DNN-BASED SPEECH PRESENCE PROBABILITY ESTIMATION FOR MULTI-FRAME SINGLE-MICROPHONE SPEECH ENHANCEMENT

Marvin Tammen, Dörte Fischer, Bernd T. Meyer, Simon Doclo

Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4all
University of Oldenburg, Germany

ABSTRACT

Multi-frame approaches for single-microphone speech enhancement, e.g., the multi-frame minimum-power-distortionless-response (MFMPDR) filter, are able to exploit speech correlations across neighboring time frames. In contrast to single-frame approaches such as the Wiener gain, it has been shown that multi-frame approaches achieve a substantial noise reduction with hardly any speech distortion, provided that an accurate estimate of the correlation matrices and especially the speech interframe correlation (IFC) vector is available. Typical estimation procedures of the IFC vector require an estimate of the speech presence probability (SPP) in each time-frequency (TF) bin. In this paper, we propose to use a bi-directional long short-term memory deep neural network (DNN) to estimate the SPP for each TF bin. Aiming at achieving a robust performance, the DNN is trained for various noise types and within a large signal-to-noise-ratio range. Experimental results show that the MFMPDR in combination with the proposed data-driven SPP estimator yields an increased speech quality compared to a state-of-the-art model-based SPP estimator. Furthermore, it is confirmed that exploiting interframe correlations in the MFMPDR is beneficial when compared to the Wiener gain especially in adverse scenarios.

***Index Terms*—** Speech Presence Probability, Deep Neural Network, Single-Microphone Speech Enhancement, Multi-Frame Filtering

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. Hence, several single- and multi-microphone speech enhancement approaches have been proposed [1, 2, 3, 4, 5]. Typical single-microphone speech enhancement approaches apply a real-valued spectro-temporal gain, e.g., the Wiener gain (WG) [1], to the noisy short-time Fourier transform (STFT) coefficients to obtain an estimate of the clean speech signal. A disadvantage of these methods is that stronger noise reduction typically goes hand-in-hand with increased speech distortion.

In contrast to these single-frame approaches, multi-frame approaches [6, 7, 8, 9, 10] apply a complex-valued filter to the noisy STFT coefficients and are able to take into account the speech correlation across consecutive time frames. Similarly to the minimum-variance-distortionless-response (MVDR) beamformer and the minimum-power-distortionless-response beamformer (MPDR) for multi-microphone speech enhancement [4, 11], a multi-frame MPDR (MFMPDR) filter has been proposed for single-microphone speech enhancement [6, 7, 10]. This multi-frame filter requires an estimate of the noisy correlation matrix and the speech interframe correlation (IFC) vector in each time-frequency (TF) bin. When oracle estimates of these quantities are available, it has been shown in [6, 12] that the MFMPDR filter achieves a good noise reduction and hardly any speech distortion in contrast to the WG. However, it has also been shown that the speech enhancement performance is very sensitive to estimation errors of the highly time-varying speech IFC vector [12].

In [7] a maximum likelihood (ML)-based approach has been proposed to estimate the speech IFC vector from the noisy microphone signals. The ML estimator typically requires an estimate of the speech presence probability (SPP) in each TF bin. Several model-based SPP estimators have been proposed [13, 14, 15, 16] based on the assumption that the speech and noise STFT coefficients are uncorrelated, complex Gaussian distributed random variables. These estimators, however, have difficulties with accurately estimating the SPP in the short STFT frames that are required to capture the highly time-varying speech IFC vector.

In recent years, data-driven supervised learning-based approaches have gained a lot of attention in a multitude of applications, including single-microphone speech enhancement [17, 18, 19, 20, 21, 22]. A common approach is to estimate real-valued TF masks, which are applied to the noisy STFT coefficients. Furthermore, mask-based approaches have been recently proposed to estimate the speech and noise correlation matrices that are required by multi-microphone speech enhancement approaches such as the MVDR beamformer or the generalized eigenvalue beamformer [23, 24].

