DNN BASED MULTIFRAME SINGLE-CHANNEL NOISE REDUCTION FILTERS

Ningning Pan[1], *Jingdong Chen*[1], *and Jacob Benesty*[2]

[1]CIAIC and Shaanxi Provincial Key Laboratory of Artificial Intelligence, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada

ABSTRACT

While multiframe noise reduction filters, e.g., the multiframe Wiener and minimum variance distortionless response (MVDR) ones, have demonstrated great potential to improve both the subband and fullband signal-to-noise ratios (SNRs) by exploiting explicitly the interframe speech correlation, the implementation of such filters requires the knowledge of the interframe correlation coefficients for every subband, which are challenging to estimate in practice. In this work, we present a deep neural network (DNN) based method to estimate the interframe correlation coefficients and the estimated coefficients are subsequently fed into multiframe filters to achieve noise reduction. Unlike existing DNN based methods, which outputs the enhanced speech directly, the presented method combines deep learning and traditional methods, which gives more flexibility to optimize or tune noise reduction performance. Experimental results are presented to justify the properties of the presented methods.

Index Terms—Single-channel noise reduction, interframe correlation, multiframe Wiener filter, multiframe MVDR filter, DNN.

1. INTRODUCTION

Single-channel noise reduction is a problem of recovering a clean speech signal of interest from its microphone observation, which is corrupted by additive noise [1–7]. The goal of noise reduction may vary from one application to another but, generally, it is to improve speech quality, intelligibility, and/or performance of some subsequent speech processors. A large number of methods have been developed and the typical ones include optimal filtering in the time and frequency domains [6,8,9], spectral subtraction techniques [2,10,11], statistical approaches [12–14], subspace methods [15,16], and deep learning based approaches [17–20]. While they are able to reduce noise thereby improving the signal-to-noise (SNR) ratio, those methods were found to introduce speech distortion. The amount of this distortion is generally proportional to the amount of noise reduction. Consequently, how to control the speech distortion has become a paramount issue in noise reduction. A single-channel minimum variance distortionless response (MVDR) filter, which exploits the interframe speech correlation to ensure the target speech component undistorted was developed in [21]. And the multiframe Wiener filter is proposed afterwards [22], which achieves more noise reduction without introducing extra speech distortion than the gain based Wiener filter. However, implementation of the multiframe Wiener and MVDR filters requires to know the interframe correlation coefficients for every subband, which are difficult to obtain in practice.

In this work, we use a deep neural network (DNN) based method to estimate the interframe correlation coefficients. Specifically, the noisy speech signals are transformed into the short-time Fourier transform (STFT) domain. The interframe correlation coefficients are then estimated through a DNN. These estimated correlation coefficients are subsequently used to construct the single-channel multiframe Wiener and MVDR filters [21, 22]. Experiments are performed to evaluate the performance of the presented method and compare them with two state-of-the-art deep learning based methods.

2. SIGNAL MODEL AND PROBLEM FORMULATION

The noise reduction problem considered in this paper is one of recovering the desired signal of interest (clean speech) from its noise corrupted (microphone) observation:

$$y(t) = x(t) + v(t), \tag{1}$$

where t is the time index, and $x(t)$ and $v(t)$ are, respectively, the clean speech signal of interest and the unwanted additive noise, which are assumed to be uncorrelated. All the signals are assumed to be of zero mean.

Transforming all the signals in (1) through STFT gives the signal model in the time-frequency domain, i.e.,

$$Y(k,m) = X(k,m) + V(k,m), \tag{2}$$

where $Y(k,m)$, $X(k,m)$, and $V(k,m)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, with $k \in \{0, 1, \ldots, K-1\}$ and m denoting, respectively, the frequency-bin and time-frame indices. Since $x(t)$ and $v(t)$ are assumed to be uncorrelated, the variance of $Y(k,m)$ is then

$$\phi_Y(k,m) \triangleq E\left[|Y(k,m)|^2\right] = \phi_X(k,m) + \phi_V(k,m),$$

where $E[\cdot]$ denotes mathematical expectation, and $\phi_X(k,m)$ and $\phi_V(k,m)$ are the variances of $X(k,m)$ and $V(k,m)$, respectively.

