, 1]$ using a hyperbolic tangent activation function. These bounds are motivated by [14].

4.2. Dataset

To train and validate the considered algorithms, we used simulated binaural room impulse responses (BRIRs) from the training subset of the first Clarity Enhancement Challenge (CEC1) dataset [27] as well as clean speech (English read book sentences) and noise from the training subset of the third Deep Noise Suppression Challenge (DNS3) dataset [21]. These BRIRs were simulated by considering a randomly positioned directed speech source and an omnidirectional noise point source captured by binaural behind-the-ear hearing aids in randomly sized rooms with ”low to moderate” reverberation, i.e., around $0.2\,\mathrm{s}$ to $0.4\,\mathrm{s}$. The speech source was always located at an angle within $\pm 30^\circ$ w.r.t. the listener, while the noise source could be positioned everywhere in the room except for less than $1\,\mathrm{m}$ from the walls or the listener. Surface absorption coefficients were varied to simulate various room characteristics such as doors, windows, curtains, rugs, or furniture. In total, 6000 room configurations were considered. Clean speech and noise were convolved with their corresponding BRIRs before being mixed at better ear SNRs from $0\,\mathrm{dB}$ to $15\,\mathrm{dB}$. In total, the training and validation datasets have a length of $80\,\mathrm{h}$ and $20\,\mathrm{h}$, respectively.

To evaluate the considered algorithms, we used measured BRIRs from the dataset proposed in [20] as well as clean speech and noise from the official test subset of the deep noise suppression (DNS) dataset [28]. The dataset in [20] comprises BRIRs measured with binaural behind-the-ear hearing aids “for multiple, realistic head and sound-source positions in four natural environments reflecting daily-life communication situations with different reverberation times”. The configuration of these hearing aids matches the configuration considered in the training and validation datasets. Clean speech and noise were convolved with the BRIRs before being mixed at better ear SNRs from $-5\,\mathrm{dB}$ to $20\,\mathrm{dB}$. In total, 100 utterances, each of length $10\,\mathrm{s}$, were considered in the evaluation. Especially due to the use of simulated vs. measured BRIRs, there is considerable mismatch between the training and validation datasets on the one hand and the evaluation dataset on the other hand. All datasets were used at a sampling frequency of $16\,\mathrm{kHz}$.

4.3. Settings

For the STFT, $\sqrt{\text{Hann}}$ windows with a relatively small frame length of $8\,\mathrm{ms}$ and $75\,\%$ overlap were used in order to increase speech interframe correlations. As input features, we used a concatenation of the logarithmic magnitude, the cosine of the phase, and the sine of the phase, of the noisy left and right STFT coefficients, i.e.,

$$\begin{aligned} \mathbf{i}_{m,t} = &\quad \left[\log_{10}|y_{m,t}| \quad \cos(\angle y_{m,t}) \quad \sin(\angle y_{m,t})\right]^{\mathsf{T}} \\ \mathbf{i}_t = &\quad [\mathbf{i}^{\mathsf{T}}_{L,t} \quad \mathbf{i}^{\mathsf{T}}_{R,t}]]^{\mathsf{T}}, \end{aligned} \tag{16}$$

where $\angle\circ$ denotes the phase of $\circ$. Note that both the cosine and sine of the noisy phase are chosen to prevent an ambiguous phase representation.

The multi-frame algorithms use $N = 5$ frames, resulting in the capability of exploiting temporal correlations within $16\,\mathrm{ms}$. To decrease distortion of the speech and residual noise components, a minimum gain of $-20\,\mathrm{dB}$ was included in all algorithms.

To estimate the required parameters of the deep binaural MFMVDR filter or the filter coefficients of the baseline algorithms, we used causal TCNs, with their hyperparameters fixed to 2 stacks of 6 layers, yielding a temporal receptive field size of $512\,\mathrm{ms}$. Since the deep binaural MFMVDR filter uses two TCNs and the number of real-valued coefficients differs per considered algorithm, the hidden dimension size of the TCNs was varied per algorithm to result in similar numbers of trainable weights for all algorithms, i.e., 6.2×10^6. While also the other hyperparameters could have been varied to this end, only varying the hidden dimension size results in TCNs with the same temporal receptive field size, which is required for a fair comparison. To prevent division by 0, a small constant was added to the denominator in (13).