Inspired by the approach in [23], in this paper we propose to use a data-driven SPP to estimate the required speech IFC vector for the MFMPDR filter. More in particular, we use a bidirectional long short-term memory (BLSTM) [25] deep neural network (DNN) to estimate the SPP in each TF bin given the noisy STFT coefficients. Aiming at achieving a robust performance, the DNN is trained on the WSJ0 [26] and NOISEX92 [27] datasets using a signal-to-noise ratio (SNR) range from 0 to 20 dB. Experimental results for non-matched noise types and partially non-matched SNRs show that using the proposed DNN-based SPP estimate yields a larger speech quality improvement compared to the model-based SPP estimate [16]. Furthermore, when utilizing either of the SPP estimates to implement an MFMPDR or a WG, the benefit of exploiting speech IFCs is confirmed [7, 9].

2. SIGNAL MODEL

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

$$Y(k,l) = X(k,l) + N(k,l), \tag{1}$$

where $X(k,\ l)$ denotes the speech component and $N(k,\ l)$ denotes the noise component at the k-th frequency bin and the l-th time frame. Multi-frame speech enhancement approaches [6, 7, 8, 10] estimate the speech component by applying a finite impulse response filter with N taps

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

to the noisy STFT coefficients, i.e.,

$$\widehat{X}(k,l)=\sum_{n=0}^{N-1}H_n^*(k,l)Y(k,l-n), \tag{2}$$

where $\widehat{\circ}$ denotes an estimate of $\circ$, $H_n(k,l)$ denotes the n-th filter coefficient, and * denotes the complex-conjugate operator. Using vector notation, (1) and (2) can be written as

$$\mathbf{y}(k,l)=\mathbf{x}(k,l)+\mathbf{n}(k,l) \tag{3}$$

$$\widehat{X}(k,l)=\mathbf{h}^H(k,l)\mathbf{y}(k,l), \tag{4}$$

where ^{H} denotes the Hermitian operator and the N-dimensional vectors $\mathbf{h}(k,l)$ and $\mathbf{y}(k,l)$ contain the filter coefficients and N consecutive STFT coefficients, i.e.,

$$\mathbf{h}(k,l)=[H_0(k,l),\,H_1(k,l),\,...,\,H_{N-1}(k,l)]^T, \tag{5}$$

$$\mathbf{y}(k,l)=[Y(k,l),\,Y(k,l-1),\,...,\,Y(k,l-N+1)]^T. \tag{6}$$

This is analogous to multi-microphone beamforming approaches [4, 5, 11] by considering the FIR filter as a spatial filter and frames as microphone inputs. Since all frequency bins are treated individually, in the remainder of this paper we omit the frequency index k.

Assuming that the speech and noise components are uncorrelated, the noisy correlation matrix $\mathbf{\Phi}_{\boldsymbol{y}}(l)=\mathcal{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\}$, with $\mathcal{E}\{\circ\}$ the expectation operator, can be written as

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

with the speech and noise correlation matrices $\mathbf{\Phi}_{\boldsymbol{x}}(l)=\mathcal{E}\{\mathbf{x}(l)\mathbf{x}^H(l)\}$ and $\mathbf{\Phi}_{\boldsymbol{n}}(l)=\mathcal{E}\{\mathbf{n}(l)\mathbf{n}^H(l)\}$. In [6], it has been proposed to exploit the speech correlation across consecutive time frames by separating the speech component into a correlated and an uncorrelated part, i.e.,

$$\mathbf{x}(l)=\underbrace{\boldsymbol{\gamma}_{\boldsymbol{x}}(l)X(l)}_{\text{correlated}}+\underbrace{\mathbf{x}'(l)}_{\text{uncorrelated}}, \tag{8}$$

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

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

with the vector $\mathbf{e}$ selecting the first column of $\mathbf{\Phi}_{\mathbf{x}}(l)$ and $\mathbf{e}^T\mathbf{\Phi}_{\mathbf{x}}(l)\mathbf{e}=\phi_X(l)=\mathcal{E}\{|X(l)|^2\}$ the speech power spectral density (PSD). Note that since $X(l)$ is fully correlated with itself, the first element of the speech IFC vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(l)$ in (9) is equal to 1, such that the first element of the uncorrelated speech vector $\mathbf{x}'(l)$ is equal to 0. Substituting (8) in (3), we obtain the multi-frame signal model

$$\mathbf{y}(l)=\boldsymbol{\gamma}_{\boldsymbol{x}}(l)X(l)+\mathbf{x}'(l)+\mathbf{n}(l), \tag{10}$$

where the uncorrelated speech component $\mathbf{x}'$ is treated as an interference.