Since speech signals are correlated across time frames [21,22], we consider the most recent L time-frames and form the following observation signal vector of length L:

$$\begin{aligned}\mathbf{y}(k,m) &\triangleq \left[\begin{array}{ccc} Y(k,m) & \cdots & Y(k,m-L+1) \end{array}\right]^T \\ &= \mathbf{x}(k,m) + \mathbf{v}(k,m),\end{aligned} \tag{3}$$

where the superscript ^{T} is the transpose operator, and $\mathbf{x}(k,m)$ and $\mathbf{v}(k,m)$ are defined similarly to $\mathbf{y}(k,m)$. Then the covariance matrix of $\mathbf{y}(k,m)$ is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,m) &\triangleq E\left[\mathbf{y}(k,m)\mathbf{y}^H(k,m)\right] \\ &= \mathbf{\Phi}_{\mathbf{x}}(k,m) + \mathbf{\Phi}_{\mathbf{v}}(k,m),\end{aligned} \tag{4}$$

where the superscript ^{H} is the conjugate-transpose operator, and $\mathbf{\Phi}_{\mathbf{x}}(k,m)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,m)$ are the covariance matrices of $\mathbf{x}(k,m)$ and $\mathbf{v}(k,m)$, respectively.

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102200, in part by the Key Program of National Science Foundation of China (NSFC) under Grant 61831019 and 62192713.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

Now, the problem of noise reduction can be formulated as one of estimating $X(k,m)$ given the observation vector $\mathbf{y}(k,m)$, which is a mixture of $\mathbf{x}(k,m)$ and $\mathbf{v}(k,m)$. The noise vector, i.e., $\mathbf{v}(k,m)$, only interferes the estimation. In contrast, the signal vector $\mathbf{x}(k,m)$ contains both the desired signal (first component) and the components $X(k,m-l)$, $l \neq 0$, which are partially correlated with $X(k,m)$. To see this clearly, let us decompose $X(k,m-l)$ into two orthogonal components, i.e.,

$$X(k,m-l) = \gamma_X(k,m,l)X(k,m) + X_{\mathrm{i}}(k,m-l), \tag{5}$$

where

$$X_{\mathrm{i}}(k,m-l) = X(k,m-l) - \gamma_X(k,m,l)X(k,m), \tag{6}$$

$$E\left[X^*(k,m)X_{\mathrm{i}}(k,m-l)\right] = 0, \tag{7}$$

and

$$\gamma_X(k,m,l) = \frac{E\left[X^*(k,m)X(k,m-l)\right]}{\phi_X(k,m)} \tag{8}$$

is the interframe correlation coefficient of the signal $X(k,m)$, with the superscript * being the complex-conjugate operator. Since it is uncorrelated with $X(k,m)$, the component $X_{\mathrm{i}}(k,m-l)$ interferes with the estimation. As a result, we call this component interference as indicated by the subscript $_{\mathrm{i}}$.

Then, we can write the vector $\mathbf{x}(k,m)$ as

$$\begin{aligned}\mathbf{x}(k,m) &= X(k,m)\boldsymbol{\gamma}_X(k,m) + \mathbf{x}_{\mathrm{i}}(k,m)\\ &= \mathbf{x}_{\mathrm{d}}(k,m) + \mathbf{x}_{\mathrm{i}}(k,m),\end{aligned} \tag{9}$$

where $\mathbf{x}_{\mathrm{d}}(k,m) = X(k,m)\boldsymbol{\gamma}_X(k,m)$ is the desired signal vector,

$$\mathbf{x}_{\mathrm{i}}(k,m) = \left[\begin{array}{ccc} X_{\mathrm{i}}(k,m) & \cdots & X_{\mathrm{i}}(k,m-L+1) \end{array}\right]^T$$

is the interference signal vector, and

$$\begin{aligned}\boldsymbol{\gamma}_X(k,m) &= \left[\begin{array}{ccc} \gamma_X(k,m,0) & \cdots & \gamma_X(k,m,L-1) \end{array}\right]^T\\ &= \frac{E\left[X^*(k,m)\mathbf{x}(k,m)\right]}{\phi_X(k,m)}\end{aligned} \tag{10}$$

is the (normalized) interframe correlation vector. Therefore, the signal model in (3) can be expressed as the sum of three mutually uncorrelated signal vectors, i.e.,

$$\mathbf{y}(k,m) = X(k,m)\boldsymbol{\gamma}_X(k,m) + \mathbf{x}_{\mathrm{i}}(k,m) + \mathbf{v}(k,m), \tag{11}$$

whose covariance matrix is

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{y}}(k,m) &= \phi_X(k,m)\boldsymbol{\gamma}_X(k,m)\boldsymbol{\gamma}_X^H(k,m)\\ &\quad + \boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i}}}(k,m) + \boldsymbol{\Phi}_{\mathbf{v}}(k,m),\end{aligned} \tag{12}$$

where $\boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i}}}(k,m)$ is the covariance matrix of $\mathbf{x}_{\mathrm{i}}(k,m)$.

