ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413775

DEEP MULTI-FRAME MVDR FILTERING FOR SINGLE-MICROPHONE SPEECH ENHANCEMENT

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

Multi-frame algorithms for single-microphone speech enhancement, e.g., the multi-frame minimum variance distortionless response (MFMVDR) filter, are able to exploit speech correlation across adjacent time frames in the short-time Fourier transform (STFT) domain. Provided that accurate estimates of the required speech interframe correlation vector and the noise correlation matrix are available, it has been shown that the MFMVDR filter yields a substantial noise reduction while hardly introducing any speech distortion. Aiming at merging the speech enhancement potential of the MFMVDR filter and the estimation capability of temporal convolutional networks (TCNs), in this paper we propose to embed the MFMVDR filter within a deep learning framework. The TCNs are trained to map the noisy speech STFT coefficients to the required quantities by minimizing the scale-invariant signal-to-distortion ratio loss function at the MFMVDR filter output. Experimental results show that the proposed deep MFMVDR filter achieves a competitive speech enhancement performance on the Deep Noise Suppression Challenge dataset. In particular, the results show that estimating the parameters of an MFMVDR filter yields a higher performance in terms of PESQ and STOI than directly estimating the multi-frame filter or single-frame masks and than Conv-TasNet.

Index Terms— Single-Microphone Speech Enhancement, Multi-Frame Filtering, Temporal Convolutional Networks

1. INTRODUCTION

In many hands-free speech communication systems such as hearing aids, mobile phones and smart speakers, ambient noise may degrade the speech quality and intelligibility of the recorded microphone signals. To alleviate this issue, several single- and multi-microphone speech enhancement algorithms have been proposed [1–5]. Single-microphone speech enhancement algorithms typically 1) transform the noisy time-domain signal to a domain that is better suited for speech enhancement, e.g., the short-time Fourier transform (STFT) domain, 2) apply a (real- or complex-valued) gain/mask to the transform-domain coefficients to obtain an estimate of the clean speech, and 3) transform the modified coefficients back to the time-domain. For such single-frame algorithms, many traditional model-based approaches [1, 2, 6, 7] as well as supervised learning-based approaches [8–12] have been proposed. A disadvantage of single-frame algorithms is that attenuation of the noise component may be accompanied by some distortion of the speech component in the enhanced signal.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 390895286 – EXC 2177/1.

In contrast to single-frame algorithms, multi-frame algorithms have been proposed which apply a complex-valued filter to the noisy speech STFT coefficients [3]. Also for multi-frame algorithms both model-based approaches such as the multi-frame minimum variance distortionless response (MFMVDR) filter [13–16] as well as supervised learning-based approaches [17, 18] have been proposed. The MFMVDR filter has been derived to explicitly take speech correlations across adjacent time frames into account and requires an estimate of the noise correlation matrix and the so-called speech interframe correlation (IFC) vector in each time-frequency bin. Although it has been shown that the MFMVDR filter can yield a good noise reduction performance and little speech distortion [13, 15], its performance is very sensitive to estimation errors of the required quantities, in particular the speech IFC vector [15].

To estimate the speech IFC vector from the noisy speech STFT coefficients, several model-based approaches have been proposed. In [14] a maximum likelihood approach has been derived, assuming that the speech and noise IFC vectors follow multi-variate Gaussian distributions. This approach requires an estimate of the a-priori signal-to-noise ratio (SNR), which can be estimated, e.g., using the decision-directed approach [6] or using a supervised learning-based approach [19]. In [16] a subspace estimator has been proposed, which is based on a low-rank speech model. However, simulation results have shown that estimating the required quantities from the noisy speech STFT coefficients using these model-based approaches typically results in a large performance degradation compared to the oracle MFMVDR filter.

In this paper we propose to embed the MFMVDR filter within a deep learning framework as shown in Fig. 1. More in particular, we propose to train temporal convolutional networks [11, 20] to map the noisy speech STFT coefficients to the required quantities, i.e., the noise correlation matrix and the a-priori SNR, by minimizing the scale-invariant signal-to-distortion ratio loss function [21] at the MFMVDR filter output. Experimental results using the INTERSPEECH 2020 Deep Noise Suppression (DNS) Challenge dataset [22] show that the proposed deep MFMVDR filter outperforms complex-valued masking as well as directly estimating the multi-frame filter without exploiting the MFMVDR structure and Conv-TasNet [11].