Similarly to the speech IFC vector in (9), the noisy IFC vector and the noise IFC vector can be defined as

$$\boldsymbol{\gamma}_{\boldsymbol{y}}(l)=\frac{\mathbf{\Phi}_{\mathbf{y}}(l)\mathbf{e}}{\mathbf{e}^T\mathbf{\Phi}_{\mathbf{y}}(l)\mathbf{e}},\quad \boldsymbol{\gamma}_{\boldsymbol{n}}(l)=\frac{\mathbf{\Phi}_{\mathbf{n}}(l)\mathbf{e}}{\mathbf{e}^T\mathbf{\Phi}_{\mathbf{n}}(l)\mathbf{e}}, \tag{11}$$

with $\mathbf{e}^T\mathbf{\Phi}_{\mathbf{y}}(l)\mathbf{e}=\mathcal{E}\{|Y(l)|^2\}$ and $\mathbf{e}^T\mathbf{\Phi}_{\mathbf{n}}(l)\mathbf{e}=\phi_N(l)=\mathcal{E}\{|N(l)|^2\}$ denoting the noisy and noise PSDs, respectively. Using (11) in (7), the speech IFC vector $\boldsymbol{\gamma}_{\mathbf{x}}(l)$ can be obtained as

$$\boldsymbol{\gamma}_{\mathbf{x}}(l)=\frac{1+\xi(l)}{\xi(l)}\boldsymbol{\gamma}_{\boldsymbol{y}}(l)-\frac{1}{\xi(l)}\boldsymbol{\gamma}_{\boldsymbol{n}}(l), \tag{12}$$

with the a-priori SNR $\xi(l)=\frac{\phi_X(l)}{\phi_N(l)}$.

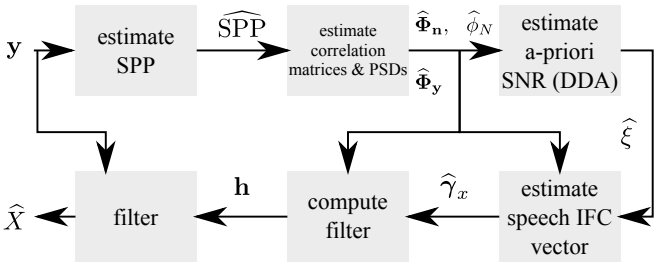

Fig. 1: Diagram of parameter estimation and multi-frame filtering.

3. MULTI-FRAME MPDR FILTER

In [6], the MFMPDR filter for single-microphone speech enhancement was proposed, which aims at minimizing the output PSD while preserving the correlated speech component. The corresponding constrained optimization problem is given by

$$\min_{\mathbf{h}(l)\in\mathbb{C}^N}\quad \mathbf{h}^H(l)\mathbf{\Phi}_{\mathbf{y}}(l)\mathbf{h}(l),\quad \text{s.t. } \mathbf{h}^H(l)\boldsymbol{\gamma}_{\boldsymbol{x}}(l)=1. \tag{13}$$

Solving this problem, the filter vector is equal to [6]

$$\boxed{\mathbf{h}_{\text{MFMPDR}}=\frac{\mathbf{\Phi}_{\mathbf{y}}^{-1}(l)\boldsymbol{\gamma}_{\boldsymbol{x}}(l)}{\boldsymbol{\gamma}_{\mathbf{x}}^H(l)\mathbf{\Phi}_{\mathbf{y}}^{-1}(l)\boldsymbol{\gamma}_{\boldsymbol{x}}(l)}} \tag{14}$$