3. FILTER DESIGN

Now, let $\mathbf{h}(k,m)$ be a complex-valued linear filter of length L. Applying this filter to the observation signal vector, we obtain an estimate of $X(k,m)$, i.e.,

$$\begin{aligned}\widehat{X}(k,m) &= \mathbf{h}^H(k,m)\mathbf{y}(k,m)\\ &= X_{\mathrm{f}}(k,m) + V_{\mathrm{rn}}(k,m),\end{aligned} \tag{13}$$

where $\widehat{X}(k,m)$ is an estimate of $X(k,m)$, $X_{\mathrm{f}}(k,m) = \mathbf{h}^H(k,m)\mathbf{x}(k,m)$ is the filtered version of the desired signal at L consecutive frames, and $V_{\mathrm{rn}}(k,m) = \mathbf{h}^H(k,m)\mathbf{v}(k,m)$ is the residual noise. Substituting the decomposition in (11) into (13), we get

$$\widehat{X}(k,m) = X_{\mathrm{fd}}(k,m) + X_{\mathrm{ri}}(k,m) + V_{\mathrm{rn}}(k,m), \tag{14}$$

where $X_{\mathrm{fd}}(k,m) = X(k,m)\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X(k,m)$ is the filtered desired signal, $X_{\mathrm{ri}}(k,m) = \mathbf{h}^H(k,m)\mathbf{x}_{\mathrm{i}}(k,m)$ denotes the residual interference.

3.1. Multiframe Wiener Filter

The mean square error (MSE) between $\hat{X}(k,m)$ and $X(k,m)$ is defined as

$$J\left[\mathbf{h}(k,m)\right] = E\left[\left|\mathbf{h}^H(k,m)\mathbf{y}(k,m) - X(k,m)\right|^2\right]. \tag{15}$$

Minimizing $J\left[\mathbf{h}(k,m)\right]$ with respect to $\mathbf{h}(k,m)$ gives the multiframe Wiener filter [22]:

$$\mathbf{h}_{\mathrm{W}}(k,m) = \boldsymbol{\Phi}_{\mathbf{y}}^{-1}(k,m)\boldsymbol{\Phi}_{\mathbf{x}}(k,m)\mathbf{i}_1, \tag{16}$$

where $\mathbf{i}_1$ is the first column of an $L \times L$ identity matrix.

3.2. Multiframe MVDR Filter

To estimate $X(k,m)$ without introducing distortion, the following distortionless constraint is needed

$$\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X(k,m) = 1. \tag{17}$$

Then, the problem becomes one of minimizing the variance of $\widehat{X}(k,m)$, i.e., $\phi_{\widehat{X}}(k,m) = \mathbf{h}^H(k,m)\boldsymbol{\Phi}_{\mathbf{y}}(k,m)\mathbf{h}(k,m)$, subject to the constraint in (17). Mathematically, it is written as

$$\arg\min_{\mathbf{h}(k,m)} \phi_{\widehat{X}}(k,m) \text{ subject to } \mathbf{h}^H(k,m)\boldsymbol{\gamma}_X(k,m) = 1, \tag{18}$$

whose solution is the single-channel MVDR filter [6], [21]:

$$\mathbf{h}_{\mathrm{MVDR}}(k,m) = \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(k,m)\boldsymbol{\gamma}_X(k,m)}{\boldsymbol{\gamma}_X^H(k,m)\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(k,m)\boldsymbol{\gamma}_X(k,m)}. \tag{19}$$

4. PERFORMANCE MEASURES

The following performance measures are used to analyze and evaluate different noise reduction filters.