2. SIGNAL MODEL

We consider an acoustic scenario with a single microphone recording one speech source and additive ambient noise. In the STFT-domain, the noisy microphone signal is given by

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

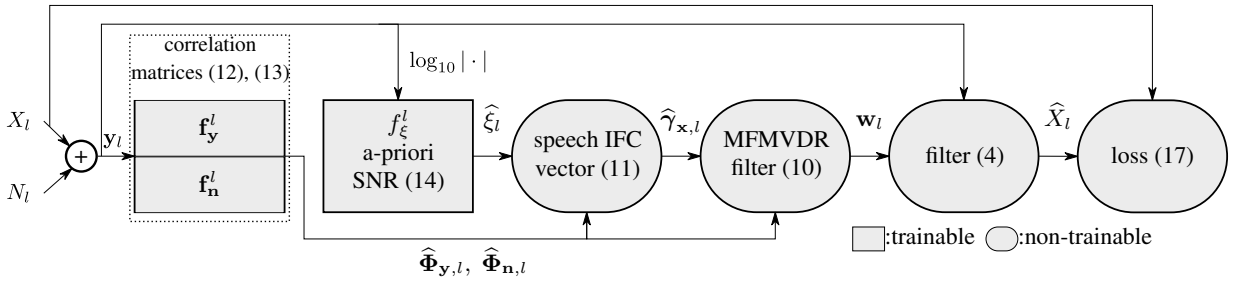

Fig. 1. Block diagram of the training process of the proposed deep MFMVDR filter. The speech enhancement-related loss function is used to update the weights of the temporal convolutional networks estimating the noisy speech and noise correlation matrices $\boldsymbol{\Phi}_{\mathbf{y},l}$ and $\boldsymbol{\Phi}_{\mathbf{n},l}$ as well as the a-priori SNR ξ_l.

where $Y_{k,l}$, $X_{k,l}$, and $N_{k,l}$ denote the noisy speech component, the speech component, and the noise component, respectively, at the k-th frequency bin and the l-th time frame. Since all frequency bins are assumed to be independent, the index k will be omitted in the remainder of this paper.

In single-frame speech enhancement algorithms [1,2,6,7,9–12], the speech component X_l is typically estimated by applying a (real- or complex-valued) gain/mask W_l to the noisy speech STFT coefficients, i.e.,

$$\widehat{X}_l = W_l Y_l. \tag{2}$$

In multi-frame speech enhancement algorithms [3], the N-dimensional noisy speech vector $\mathbf{y}_l$ is defined as

$$\mathbf{y}_l = [Y_l,\ Y_{l-1},\ \ldots,\ Y_{l-N+1}]^{\mathsf{T}}, \tag{3}$$

where $\circ^{\mathsf{T}}$ denotes the transpose operator. Using (1), the noisy speech vector $\mathbf{y}_l$ can be written as $\mathbf{y}_l = \mathbf{x}_l + \mathbf{n}_l$, where the speech vector $\mathbf{x}_l$ and the noise vector $\mathbf{n}_l$ are defined similarly as $\mathbf{y}_l$ in (3). The speech component X_l is then estimated by applying a (complex-valued) finite impulse response filter $\mathbf{w}_l$ with N taps to the noisy speech STFT coefficients, i.e.,

$$\widehat{X}_l = \mathbf{w}_l^{\mathsf{H}} \mathbf{y}_l, \tag{4}$$

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

Assuming that the speech and noise components are uncorrelated, the $N \times N$-dimensional noisy speech correlation matrix $\boldsymbol{\Phi}_{\mathbf{y},l} = \mathcal{E}\left\{\mathbf{y}_l \mathbf{y}_l^{\mathsf{H}}\right\}$, with $\mathcal{E}\{\circ\}$ the expectation operator, can be written as