4. PARAMETER ESTIMATION

In practice, the performance of the MFMPDR filter depends on how well the time-varying correlation matrix $\mathbf{\Phi}_{\boldsymbol{y}}(l)$ as well as the highly time-varying speech IFC vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(l)$ can be estimated from the noisy microphone signals. In [12] it has been shown that the performance of the MFMPDR filter is very sensitive to estimation errors of the speech IFC vector. Whereas estimating the noisy correlation matrix $\mathbf{\Phi}_{\mathbf{y}}(l)$ is rather straightforward, accurately estimating the speech IFC vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(l)$ is not so trivial [7, 8, 10, 12]. Typically, this vector requires an estimate of the a-priori SNR $\xi(l)$ and the noise correlation matrix $\mathbf{\Phi}_{\mathbf{n}}(l)$, which in turn require an estimate of the SPP in each TF bin [7]. The following subsections discuss the estimation of the noisy and noise correlation matrices, the speech IFC vector, as well as the a-priori SNR using either a state-of-the-art model-based SPP estimator or the proposed DNN-based SPP estimator. Fig. 1 depicts the parameter estimation and multi-frame filtering process.

4.1. Correlation Matrices Estimation

The noisy correlation matrix $\mathbf{\Phi}_{\mathbf{y}}(l)$ is estimated using recursive smoothing with smoothing constant λ_y, i.e.,

$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}(l)=\lambda_y\widehat{\mathbf{\Phi}}_{\mathbf{y}}(l-1)+(1-\lambda_y)\mathbf{y}(l)\mathbf{y}^H(l). \tag{15}$$

To estimate the noise correlation matrix $\mathbf{\Phi}_{\boldsymbol{n}}(l)$, similarly to [28] we apply a recursive smoothing procedure to the noisy microphone signals, where the smoothing factor for each TF bin depends on a time-varying SPP estimate $\widehat{\text{SPP}}(l)$ and a smoothing constant α_n, i.e.,

$$\widehat{\mathbf{\Phi}}_{\mathbf{n}}(l)=\lambda_n(l)\widehat{\mathbf{\Phi}}_{\mathbf{n}}(l-1)+(1-\lambda_n(l))\mathbf{y}(l)\mathbf{y}^H(l) \tag{16}$$

$$\lambda_n(l)=\alpha_n+(1-\alpha_n)\widehat{\text{SPP}}(l). \tag{17}$$

In the limiting cases, we have

$$\begin{cases}\widehat{\text{SPP}}(l)=0\Rightarrow\lambda_n(l)=\alpha_n & (18)\\ \widehat{\text{SPP}}(l)=1\Rightarrow\lambda_n(l)=1\Rightarrow\widehat{\mathbf{\Phi}}_{\mathbf{n}}(l)=\widehat{\mathbf{\Phi}}_{\mathbf{n}}(l-1). & (19)\end{cases}$$

We consider two approaches to estimate the SPP for each TF bin required in (17). As the reference, denoted with subscript $\circ_{\mathrm{R}}$, we use the model-based approach from [16], which assumes that the speech and noise STFT coefficients are complex Gaussian distributed. Using this assumption, likelihood functions for speech presence and speech absence can be derived, yielding the SPP estimate

$$\boxed{\widehat{\mathrm{SPP}}_{\mathrm{R}}(l)=\left(1+\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}(1+\xi_{\mathcal{H}_1})e^{-\frac{|Y(l)|^2}{\widehat{\phi}_N(l-1)}\frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)^{-1}} \tag{20}$$

where $P(\mathcal{H}_1)$ and $P(\mathcal{H}_0)$ denote the prior probability of speech presence and absence, respectively, and the parameter $\xi_{\mathcal{H}_1}$ denotes a typical a-priori SNR encountered during speech presence. Note that this method relies on the noise PSD estimate of the *previous* frame $\widehat{\phi}_N(l-1)=\mathbf{e}^T\widehat{\mathbf{\Phi}}_{\mathbf{n}}(l-1)\mathbf{e}$.