As loss function, the MSAE proposed in [19] was used, where the loss was averaged across the batch, the left and right output signals, and the frequency bins and time frames, i.e.,

$$L_{b,m,f,t} = \beta|x_{b,m,f,t} - \widehat{x}_{b,m,f,t}| + (1-\beta)\big||x_{b,m,f,t}| - |\widehat{x}_{b,m,f,t}|\big|$$
$$L = \frac{1}{2BFT}\sum_{b=0}^{B-1}\sum_{m\in\{L,R\}}\sum_{f=0}^{F-1}\sum_{t=0}^{T-1} L_{b,m,f,t}, \tag{17}$$

where B denotes the batch size, F and T denote the numbers of frequency bins and time frames in an utterance, and $\beta = 0.4$ [19].

The TCNs were implemented based on the official Conv-TasNet implementation[1], and they were trained for a maximum of 150 epochs with early stopping using the AdamW optimizer [29]. The learning rate was initialized as 3×10^{-4}, and it was halved after 3 epochs without an improvement on the validation dataset. Gradient ℓ_2-norms were clipped to 5, and the batch size was 8.

The simulations were implemented using PyTorch 1.10 [30] and performed on NVIDIA GeForce®RTX A5000 graphics cards. A PyTorch implementation of the compared algorithms as well as the model weights used in the evaluation will be made publicly available upon publication.

4.4. Results

For all considered algorithms, Fig. 2 depicts the improvement in terms of PESQ and FWSSNR w.r.t. the noisy microphone signals on the evaluation dataset. Note that, similarly as for the MSAE loss function in (17), PESQ and FWSSNR improvements are simply averaged across the left and right output signals [10].

First, a considerable improvement in terms of PESQ and FWSSNR can be observed for all algorithms, with the deep binaural MFMVDR filter outperforming the baseline algorithms. Second, comparing the baseline algorithms, it can be observed that increasing the degrees of freedom of the filter, i.e., by allowing for a binaural multi-frame vs. a binaural single-frame filter, improves binaural noise reduction performance. Third, by enforcing the binaural MFMVDR structure on the binaural multi-frame filter, binaural noise reduction performance is further increased.

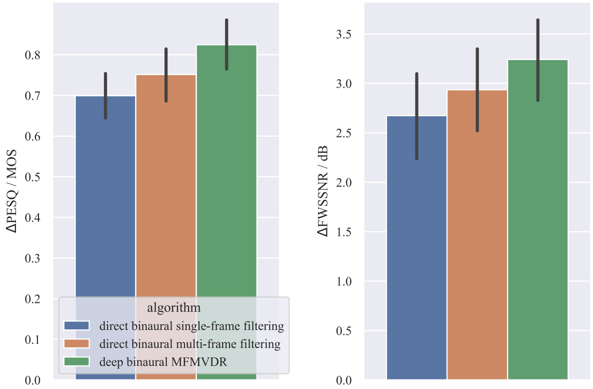

Fig. 2. Mean and standard deviation of the PESQ and FWSSNR improvements obtained on the evaluation dataset. The mean noisy PESQ score is $1.74\,\mathrm{MOS}$ and the mean noisy FWSSNR score is $14.08\,\mathrm{dB}$.

Audio examples for the compared algorithms are available online[2].

5. CONCLUSION

In this paper we proposed a binaural extension of the MFMVDR filter, which is capable of utilizing both spatial and temporal correlations of the speech and noise components. To estimate the speech STCVs as well as the inverse noise STCM required by the binaural MFMVDR filter, we use TCNs, which are trained by embedding the binaural MFMVDR filter in an end-to-end supervised learning framework and minimizing the MSAE loss function. Simulations comprising measured binaural room impulse responses as well as diverse noise sources at SNRs in $-5\,\mathrm{dB}$ to $20\,\mathrm{dB}$ demonstrate the advantage of binaural multi-frame filtering over binaural single-frame filtering as well as employing the binaural MFMVDR structure over directly estimating the single- or multi-frame binaural filters using TCNs.

References

[1] 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 Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[2] E. Hadad, S. Doclo, and S. Gannot, "The Binaural LCMV Beamformer and its Performance Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 543–558, Mar. 2016.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Jan. 2017.