- Fullband input SNR. It is defined according to the signal model in (2) as

$$\mathrm{iSNR}(m) \triangleq \frac{\sum_{k=0}^{K-1}\phi_X(k,m)}{\sum_{k=0}^{K-1}\phi_V(k,m)}. \tag{20}$$

- Fullband output SNR. At time frame m, the fullband output SNR is defined according to the filtering model in (11) as

$$\mathrm{oSNR}[\mathbf{h}(m)] \triangleq \frac{\sum_{k=0}^{K-1}\phi_X(k,m)\left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X(k,m)\right|^2}{\sum_{k=0}^{K-1}\mathbf{h}^H(k,m)\boldsymbol{\Phi}_{\mathrm{in}}(k,m)\mathbf{h}(k,m)}, \tag{21}$$

where $\boldsymbol{\Phi}_{\mathrm{in}}(k,m) = \boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i}}}(k,m) + \boldsymbol{\Phi}_{\mathbf{v}}(k,m)$.

- Fullband speech-distortion index. It is defined as

$$\upsilon_{\mathrm{sd}}[\mathbf{h}(m)] = \frac{\sum_{k=0}^{K-1}E\left\{\left|X_{\mathrm{fd}}(k,m) - X(k,m)\right|^2\right\}}{\sum_{k=0}^{K-1}\phi_X(k,m)}. \tag{22}$$

The speech-distortion index is always greater than or equal to 0 and should not exceed 1. The higher is its value, the more distorted is the speech signal. For the MVDR filter, $\upsilon_{\mathrm{sd}}[\mathbf{h}_{\mathrm{MVDR}}(m)] = 0$.

5. STATISTICS ESTIMATION BY DNN

In this section, we introduce a DNN based method to estimate the needed statistics, $\boldsymbol{\Phi}_{\mathbf{x}}(k,m)$, for implementing the multiframe Wiener and MVDR filters. The DNN framework is illustrated in Fig. 1, which is a fusion of the fullband and subband models, denoted as FullSubNet [23]. We introduce the network in three parts, i.e., the input of the network, learning target, and model structure.

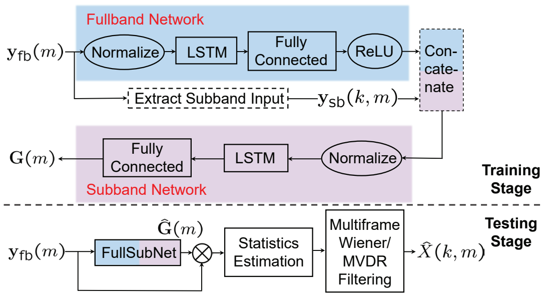

Fig. 1. Framework of DNN estimation.

5.1. Input

Let us first define a fullband spectral vector, $\mathbf{y}_{\mathrm{fb}}(m)$, and a subband spectral vector, $\mathbf{y}_{\mathrm{sb}}(k,m)$, as follows:

$$\mathbf{y}_{\mathrm{fb}}(m) \triangleq \begin{bmatrix} |Y(0,m)| & |Y(1,m)| & \cdots & |Y(K-1,m)| \end{bmatrix}^T$$

is the fulband spectral magnitude of the noisy observation in the mth frame and

$$\mathbf{y}_{\mathrm{sb}}(k,m) \triangleq \\ \begin{bmatrix} |Y(k-P,m)| & \cdots & |Y(k,m)| & \cdots & |Y(k+P,m)| \end{bmatrix}^T$$

is the subband spectral magnitude vector, which takes the adjacent P frequency bins centered in (k,m) from both sides in order to exploit the interband information. For frequency bins $k < P$ and $k > K-1-P$, circular Fourier frequencies are used to pad $\mathbf{y}_{\mathrm{sb}}(k,m)$.

The fullband network takes $\mathbf{y}_{\mathrm{fb}}(m)$ as its input and outputs a fullband feature in the mth frame, denoted as $f[\mathbf{y}_{\mathrm{fb}}(m)] \triangleq \begin{bmatrix} Y_{\mathrm{fb}}(0,m) & \cdots & Y_{\mathrm{fb}}(K-1,m) \end{bmatrix}^T$, by exploiting the fullband information, where $f[\cdot]$ denotes all operations in the fullband network. While the subband network takes the concatenation of $\mathbf{y}_{\mathrm{sb}}(k,m)$ and the corresponding fullband feature $Y_{\mathrm{fb}}(k,m)$ as input, denoted as $\begin{bmatrix} \mathbf{y}_{\mathrm{sb}}(k,m) \\ Y_{\mathrm{fb}}(k,m) \end{bmatrix}$. All frequency bins share the same subband network.