Alternatively, in this paper we propose to exploit the capabilities of a BLSTM DNN to capture temporal and spectral structures in order to estimate the SPP. The DNN is trained to perform a mapping between the noisy STFT coefficient magnitudes and the SPP, i.e.,

$$\boxed{\widehat{\mathrm{SPP}}_{\mathrm{DNN}}(l)=f_{\mathbf{\Theta}}\{|\mathbf{Y}|\}(l)} \tag{21}$$

with $|\mathbf{Y}|\in\mathbb{R}^{K\times L}$ containing all K frequency bins and L time frames of the noisy STFT coefficient magnitudes of the considered signal, $f_{\mathbf{\Theta}}$ the trained DNN with parameters $\mathbf{\Theta}$, and $\widehat{\mathrm{SPP}}_{\mathrm{DNN}}(l)$ the DNN-based SPP estimate. The training process is detailed in Sec. 5.

4.2. Speech IFC Vector Estimation

Similarly to (12), the ML-based approach in [7] estimates the speech IFC vector as

$$\boxed{\widehat{\boldsymbol{\gamma}}^{\boldsymbol{\mu}}_{\mathbf{x}}(l)=\frac{1+\widehat{\xi}(l)}{\widehat{\xi}(l)}\widehat{\boldsymbol{\gamma}}_{\mathbf{y}}(l)-\frac{1}{\widehat{\xi}(l)}\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}} \tag{22}$$

where $\widehat{\xi}(l)$ is an estimate of the a-priori SNR and $\widehat{\boldsymbol{\gamma}}_{\mathbf{y}}(l)$ is an estimate of the noisy IFC vector obtained similarly as in (11) using $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(l)$ from (15). The fixed mean noise IFC vector $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ can be computed based on the analysis window and overlap settings [7].

Alternatively, by replacing the fixed mean noise IFC vector by a TF-varying noise IFC vector estimate $\widehat{\boldsymbol{\gamma}}_{\mathbf{n}}(l)$, the speech IFC vector can be computed as

$$\boxed{\widehat{\boldsymbol{\gamma}}^{\boldsymbol{\gamma}}_{\mathbf{x}}(l)=\frac{1+\widehat{\xi}(l)}{\widehat{\xi}(l)}\widehat{\boldsymbol{\gamma}}_{\mathbf{y}}(l)-\frac{1}{\widehat{\xi}(l)}\widehat{\boldsymbol{\gamma}}_{\mathbf{n}}(l)} \tag{23}$$

where $\widehat{\boldsymbol{\gamma}}_{\mathbf{n}}(l)$ is obtained similarly to (11) using $\widehat{\mathbf{\Phi}}_{\mathbf{n}}(l)$ from (16).

To estimate the a-priori SNR $\xi(l)$, we apply the well-known decision-directed approach (DDA) [29], i.e.,

$$\widehat{\xi}(l)=\lambda_{\mathrm{DDA}}\frac{\widehat{X}(l-1)}{\widehat{\phi}_N(l-1)}+(1-\lambda_{\mathrm{DDA}})\frac{|Y(l)|^2}{\widehat{\phi}_N(l-1)}, \tag{24}$$

with weighting constant λ_{DDA} and $\widehat{X}(l-1)$ denoting the speech estimate of the previous frame.

5. DNN TRAINING PROCESS

As described in (21), the DNN is trained to map the input features, i.e., the noisy STFT coefficient magnitudes, to the SPP. More specifically, we train the DNN with the target defined as

$$\mathrm{SPP}_{\mathrm{DNN}}(l)=\left(1+\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}(1+\xi_{\mathcal{H}_1})e^{-\frac{|Y(l)|^2}{\phi_N(l)}\frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)^{-1}. \tag{25}$$

