SINGLE-MICROPHONE SPEECH ENHANCEMENT USING MVDR FILTERING AND WIENER POST-FILTERING

Dörte Fischer and Timo Gerkmann

Speech Signal Processing Group, Department of Medical Physics and Acoustics,
Cluster of Excellence Hearing4all, University of Oldenburg, Germany
{doerte.fischer, timo.gerkmann}@uni-oldenburg.de

ABSTRACT

For single-microphone noise reduction, a minimum variance distortionless response (MVDR) filter has been proposed recently. This filter takes the speech correlations of consecutive time frames into account and achieves impressive results in terms of speech distortions even in a blind implementation where we only have access to the noisy speech signal. However, compared to conventional approaches less noise reduction is achieved. Therefore, we propose to combine the single-microphone MVDR with a Wiener post-filter as the minimum-mean-square error optimal solution when multiple time frames are considered. We propose to pre-train the required interframe coherence matrices of the interferences for a large database, while speech correlations and interference power spectral densities are estimated online. In an experimental study based on instrumental measures, the proposed approach achieves a good trade-off between a single-channel Wiener filter and a multi-frame MVDR.

Index Terms— Speech enhancement, interframe correlation, MVDR, post-filter, Wiener filter

1. INTRODUCTION

Most speech communication systems like mobile phones or hearing aids are affected by ambient noise. The quality and intelligibility of speech decreases, especially at low signal-to-noise ratios (SNR). Noise reduction algorithms intend to suppress such additive noise. Generally, in a single-microphone scenario, noise reduction algorithms can improve the perceptual quality of speech while improving intelligibility remains difficult. This is because noise reduction is often accompanied by distortions rendering processed speech less intelligible [1].

Single-channel noise reduction algorithms are often formulated in the time-frequency domain based on the short-time Fourier transform (STFT). In order to obtain an estimate of the clean speech STFT coefficients, a multiplicative gain function is applied to the noisy speech signal at each time-frequency point. Examples are the single-channel Wiener filter (WF) gain [2], the minimum-mean-square error (MMSE) based amplitude estimator [3] and the MMSE log-amplitude estimator [4]. All these approaches assume that adjacent time frames are uncorrelated and that each time-frequency point can be processed independently. However, it is well known that speech is highly correlated over time and frequency. In the STFT domain, interframe correlations (IFC) result both from signal correlations and overlapping spectral analysis frames. While the Fourier transform generally reduces the correlation of signal samples, neighboring bands remain correlated [5]. Hence, it may be reasonable to also take spectral correlations into account. Here, we will focus on the IFC.

This work was supported by the DFG Cluster of Excellence EXC 1077/1 Hearing4all.

Benesty and Huang [6] [7] exploit the IFC and derived a multi-frame Wiener filter (MFWF) and a multi-frame MVDR (MFMVDR) finite impulse response (FIR) filter. For this, the current time frame of a signal as well as the previous frames are used. This frame array is similar to a multimicrophone system since each frame can be interpreted as a microphone input.

In [6] [7], the MFMVDR filter achieves an extraordinary performance with almost no speech distortions also because the authors evaluate their algorithms with an oracle estimate of the speech IFC. Schasse and Martin [8] derived a maximum likelihood (ML) estimator for the speech IFC and show that the instrumentally predicted speech quality and intelligibility can be improved also in a blind setup where only access to the noisy speech signal is available. However, it is known from microphone arrays that the MVDR filter often does not provide sufficient noise reduction [9] which is also the case for the MFMVDR with blindly estimated parameters. Inspired by [9], we add the WF as post-filter to the MFMVDR as this combination is MMSE optimal. We propose to pre-train the interframe coherence matrices of typical interferences offline for a wide range of noise and speech sources while speech correlations and interference power spectral densities are estimated online. The evaluation of the proposed MFMVDR+WF takes place in terms of PESQ [10] and STOI [11].

The paper is structured as follows. In the next sections we summarize the MFMVDR filter proposed in [6] [7] and derive the MFMVDR+WF. In Section 4 and 5, we evaluate the algorithms in a blind implementation given only the noisy speech signal. We conclude our work in Section 6.

2. MULTI-FRAME MVDR FILTER

In this section, we first summarize the interframe signal model and the MFMVDR filter presented in [6] [7].

We assume that a speech signal $X(k,m)$ is corrupted by an additive noise $V(k,m)$ in the STFT domain. The indexes k and m denote the frequency bin and time frame, respectively. It is assumed that the speech and noise processes are uncorrelated and that $X(k,m)$ and $V(k,m)$ are complex-valued, zero-mean random variables. The complex spectral noisy observation $Y(k,m)$ is thus given by

$$Y(k,m) = X(k,m) + V(k,m). \tag{1}$$

The estimate of the clean speech spectral component $X(k,m)$ is obtained by applying an FIR filter $H(k,m,l)$ of order $L-1$ to the noisy speech signal at each time-frequency point as

$$\hat{X}(k,m) = \sum_{l=0}^{L-1} H^*(k,m,l) Y(k,m-l) \tag{2}$$

$$= \mathbf{h}^H(k,m)\mathbf{y}(k,m). \tag{3}$$

Here, L is the number of consecutive time-frames, * indicates the complex-conjugate operator and ^{H} the hermitian operator. The vectors $\mathbf{h}(k,m)$ and $\mathbf{y}(k,m)$ contain the time-varying filter coefficients and the last $L-1$ noisy speech samples, respectively, i.e,

$$\mathbf{h}(k,m) = [H(k,m,0),\, H(k,m,1),\, ...,\, H(k,m,L-1)]^T, \quad (4)$$
$$\mathbf{y}(k,m) = [Y(k,m),\, Y(k,m-1),\, ...,\, Y(k,m-L+1)]^T.$$

The superscript ^{T} denotes the vector transpose. Note that a conventional multiplicative gain can be realized with $L=1$.