5.2. Target

According to the literature of noise reduction, there exists a complex ratio $G(k,m)$, which makes

$$X(k,m) = G(k,m)Y(k,m). \tag{23}$$

Then, (23) can be rewritten as

$$\begin{aligned} & X_{\mathrm{re}}(k,m) + \jmath X_{\mathrm{im}}(k,m) \\ & = \left[G_{\mathrm{re}}(k,m) + \jmath G_{\mathrm{im}}(k,m)\right]\left[Y_{\mathrm{re}}(k,m) + \jmath Y_{\mathrm{im}}(k,m)\right], \end{aligned} \tag{24}$$

where $\jmath$ is the imaginary unit, the subscripts $_{\mathrm{re}}$ and $_{\mathrm{im}}$ indicate the real and imaginary components, respectively, and $G(k,m) = G_{\mathrm{re}}(k,m) + \jmath G_{\mathrm{im}}(k,m)$ is the complex ideal ratio mask (cIRM) [24]. By solving the equation in (24), the real and imaginary components of the complex ratio are

$$G_{\mathrm{re}}(k,m) = \frac{Y_{\mathrm{re}}(k,m)X_{\mathrm{re}}(k,m) + Y_{\mathrm{im}}(k,m)X_{\mathrm{im}}(k,m)}{Y_{\mathrm{re}}^2(k,m) + Y_{\mathrm{im}}^2(k,m)}, \tag{25}$$

$$G_{\mathrm{im}}(k,m) = \frac{Y_{\mathrm{re}}(k,m)X_{\mathrm{im}}(k,m) - Y_{\mathrm{im}}(k,m)X_{\mathrm{re}}(k,m)}{Y_{\mathrm{re}}^2(k,m) + Y_{\mathrm{im}}^2(k,m)}. \tag{26}$$

The training target of the network in the mth frame is $\mathbf{G}(m) = \begin{bmatrix} \mathbf{q}(0,m) & \cdots & \mathbf{q}(k,m) & \cdots & \mathbf{q}(K-1,m) \end{bmatrix}$, where $\mathbf{q}(k,m) = \begin{bmatrix} G_{\mathrm{re}}(k,m) \\ G_{\mathrm{im}}(k,m) \end{bmatrix}$.

5.3. Model Structure

We adopt the same model structure for the fullband and subband networks, which consist of 2 stacked unidirectional long-short-term-memory (LSTM) layers and a fully connected layer. LSTM of the fullband network has 512 hidden units, while the subband network has 384 units.

5.4. Statistics Estimation

We denote $\widehat{G}(k,m) \triangleq \widehat{G}_{\mathrm{re}}(k,m) + \jmath\widehat{G}_{\mathrm{im}}(k,m)$ and $\widehat{\mathbf{g}}(k,m) \triangleq \begin{bmatrix} \widehat{G}(k,m) & \cdots & \widehat{G}(k,m-L+1) \end{bmatrix}^T$, which are estimated by the aforementioned networks. The estimate of $\mathbf{\Phi}_{\mathbf{x}}(k,m)$ is then obtained as

$$\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,m) = \widehat{\mathbf{g}}(k,m)\widehat{\mathbf{g}}^H(k,m) \odot \mathbf{\Phi}_{\mathbf{y}}(k,m), \tag{27}$$

where $\odot$ denotes the Hadamard product. Then we get the estimate of $\boldsymbol{\gamma}_X(k,m)$ as

$$\widehat{\boldsymbol{\gamma}}_X(k,m) = \widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,m)/\widehat{\phi}_X(k,m)\mathbf{i}_1, \tag{28}$$

where $\widehat{\phi}_X(k,m)$ is the first element of $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,m)$.

6. EXPERIMENTAL RESULTS

In this section, we study through experiments the single-channel multiframe Wiener and MVDR filters with DNN statistics estimation and compare them with two state-of-the-art deep learning based methods, i.e., FullSubNet and deep MFMVDR [25].

6.1. Experimental Settings

Training and evaluation speech signals are from different dataset. All signals are resampled to 16 kHz. The FFT length is 8 ms. The overlap-add method is adopted for signal reconstruction with an overlap factor of 75%.

6.1.1. Training Configurations for FullSubNet

We use the DNS challenge dataset [26] to retrain the network. The clean speech dataset consists of more than 1000 hours clean speech signals, including English and non-English languages. There are 181 hours of noise data in the DNS noise dataset, which includes 60000 clips belonging to 150 audio classes. The training pairs are generated by dynamically mixing noise and speech with a random SNR between -10 dB and 30 dB at a step of 1 dB. Before adding noise, the speech level is randomly normalized to the range between -35 dB and -15 dB at a step of 1 dB.