$$\boldsymbol{\Phi}_{\mathbf{y},l} = \boldsymbol{\Phi}_{\mathbf{x},l} + \boldsymbol{\Phi}_{\mathbf{n},l}, \tag{5}$$

with the speech and noise correlation matrices $\boldsymbol{\Phi}_{\mathbf{x},l} = \mathcal{E}\left\{\mathbf{x}_l \mathbf{x}_l^{\mathsf{H}}\right\}$ and $\boldsymbol{\Phi}_{\mathbf{n},l} = \mathcal{E}\left\{\mathbf{n}_l \mathbf{n}_l^{\mathsf{H}}\right\}$. In [13], it has been proposed to exploit the speech correlation across adjacent time frames by decomposing the speech vector into a temporally correlated and a temporally uncorrelated part, i.e.,

$$\mathbf{x}_l = \underbrace{\boldsymbol{\gamma}_{\mathbf{x},l} X_l}_{\text{correlated}} + \underbrace{\mathbf{x}'_l}_{\text{uncorrelated}}, \tag{6}$$

where the (highly time-varying) speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x},l}$ describes the correlation between the current and previous time frames w.r.t. the speech STFT coefficient X_l, i.e.,

$$\boldsymbol{\gamma}_{\mathbf{x},l} = \frac{\mathcal{E}\left\{\mathbf{x}_l X_l^*\right\}}{\mathcal{E}\left\{|X_l|^2\right\}} = \frac{\boldsymbol{\Phi}_{\mathbf{x},l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\boldsymbol{\Phi}_{\mathbf{x},l}\mathbf{e}}, \tag{7}$$

where $\circ^*$ denotes the complex-conjugate operator, $\mathbf{e} = [1,\ 0,\ \ldots,\ 0]^{\mathsf{T}}$ is an N-dimensional selection vector, and $\mathbf{e}^{\mathsf{T}}\boldsymbol{\Phi}_{\mathbf{x},l}\mathbf{e} = \mathcal{E}\left\{|X_l|^2\right\} = \phi_{X,l}$ corresponds to the speech power spectral density (PSD). Using (5) and (7), the speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x},l}$ can be written as

$$\boxed{\boldsymbol{\gamma}_{\mathbf{x},l} = \frac{1+\xi_l}{\xi_l}\frac{\boldsymbol{\Phi}_{\mathbf{y},l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\boldsymbol{\Phi}_{\mathbf{y},l}\mathbf{e}} - \frac{1}{\xi_l}\underbrace{\frac{\boldsymbol{\Phi}_{\mathbf{n},l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\boldsymbol{\Phi}_{\mathbf{n},l}\mathbf{e}}}_{\boldsymbol{\gamma}_{\mathbf{n},l}}} \tag{8}$$

where $\xi_l = \frac{\phi_{X,l}}{\phi_{N,l}}$ denotes the a-priori SNR, with $\phi_{N,l} = \mathcal{E}\left\{|N_l|^2\right\} = \mathbf{e}^{\mathsf{T}}\boldsymbol{\Phi}_{\mathbf{n},l}\mathbf{e}$ the noise PSD, and $\boldsymbol{\gamma}_{\mathbf{n},l}$ denotes the noise IFC vector.

3. DEEP MULTI-FRAME MVDR FILTER

In [3, 13] the MFMVDR filter for single-microphone speech enhancement has been proposed, which aims at minimizing the output noise PSD while not distorting the correlated speech component $\boldsymbol{\gamma}_{\mathbf{x},l} X_l$, i.e.,

$$\min_{\mathbf{w}_l \in \mathbb{C}^N} \quad \mathbf{w}_l^{\mathsf{H}} \boldsymbol{\Phi}_{\mathbf{n},l} \mathbf{w}_l, \quad \text{s.t. } \mathbf{w}_l^{\mathsf{H}} \boldsymbol{\gamma}_{\mathbf{x},l} = 1. \tag{9}$$

Solving this constrained optimization problem yields the MFMVDR filter vector:

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

To implement the MFMVDR filter, estimates of the noise correlation matrix $\boldsymbol{\Phi}_{\mathbf{n},l}$ as well as the speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x},l}$ are required. In [14, 15] it has been shown that the speech enhancement performance of the MFMVDR filter strongly depends on how well these quantities can be estimated from the noisy speech STFT coefficients. Previously proposed model-based approaches [14, 16] simply use an estimate of the noisy speech correlation matrix $\boldsymbol{\Phi}_{\mathbf{y},l}$ instead of the noise correlation matrix $\boldsymbol{\Phi}_{\mathbf{n},l}$ (leading to the multi-frame minimum *power* distortionless response filter) and estimate the speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x},l}$ based on (8) by using the decision-directed approach [6] to estimate the a-priori SNR ξ_l and assuming the noise IFC vector $\boldsymbol{\gamma}_{\mathbf{n},l}$ to be constant for all time-frequency points.

In this paper we propose to estimate the required quantities for the MFMVDR filter in (10) from the noisy speech STFT coefficients using a supervised learning-based approach by minimizing a speech enhancement-related loss function at the MFMVDR filter output.

3.1. Speech IFC Vector

Due to its highly time-varying nature, the speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x},l}$ is difficult to estimate accurately. Since preliminary experiments have

shown that using (8) instead of directly estimating $\boldsymbol{\gamma}_{\mathbf{x},l}$ with a DNN yields a higher speech enhancement performance, we propose to estimate $\boldsymbol{\gamma}_{\mathbf{x},l}$ using the estimated correlation matrices $\widehat{\boldsymbol{\Phi}}_{\mathbf{y},l}$ and $\widehat{\boldsymbol{\Phi}}_{\mathbf{n},l}$ as well as the estimated a-priori SNR $\widehat{\xi}_l$, i.e.,

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\mathbf{x},l} = \frac{1+\widehat{\xi}_l}{\widehat{\xi}_l} \frac{\widehat{\boldsymbol{\Phi}}_{\mathbf{y},l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\widehat{\boldsymbol{\Phi}}_{\mathbf{y},l}\mathbf{e}} - \frac{1}{\widehat{\xi}_l} \frac{\widehat{\boldsymbol{\Phi}}_{\mathbf{n},l}\mathbf{e}}{\mathbf{e}^{\mathsf{T}}\widehat{\boldsymbol{\Phi}}_{\mathbf{n},l}\mathbf{e}}} \tag{11}$$

3.2. Correlation Matrices

Since the $N \times N$-dimensional correlation matrices $\boldsymbol{\Phi}_{\mathbf{y},l}$ and $\boldsymbol{\Phi}_{\mathbf{n},l}$ can be assumed to be Hermitian positive semidefinite (PSD), each matrix consists of a total of N^2 real-valued coefficients, denoted by $\mathbf{h}_{\mathbf{y},l}$ and $\mathbf{h}_{\mathbf{n},l}$. As illustrated in Fig. 1, we propose to estimate these coefficients using separate DNNs $\mathbf{f}_{\mathbf{y}}^l$ and $\mathbf{f}_{\mathbf{n}}^l$ in state l, i.e.,

$$\begin{aligned} \widehat{\mathbf{h}}_{\mathbf{y},l} &= \mathbf{f}_{\mathbf{y}}^l\left(\mathbf{y}_{c,l}\right) \\ \widehat{\mathbf{h}}_{\mathbf{n},l} &= \mathbf{f}_{\mathbf{n}}^l\left(\mathbf{y}_{c,l}\right), \end{aligned} \tag{12}$$

where $\widehat{\circ}$ denotes an estimate of $\circ$, and $\mathbf{y}_{c,l} = [\Re Y_l, \Im Y_l]^{\mathsf{T}}$ denotes the vector of the real and imaginary parts of the noisy speech STFT coefficient Y_l. Since the coefficients $\mathbf{h}_{\mathbf{y},l}$ and $\mathbf{h}_{\mathbf{n},l}$ are unbounded, a linear activation is used for $\mathbf{f}_{\mathbf{y}}^l$ and $\mathbf{f}_{\mathbf{n}}^l$. The Hermitian PSD correlation matrix estimates are obtained as