For this target, we compute the noise PSD $\phi_N(l)$ via recursive averaging of the noise component, which is available during training, i.e.,

$$\phi_N(l)=\alpha_n\phi_N(l-1)+(1-\alpha_n)|N(l)|^2. \tag{26}$$

As loss function, we use the mean-squared difference between the target SPP defined in (25) and the estimated SPP $\widetilde{\mathrm{SPP}}_{\mathrm{DNN}}(k,l)$, i.e.,

$$\frac{1}{LK}\sum_{l=0}^{L-1}\sum_{k=0}^{K-1}\left(\widetilde{\mathrm{SPP}}_{\mathrm{DNN}}(k,l)-\mathrm{SPP}_{\mathrm{DNN}}(k,l)\right)^2, \tag{27}$$

where $\widetilde{\mathrm{SPP}}_{\mathrm{DNN}}(k,l)=f_{\tilde{\mathbf{\Theta}}}\{|\mathbf{Y}|\}(k,l)$ uses the current set of parameters $\tilde{\mathbf{\Theta}}$. The DNN is composed of an input layer with 33 input nodes, a hidden BLSTM layer with 256 nodes for each direction, two hidden fully-connected layers with 513 nodes each, and an output layer with 33 nodes. The corresponding activation functions of the hidden and output layers are $\tanh$, rectifying linear unit (ReLU), ReLU, and $\mathrm{sigmoid}$, respectively, inherently restricting the SPP estimates to $]0,\,1[$. This network architecture is inspired by the DNN used in [23] and has been tested for various sets of hyperparameters.

The network weights are initialized using a uniform distribution $U(-a,\,a)$, with $a=\sqrt{6/(n_{\mathrm{in}}+n_{\mathrm{out}})}$, and n_{in} and n_{out} the number of input and output neurons of the layer, respectively [30]. All bias values are initialized with 0. To decrease the dynamic range of the input data and to stabilize the training process, we apply batch normalization to the input and before the activations of the hidden layers [31]. To optimize the network parameters, the Adam optimizer is utilized with parameters as proposed in [32], with the learning rate set to 10^{-3} and the smoothing parameters for the gradient and the squared gradient set to 0.9 and 0.999, respectively. If the l^2-norm of a gradient is larger than 1, the gradient is divided by this norm.

To evaluate the model performance, we make use of a separate validation set as described in Sec. 6.1. The training is stopped either after 100 epochs or after the validation loss as measured by (27) has not decreased for 5 epochs. The DNN is implemented in PyTorch 1.2.0 [33], and training and evaluation are performed on a multi-GPU system utilizing 3 NVIDIA GeForce® GTX 1080 Ti graphics cards.

6. EXPERIMENTAL RESULTS

In this section, we compare the speech enhancement performance of the MFMPDR filter in (14) using

1. to estimate the SPP required in (17): either the model-based SPP estimator $\widehat{\mathrm{SPP}}_{\mathrm{R}}$ in (20) or the proposed DNN-based estimator $\widehat{\mathrm{SPP}}_{\mathrm{DNN}}$ in (21).
2. to estimate the speech IFC vector: either the fixed mean noise IFC vector $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ in (22) or the estimated time-varying noise IFC vector $\widehat{\boldsymbol{\gamma}}_{\mathbf{n}}(l)$ in (23).

In addition, to investigate the impact of exploiting speech IFCs, we also use the SPP estimators in (20) and (21) in a (single-frame) Wiener gain (WG), resulting in a total of 6 compared methods.

SNR / dB	-5	0	5	10	15	20
$\text{MFMPDR}_{\text{R},\mu}$	0.09	0.27	0.31	0.31	0.27	0.24
$\text{MFMPDR}_{\text{DNN},\mu}$	**0.22**	**0.33**	**0.41**	**0.35**	**0.33**	0.22
$\text{MFMPDR}_{\text{R},\gamma}$	-0.01	0.15	0.24	0.29	0.29	0.24
$\text{MFMPDR}_{\text{DNN},\gamma}$	0.04	0.21	0.29	0.32	0.29	0.21
WG_{R}	0.04	0.13	0.18	0.23	0.26	**0.28**
WG_{DNN}	0.06	0.16	0.20	0.24	0.27	0.23

Table 1: PESQ / MOS improvements vs. input SNR / dB, averaged over all evaluation set utterances and noise types.