The learning rate is set to 10^{-3} for the first epoch, and decreased to be half if the loss in validation set does not decrease in the next 3 consecutive epoches. ADAM is used as the optimizer.

6.1.2. Filters Implementation

Besides the statistics $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,m)$ and $\widehat{\boldsymbol{\gamma}}_X(k,m)$ obtained from (27) and (28), we also need to know the $\mathbf{\Phi}_{\mathbf{y}}(k,m)$ matrix to implement the studied filters, which is estimated as follows. We use the first 100 frames to initialize $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m)$ by averaging with a batch method. Then, $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m)$ is recursively updated according to

$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m) = \alpha_y\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m-1) + (1-\alpha_y)\mathbf{y}(k,m)\mathbf{y}^H(k,m), \tag{29}$$

where $\alpha_y \in (0,1)$ is a forgetting factor. The inverse of the $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m)$ matrix is then computed according to

$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}^{-1}(k,m) = \left\{ \widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m) + \frac{\delta \cdot \mathrm{tr}[|\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m)|]}{L} \mathbf{I}_{L\times L} \right\}^{-1}, \tag{30}$$

where $\delta > 0$ is a regularization factor which is empirically chosen as 0.01, $\mathrm{tr}[\cdot]$ denote the trace of a square matrix and $\mathbf{I}_{L\times L}$ is an $L \times L$ identity matrix.

6.1.3. *Evaluation*

We take 20 speech signals (from 10 female speakers and 10 male speakers) from the WSJ0 dataset. WSJ0 is not included in the training set for evaluation. The white noise and babble noise from the NoiseX92 database [27] are added to the 20 clean speech signals at an SNR of 10 dB separately, leading to 40 noisy speech signals for evaluation.

The fullband SNR defined in (21) and the fullband speech-distortion index in (22) are used to evaluate the MVDR filter. While for the Wiener filter, the conventional definitions of these two measures [6, 8] are used. Reasons for adopting different sets of performance metrics for the MVDR and Wiener filters can be found in [22]. Besides, the mean opinion score of perceptual evaluation of speech quality (PESQ-MOS) is used for evaluation of all the studied algorithms. For FullSubNet, along with PESQ-MOS, we compute the speech-distortion index between the FullSubNet's output signal and the clean speech. As for the deep MFMVDR, the open source code[1] is adopted to train model for comparison. Note that the signal amplitude filtered by deep MFMVDR are changed nonlinearly, which makes it difficult to compute the fullband SNR and speech distortion. So, only PESQ-MOS is used to evaluate deep MFMVDR.

6.2. Experimental Results

Figure 2 plots the fullband output SNR, the fullband speech-distortion index, and PESQ-MOS of the Wiener and MVDR filters (with $L = 5,\ 10$), all as a function of the forgetting factor α_y. For comparison, the results from two deep learning based methods, i.e., FullSubNet and deep MFMVDR, are also presented. One can see that oSNR, υ_{sd}, and PESQ-MOS of the Wiener filter are larger than those of the MVDR filter. The speech-distortion index of the MVDR filter is around -87 dB, which clearly indicates that the speech distortionless constraint is satisfied with the proposed DNN based statistics estimation, regardless of the filter length L, which is achieved at the expense of less noise reduction. In comparison, PESQ-MOS of the multiframe Wiener filter is higher than that of the deep MFMVDR while marginally less than that of FullSubNet. Nevertheless, postprocessing with multiframe Wiener filter can help control the compromise between speech distortion and noise reduction. The multiframe MVDR filter falls short of expectation in PESQ-MOS than other methods due to the small amount of noise reduction. Linear constraints could be added in filter design to achieve more noise reduction, e.g., linearly constrained minimum variance (LCMV) filter [6], which could also be fulfilled by the proposed framework, which will not be discussed due to space limitation.

Figure 3 plots the time-domain filtered desired signal $x_{\mathrm{fd}}(t)$ of the proposed and deep MFMVDR methods. Figure 3(a) plots a zoomed segment of the signals (in the red box) in Fig. 3(b). All the signals were normalized to the same level. For the proposed method (dashed blue), the parameters are: $L = 5$ and $\alpha_y = 0.4$. It is seen from Fig. 3(a) that the proposed method is able to maintain the speech distortionless constraint but the deep MFMVDR method (green line) is unable to do this.