$$\boxed{\begin{aligned} \widehat{\boldsymbol{\Phi}}_{\mathbf{y},l} &= \widehat{\mathbf{H}}_{\mathbf{y},l}\widehat{\mathbf{H}}_{\mathbf{y},l}^{\mathsf{H}}, \quad \widehat{\mathbf{H}}_{\mathbf{y},l} = \text{Hermitian}\left\{\widehat{\mathbf{h}}_{\mathbf{y},l}\right\} \\ \widehat{\boldsymbol{\Phi}}_{\mathbf{n},l} &= \widehat{\mathbf{H}}_{\mathbf{n},l}\widehat{\mathbf{H}}_{\mathbf{n},l}^{\mathsf{H}}, \quad \widehat{\mathbf{H}}_{\mathbf{n},l} = \text{Hermitian}\left\{\widehat{\mathbf{h}}_{\mathbf{n},l}\right\} \end{aligned}} \tag{13}$$

where Hermitian $\{\mathbf{h}\}$ assembles an $N \times N$-dimensional Hermitian matrix from the (real-valued) N^2-dimensional vector $\mathbf{h}$, and the matrix multiplication ensures that the correlation matrix estimates are Hermitian PSD. It should be noted that no correlation matrix labels are used in the training process — instead, the DNNs are trained to minimize the speech enhancement-related loss function at the MFMVDR filter output (see Section 4.4).

3.3. A-Priori SNR

Similarly to the approach described in the previous section, we propose to use a DNN f_ξ^l to map noisy speech features to an a-priori SNR estimate $\widehat{\xi}_l$, i.e.,

$$\boxed{\widehat{\xi}_l = f_\xi^l\left(\log_{10}|Y_l|\right)} \tag{14}$$

Since $\xi_l \geq 0$, a softmax activation is used for f_ξ^l. Similarly as for the correlation matrices, the DNN is trained to output an a-priori SNR estimate $\widehat{\xi}_l$ such that the speech enhancement-related loss function at the MFMVDR filter output is minimized (see Section 4.4).

4. EXPERIMENTAL RESULTS

In this section, the speech enhancement performance of the proposed deep MFMVDR filter is compared to several baseline deep learning-based speech enhancement algorithms (see Section 4.1). In Sections 4.2 - 4.4 we discuss the used dataset, DNN architecture, and algorithm settings. In Section 4.5 we present the results in terms of the perceptual evaluation of speech quality (PESQ) [23] and short-time objective intelligibility (STOI) [24] improvement.

4.1. Baseline Algorithms

As baseline single-microphone speech enhancement algorithms, we consider three deep learning-based algorithms:

1. *Masking*: in order to investigate the possible benefit of multi-frame filtering, i.e., $N > 1$, we also consider the complex-valued gain/mask in (2), i.e., $N = 1$:

$$W_{M,l} = f_M^l\left(\mathbf{y}_{c,l}\right) \in \mathbb{C};\ \Re W_{\mathrm{M},l}, \Im W_{\mathrm{M},l} \in [-2, 2], \tag{15}$$

   where the bounds for the real and imaginary parts are motivated by [25].

2. *Direct filtering*: in order to investigate the possible benefit of using the MFMVDR filter structure in (10), we also consider directly estimating the complex-valued coefficients of the N-dimensional multi-frame filter $\mathbf{w}_l$ in (4), similarly to [17]:

$$\mathbf{w}_{F,l} = \mathbf{f}_{\mathbf{F}}^l\left(\mathbf{y}_{c,l}\right) \in \mathbb{C}^N;\ \Re \mathbf{w}_{\mathrm{F},l}, \Im \mathbf{w}_{F,l} \in [-1, 1], \tag{16}$$

   where the bounds for the real and imaginary parts are motivated by [17].

3. *Conv-TasNet* [11]: instead of considering the STFT-domain as the transform-domain for speech enhancement, Conv-TasNet uses learnable transformations and applies a real-valued mask in the transform-domain. For a fair comparison with the other considered algorithms, we considered the causal version of Conv-TasNet [11].