6.1. Dataset

As clean speech material, we have used the training, development, and test sets of the WSJ0 corpus [26] for training, model validation, and evaluation, respectively. The noisy microphone signals have been generated by adding scaled (randomly chosen) noise segments to the clean speech signals at a sampling frequency of 16 kHz. Regarding noise, we have used the NOISEX92 database [27] for training and the Aurora database [34] for evaluation, resulting in a strong mismatch between training and evaluation conditions in order to evaluate the generalization capability of the proposed method. For each training utterance, the corresponding broadband SNR has been uniformly sampled from $[0, 20]\,\text{dB}$. For evaluation, 4 random utterances from the WSJ0 test set have been used at broadband SNRs $\in \{-5,\ 0,\ 5,\ 10,\ 15,\ 20\}\,\text{dB}$ for each of the 8 noise types in the Aurora database [34]. In total, this results in 12776, 2348, and 192 utterances for training, validation, and evaluation, respectively.

6.2. Simulation Settings

Since the speech IFC vector is highly time-varying, we employ an STFT with a high temporal resolution, i.e., a frame length of 4 ms and a frame shift of 1 ms, similarly as in [6, 7, 9, 10]. A Hann window is used for both STFT analysis and synthesis. The parameters of both the model-based SPP estimator $\widehat{\text{SPP}}_{\text{R}}$ in (20) and the DNN-based SPP estimator $\widehat{\text{SPP}}_{\text{DNN}}$ in (21) are set as proposed in [16], i.e., $P(\mathcal{H}_1) = P(\mathcal{H}_0) = 0.5$ and $\xi_{\mathcal{H}_1} = 15\,\text{dB}$. As recursive smoothing constants, we use $\alpha_n = 0.98$, $\lambda_y = 0.92$, and $\lambda_{\text{DDA}} = 0.97$. The MFMPDR filters use a filter length of $N = 18$, such that correlations within a window of 21 ms can be exploited. To be more comparable to the MFMPDR filters, the WG methods are used with the same settings, except for $N = 1$. To improve numerical stability when inverting a matrix, we perform regularization using diagonal loading as in [6, 7] with regularization parameter $\delta = 10^{-3}$. Finally, all compared methods use a minimum gain of -17 dB.

6.3. Results

For the 6 considered methods, Tab. 1 depicts the improvements in terms of the perceptual evaluation of speech quality (PESQ) [35] measure w.r.t. the noisy microphone signals as a function of the input SNR. The clean speech signal has been used as the reference signal. Subscripts denote which SPP estimator was used and, in the case of the MFMPDR filters, whether the mean IFC noise vector $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ or the time-varying noise IFC vector $\boldsymbol{\gamma}_{\boldsymbol{n}}(l)$ was utilized. The presented values are averaged over all utterances and noise types included in the evaluation set.

First, it can be observed that the MFMPDR filter utilizing the proposed DNN-based SPP estimate $\widehat{\text{SPP}}_{\text{DNN}}(l)$ and the fixed mean noise IFC vector $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ ($\text{MFMPDR}_{\text{DNN},\mu}$), yields the highest PESQ improvements for all input SNRs except 20 dB. Second, comparing the methods utilizing either the model-based SPP estimate $\widehat{\text{SPP}}_{\text{R}}(l)$ or the DNN-based SPP estimate $\widehat{\text{SPP}}_{\text{DNN}}(l)$, the advantages of using the DNN-based estimator are evident. This may be explained by the fact that, in contrast to the model-based estimator, the DNN can exploit spectral structures of speech and noise. Third, contrasting the MFMPDR filters and the WG, it can be confirmed that exploiting speech IFCs may yield higher speech quality improvements than directly using the SPP estimate in a WG approach [7, 9] (except for an input SNR of 20 dB). The difference between the MFMPDR-based methods and the WG-based methods increases for lower input SNRs, suggesting that exploiting the speech IFCs is especially helpful in adverse scenarios. Fourth, using the fixed mean noise IFC vector in the MFMPDR filters consistently leads to larger PESQ improvements than using the estimated time-varying noise IFC vector. Considering the results in [7, 36], this suggests that a filter bank with higher frequency resolution is required to effectively incorporate an estimate of the time-varying noise IFC vector into the MFMPDR filter.