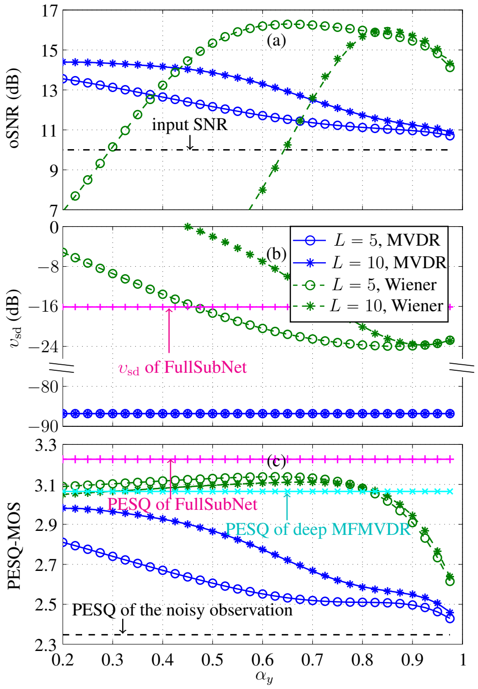

Fig. 2. Performance of different algorithms as a function of the forgetting factor α_y: (a) fullband output SNR, (b) fullband speech-distortion index, and (c) PESQ-MOS.

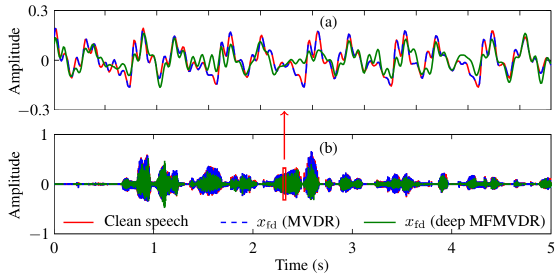

Fig. 3. The time-domain speech signals: the ground truth clean speech signal (red line), the filtered desired speech signal of the proposed MVDR (dashed blue line), and the deep MFMVDR (green line).

7. CONCLUSIONS

This paper studied the problem of single-channel noise reduction, which was formulated as one of multiple-frame filtering in the STFT domain. We presented a DNN based method with an LSTM structure to estimate the interframe correlation coefficients. The estimate coefficients were then fed to the single-channel multiframe Wiener and MVDR filters to achieve noise reduction. Experimental results showed that, with the experimental setup, the performances of the DNN based multiframe Wiener and MVDR filters are comparable to those of FullSubNet and deep MFMVDR; but the presented multiframe filters offer the flexibility to control the tradeoff between the amount of noise reduction and the level of speech distortion if needed.

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

8. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] M. Brandstein and Eds. D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Berlin, Germany: Springer-Verlag, 2001.

[4] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Berlin, Germany: Springer-Verlag, 2005.

[5] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of Noise Reduction," in *Springer Handbook on Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., pp. 843–871. Berlin, Germany: Springer-Verlag, 2007.

[6] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*, Berlin, Germany: Springer-Verlag, 2011.

[7] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second edition*, Boca Raton Florida: CRC Press, 2013.

[8] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[9] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters: a Theoretical Study*, Berlin, Germany: Springer-Verlag, 2011.

[10] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1770–1779, Aug. 2011.

[11] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 7, pp. 2080–2094, Apr. 2012.

[12] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

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

[14] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *ELSEVIER Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[15] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[16] N. Pan, J. Benesty, and J. Chen, "On single-channel noise reduction with rank-deficient noise correlation matrix," *ELSEVIER Appl. Acoust.*, vol. 126, pp. 26–35, Nov. 2017.

[17] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2020, pp. 46–50.

[20] X. Wang, J. Du, A. Cristia, L. Sun, and C. Lee, "A study of child speech extraction using joint speech enhancement and separation in realistic conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2020, pp. 7304–7308.

[21] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jul. 2011, pp. 273–276.

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

[23] X. Hao, X. Su, R. Horaud, and X. Li, "FullSubNet: A full-band and sub-band fusion model for real-time single channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2021, pp. 6618–6622.

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

[25] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2021, pp. 8428–8432.

[26] C. K. A Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2021, pp. 6608–6612.

[27] A. Varga and H. J. M. 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, Jul. 1993