4.2. Dataset

All considered algorithms were trained and evaluated on the DNS Challenge dataset [22]. In total, this dataset contains more than 500 h of speech from 2150 speakers and 180 h of noise from 150 different noise classes at a sampling frequency of 16 kHz. For training and validation, we randomly selected a subset of 45000 utterances of length 4 s, with SNRs uniformly sampled from [0, 20] dB. Using a validation split of 20 %, this resulted in 40 h for training and 10 h for validation, respectively. Evaluation was performed on the DNS Challenge synthetic test set without reverberation. This test set is disjoint from the training and validation set and includes 20 speakers, 12 VoIP-relevant noise sources, and SNRs uniformly sampled from [0, 25] dB, in total consisting of 150 utterances of length 10 s.

4.3. DNN Architecture

As the DNN architecture for all estimators, we used temporal convolutional networks (TCNs) [20][1], which have been demonstrated to exhibit strong temporal and spectral modeling capabilities [11]. Without performing extensive hyperparameter optimization, we fixed the hyperparameters of all TCN modules (except for the Conv-TasNet baseline, for which we used the hyperparameters proposed in [11]) to 2 stacks of 4 layers each, with a kernel size of 3, resulting in a temporal receptive field of 128 ms. Aiming at a fair comparison, the number of hidden dimensions was varied to obtain a similar total number of trainable weights for all considered algorithms (cf. Table 1). Note that this hyperparameter was varied as opposed to the number of stacks/layers or the kernel size, since increasing the temporal receptive field might give an unfair advantage.

[1] We used the implementation provided by the authors of [11], available at https://github.com/naplab/Conv-TasNet.

algorithm	hidden dimension	trainable weights
masking	226	5.0 M
direct filtering	225	5.1 M
Conv-TasNet	128	5.0 M
deep MFMVDR	128	5.3 M

Table 1. TCN module hyperparameters.

algorithm	ΔPESQ / MOS	ΔSTOI	real-time factor
masking	0.65	0.037	**0.068**
direct filtering	0.67	0.038	0.070
Conv-TasNet	0.67	0.041	0.194
deep MFMVDR	**0.76**	**0.042**	0.176

Table 2. PESQ and STOI improvement as well as real-time factors obtained on the DNS Challenge synthetic test set without reverberation, averaged over all utterances.

4.4. Algorithm Settings

In order to be able to exploit speech correlation, an STFT with high temporal resolution, i.e., a frame length of 8 ms and a frame shift of 2 ms, was employed for all STFT-based algorithms, similarly as in [13]. A Hann window was used both as analysis and synthesis window. The multi-frame algorithms (i.e., the proposed deep MFMVDR filter and direct filtering) use a filter length of $N = 5$, such that speech correlation within 16 ms can be exploited. To improve the numerical stability of the matrix inversion in (10), Tikhonov regularization with a regularization constant $\delta = 10^{-3}$ was used [13, 14]. Finally, a minimum gain of -17 dB was included in all algorithms except Conv-TasNet.

The TCNs were trained for 50 epochs using the Adam optimizer [26] 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 early in case the validation loss did not decrease for 10 consecutive epochs. The gradient norms were clipped to 5, and the batch size was set to 6 to maximize the usage of graphics card memory. As loss function, we used the negative scale-invariant signal-to-distortion ratio (SI-SDR) [21], i.e.,

$$\text{SI-SDR} = 10\log_{10}\left(\frac{|\alpha\tilde{\mathbf{x}}|^2}{|\alpha\tilde{\mathbf{x}} - \widehat{\tilde{\mathbf{x}}}|^2}\right), \quad \alpha = \frac{\widehat{\tilde{\mathbf{x}}}^{\mathsf{T}}\tilde{\mathbf{x}}}{||\tilde{\mathbf{x}}||^2}, \tag{17}$$

where $\tilde{\mathbf{x}}$ and $\widehat{\tilde{\mathbf{x}}}$ denote the speech signal and the estimated speech signal in the time-domain. All algorithms were implemented in PyTorch 1.6.0, and training and evaluation were performed on an NVIDIA GeForce® RTX 2080 Ti graphics card.