[4] S. Doclo, S. Gannot, D. Marquardt, and E. Hadad, "Binaural Speech Processing with Application to Hearing Devices," in

[1] https://github.com/naplab/Conv-TasNet

[2] https://uol.de/en/sigproc/research/audio-demos/binaural-noise-reduction/deep-bmfmvdr

Audio Source Separation and Speech Enhancement, John Wiley & Sons, Ltd, Aug. 2018, pp. 413–442.

[5] A. H. Moore, L. Lightburn, W. Xue, P. A. Naylor, and M. Brookes, "Binaural Mask-Informed Speech Enhancement for Hearing AIDS with Head Tracking," in *Proc. 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 461–465.

[6] X. Sun, R. Xia, J. Li, and Y. Yan, "A Deep Learning Based Binaural Speech Enhancement Approach with Spatial Cues Preservation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019.

[7] C. Han, Y. Luo, and N. Mesgarani, "Real-Time Binaural Speech Separation with Preserved Spatial Cues," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6404–6408.

[8] Z. Sun, Y. Li, H. Jiang, F. Chen, X. Xie, and Z. Wang, "A Supervised Speech Enhancement Method for Smartphone-Based Binaural Hearing Aids," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 5, pp. 951–960, Oct. 2020.

[9] J.-H. Kim, J. Choi, J. Son, G.-S. Kim, J. Park, and J.-H. Chang, "MIMO Noise Suppression Preserving Spatial Cues for Sound Source Localization in Mobile Robot," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, Daegu, Korea, May 2021.

[10] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, and C. J. Smalt, "Speaker separation in realistic noise environments with applications to a cognitively-controlled hearing aid," *Neural Networks*, pp. 136–147, Mar. 2021.

[11] T. Green, G. Hilkhuysen, M. Huckvale, *et al.*, "Speech recognition with a hearing-aid processing scheme combining beamforming with mask-informed speech enhancement," *Trends in Hearing*, vol. 26, Jan. 2022.

[12] Y. A. Huang and J. Benesty, "A Multi-Frame Approach to the Frequency-Domain Single-Channel Noise Reduction Problem," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[13] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 305–308.

[14] W. Mack and E. A. P. Habets, "Deep Filtering: Signal Extraction and Reconstruction Using Complex Time-Frequency Filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2020.

[15] A. Aroudi, M. Delcroix, T. Nakatani, K. Kinoshita, S. Araki, and S. Doclo, "Cognitive-Driven Convolutional Beamforming Using EEG-Based Auditory Attention Decoding," in *Proc. IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, Espoo, Finland, Sep. 2020.

[16] M. Tammen and S. Doclo, "Deep Multi-Frame MVDR Filtering for Single-Microphone Speech Enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, Jun. 2021, pp. 8443–8447.

[17] Z. Zhang, Y. Xu, M. Yu, *et al.*, "Multi-Channel Multi-Frame ADL-MVDR for Target Speech Separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 3526–3540, Nov. 2021.

[18] Z.-Q. Wang, H. Erdogan, S. Wisdom, *et al.*, "Sequential Multi-Frame Neural Beamforming for Speech Separation and Enhancement," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, Jan. 2021, pp. 905–911.

[19] 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, and Language Processing*, vol. 28, pp. 1778–1787, May 2020.

[20] H. Kayser, S. D. Ewert, J. Anemuller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, Jun. 2009.

[21] C. K. Reddy, H. Dubey, K. Koishida, *et al.*, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. Interspeech*, Brno, Czech Republic, Aug. 2021, pp. 2796–2800.

[22] D. Fischer and S. Doclo, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *Proc. European Signal Processing Conference (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 603–607.

[23] A. Schasse and R. Martin, "Estimation of Subband Speech Correlations for Noise Reduction via MVDR Processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1355–1365, Sep. 2014.

[24] A.-L. Cholesky, "Note sur une méthode de resolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations lineaires en nombre inferieure à celui des inconnues," *Bulletin Géodésique*, vol. 2, no. 1, pp. 67–77, Apr. 1924.

[25] 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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, May 2001, pp. 749–752.

[26] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[27] S. Graetzer, J. Barker, T. J. Cox, *et al.*, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. Interspeech*, Brno, Czech Republic, Aug. 2021, pp. 686–690.

[28] C. K. Reddy, V. Gopal, R. Cutler, *et al.*, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2492–2496.

[29] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, May 2019.

[30] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, Dec. 2019.