7. CONCLUSION

In this paper we considered a DNN-based SPP estimator for multi-frame approaches in single-microphone speech enhancement. Since the MFMPDR filter requires accurate estimates of the time-varying noisy correlation matrix and especially the speech IFC vector, in this paper we propose to use a DNN to improve the estimation of the speech IFC vector. The DNN is trained to map noisy STFT coefficient magnitudes to an SPP on a database comprising multiple noise types within a large SNR range to improve the generalization capability of the DNN. We demonstrate a higher objective speech quality improvement when using the proposed DNN-based SPP estimator instead of a state-of-the-art model-based estimator. Furthermore, by comparing the MFMPDR filters with Wiener gains based on equal SPP estimates, we confirm that utilizing interframe correlations can be beneficial especially in adverse scenarios.

8. REFERENCES

[1] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, John Wiley & Sons, 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] Jacob Benesty, Jingdong Chen, and Emanuël AP 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. A. Huang and J. Benesty, "A Multi-Frame Approach to the Frequency-Domain Single-Channel Noise Reduction Problem," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1256–1269, May 2012.

[7] 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, Sept. 2014.

[8] K. T. Andersen and M. Moonen, "Robust Speech-Distortion Weighted Interframe Wiener Filters for Single-Channel Noise Reduction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 97–107, Jan. 2018.

[9] D. Fischer, S. Doclo, E. A. P. Habets, and T. Gerkmann, "Combined Single-Microphone Wiener and MVDR Filtering based on Speech Interframe Correlations and Speech Presence Probability," in *Proc. ITG Conference on Speech Communication*, Paderborn, Germany, 2016, pp. 292–296.

[10] D. Fischer and S. Doclo, "Robust Constrained MFMVDR Filtering for Single-Microphone Speech Enhancement," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sept. 2018, pp. 41–45.

[11] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[12] 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.

[13] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[14] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved *A Posteriori* Speech Presence Probability Estimation Based on a Likelihood Ratio With Fixed Priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 910–919, July 2008.

[15] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian Model-Based Multichannel Speech Presence Probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, July 2010.

[16] T. Gerkmann and R. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[17] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. 2005, pp. 181–197, Springer.

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

[19] D. Williamson, Y. Wang, and D. L. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[20] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[21] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and Björn Schuller, "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," in *Latent Variable Analysis and Signal Separation*. 2015, pp. 91–99, Springer.

[22] S. E. Chazan, J. Goldberger, and S. Gannot, "Deep recurrent mixture of experts for speech enhancement," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, Oct. 2017, pp. 359–363.

[23] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 196–200.

[24] W. Jiang, F. Wen, and P. Liu, "Robust Beamforming for Speech Recognition Using DNN-based Time-Frequency Masks Estimation," *IEEE Access*, vol. 6, pp. 52385 – 52392, Sept. 2018.

[25] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, July 2005.

[26] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *Speech and Natural Language: Proceedings of a Workshop*, New York, USA, Feb. 1992.

[27] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[28] M. Souden, J. Chen, J. Benesty, and S. Affes, "An Integrated Solution for Online Multichannel Noise Tracking and Reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2159–2169, Sept. 2011.

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

[30] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, May 2010, pp. 249–256.

[31] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]*, Feb. 2015, arXiv: 1502.03167.

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

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. Neural Information Processing Systems (NIPS) Workshop*, Long Beach, USA, Dec. 2017.

[34] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[35] 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*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[36] D. Fischer, K. Brümann, and S. Doclo, "Comparison of Parameter Estimation Methods for Single-Microphone Multi-Frame Wiener Filtering," in *Proc. European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, Sept. 2019, pp. 1809–1813.