4.5. Results

For all considered algorithms, Table 2 shows the average improvement in terms of PESQ and STOI w.r.t. the noisy microphone signals using the speech signal as reference signal. As can be observed, all considered algorithms yield a significant PESQ and STOI improvement, where the proposed deep MFMVDR filter outperforms all other algorithms. A minor performance improvement can be observed between direct filtering ($N = 5$) and masking ($N = 1$), hinting at the potential of exploiting multiple frames. A much larger improvement can be observed between deep MFMVDR filtering and direct filtering, showing that exploiting the MFMVDR filter structure and guiding the TCN training to estimate the required quantities (correlation matrices and a-priori SNR) instead of directly estimating the filter coefficients is advantageous. Exemplary audio examples for all considered algorithms are available online[2].

As a measure for computational complexity, Table 2 also shows the average real-time factor (RTF) for all considered algorithms, defined as RTF = (processing time)/(signal length), processed using 4 cores of an Intel® Xeon® CPU clocked at 2.6 GHz. Although the proposed deep MFMVDR filter results in a larger computational complexity than directly estimating the filter coefficients, its computational complexity is similar to that of Conv-TasNet. A PyTorch implementation of the deep MFMVDR filter is available online[3].

5. CONCLUSION

In this paper we proposed a supervised learning-based approach to estimate the required parameters of an MFMVDR filter for single-microphone speech enhancement. Because the MFMVDR filter requires accurate estimates of the noisy speech and noise correlation matrices as well as the speech IFC vector, we proposed to utilize the temporal and spectral modeling capabilities of TCNs for this estimation task. The TCNs are trained to map the noisy speech STFT coefficients to the required parameters by minimizing the SI-SDR loss function at the output of the MFMVDR filter. Experiments on the DNS Challenge dataset demonstrate the benefits of (i) using a multi-frame algorithm as compared to a single-frame algorithm, and (ii) guiding the TCN training by using the MFMVDR filter structure instead of directly estimating the filter coefficients.

6. REFERENCES

[1] P. Vary and R. Martin, *Digital speech transmission: enhancement, coding and error concealment*, John Wiley, Chichester, England; Hoboken, NJ, 2006.

[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*, Morgan & Claypool Publishers, 2013.

[3] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, Springer Science & Business Media, 2011.

[4] 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.

[5] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*, John Wiley & Sons, 2018.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[7] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[2] https://uol.de/en/mediphysics-acoustics/sigproc/research/audio-demos

[3] https://uol.de/en/mediphysics-acoustics/sigproc/research/code-examples

[8] Y. Xu, J. Du, L. Dai, and C. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[9] Y. Wang, A. Narayanan, and D. L. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[10] M. Kolbæk, Z. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network Based Speech Enhancement Systems," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[11] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[12] K. Tan, J. Chen, and D. Wang, "Gated Residual Networks With Dilated Convolutions for Monaural Speech Enhancement," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 189–198, Jan. 2019.

[13] 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.

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

[15] 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.

[16] D. Fischer and S. Doclo, "Subspace-Based Speech Correlation Vector Estimation for Single-Microphone Multi-Frame MVDR Filtering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 856–860.

[17] W. Mack and E. A. P. Habets, "Deep Filtering: Signal Extraction and Reconstruction Using Complex Time-Frequency Filters," *IEEE Signal Processing Letters*, vol. 27, Nov. 2019.

[18] Y. Xu, M. Yu, S.-X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu, "Neural Spatio-Temporal Beamformer for Target Speech Separation," *arXiv:2005.03889 [cs, eess]*, May 2020.

[19] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 191–195.

[20] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv:1803.01271 [cs]*, Mar. 2018.

[21] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 626–630.

[22] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," *arXiv:2005.13981 [cs, eess]*, May 2020.

[23] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862," Tech. Rep., International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[24] 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, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[25] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and Friends: Leveraging Discrete Representations for Source Separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 370–382, May 2019.

[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014.