According to equation (1), the L-dimensional vector $\mathbf{y}(k,m)$ can be formulated by

$$\mathbf{y}(k,m) = \mathbf{x}(k,m) + \mathbf{v}(k,m), \quad (5)$$

where the clean speech vector $\mathbf{x}(k,m)$ and the noise vector $\mathbf{v}(k,m)$ are similarly defined as $\mathbf{y}(k,m)$ in (4). In order to take the speech IFC into account, the vector $\mathbf{x}(k,m)$ is decomposed into correlated and uncorrelated components with respect to the desired signal $X(k,m)$. Thus, equation (5) can be rewritten as

$$\mathbf{y}(k,m) = \gamma_{\mathbf{x}}(k,m)X(k,m) + \mathbf{x}'(k,m) + \mathbf{v}(k,m), \quad (6)$$
$$= \gamma_{\mathbf{x}}(k,m)X(k,m) + \mathbf{n}(k,m). \quad (7)$$

Here, the vector $\mathbf{x}'(k,m)$ represents the speech components uncorrelated to the local speech coefficients $X(k,m)$. Since we consider $\mathbf{x}'(k,m)$ as an interference, we replace $\mathbf{x}'(k,m)+\mathbf{v}(k,m)$ by $\mathbf{n}(k,m)$ as the undesired signal vector in (7). The term $\gamma_{\mathbf{x}}(k,m)X(k,m)$ indicates the correlated speech samples, with the speech IFC coefficient vector $\gamma_{\mathbf{x}}(k,m)$ defined by the normalized speech correlation vector $\mathbf{\Phi}_{\mathbf{x}X}(k,m) = E[\mathbf{x}(k,m)X^*(k,m)]$, as

$$\gamma_{\mathbf{x}}(k,m) = \frac{E[\mathbf{x}(k,m)X^*(k,m)]}{E[|X(k,m)|^2]} = \frac{\mathbf{\Phi}_{\mathbf{x}X}(k,m)}{\phi_X(k,m)}. \quad (8)$$

The operator $E[\cdot]$ denotes the expectation and $\phi_X = E[|X(k,m)|^2]$ the speech power spectral density (PSD). Due to the normalization, the first element of vector $\gamma_{\mathbf{x}}(k,m)$ is always 1 as $X(k,m)$ is obviously fully correlated with itself. Hence, the first element of the uncorrelated speech vector $\mathbf{x}'(k,m)$ is 0.

Based on the definition of (3) and (7) the MFMVDR filter [6] [7] is given by

$$\mathbf{h}_{\text{MFMVDR}}(k,m) = \frac{\mathbf{\Phi}_{\mathbf{yy}}^{-1}(k,m)\gamma_{\mathbf{x}}(k,m)}{\gamma_{\mathbf{x}}^{\mathbf{H}}(k,m)\mathbf{\Phi}_{\mathbf{yy}}^{-1}(k,m)\gamma_{\mathbf{x}}(k,m)}. \quad (9)$$

This filter minimizes the mean-squared error of the filtered undesired signal $\mathbf{n}(k,m)$ while the correlated speech components are not distorted. Here, $\mathbf{\Phi}_{\mathbf{yy}}(k,m) = E[\mathbf{y}(k,m)\mathbf{y}^H(k,m)]$ denotes the correlation matrix of the noisy speech $\mathbf{y}(k,m)$.

In analogy to multi-microphone algorithms, the clean speech IFC $\gamma_{\mathbf{x}}(k,m)$ acts as a steering vector. The main difference is that $\gamma_{\mathbf{x}}(k,m)$ needs to be determined for each time-frame m, while in multi-microphone speech enhancement the steering vector reflects the spatial location of the target and is typically more stationary.

3. MULTI-FRAME MVDR WITH POST-FILTER

Similar as for multi-microphone beamforming algorithms, the multi-frame MVDR filter often does not provide sufficient noise reduction in a blind implementation due to reverberation, diffuse noise and/or estimation errors [9]. Therefore, we propose to add a post-filter to the single-channel output of the MFMVDR in order to achieve more noise reduction while keeping speech distortions low. We show that, as in the multi-microphone case [9], the MMSE optimal post-filter is the single-channel WF. For a better readability, we omit the time and frequency indexes m and k for the derivation.

Minimizing the mean-square error between the desired signal $X(k,m)$ and the estimated desired signal $\hat{X}(k,m)$ leads to the MFWF [7], given by

$$\mathbf{h}_{\text{MFWF}} = \mathbf{\Phi}_{\mathbf{yy}}^{-1}\mathbf{\Phi}_{\mathbf{y}X}. \quad (10)$$

Here, $\mathbf{\Phi}_{\mathbf{y}X} = E[\mathbf{y}X^*]$ is the cross-correlation vector between the noisy speech vector $\mathbf{y}$ and the desired signal X. In contrast to the MFMVDR, the MFWF does allow for speech distortions.The resulting filter may lead to more noise reduction but also to more speech distortions compared to the MFMVDR.

Assuming that the desired speech and undesired signals are uncorrelated, the cross-correlation $\mathbf{\Phi}_{\mathbf{y}X}$ can be replaced by $\mathbf{\Phi}_{\mathbf{x}X}$. Since the speech IFC $\gamma_{\mathbf{x}}$ depends on the correlation vector $\mathbf{\Phi}_{\mathbf{x}X}$ in (8), the MFWF can be rewritten as

$$\mathbf{h}_{\text{MFWF}} = \phi_X \mathbf{\Phi}_{\mathbf{yy}}^{-1}\gamma_{\mathbf{x}}. \quad (11)$$

Further, we apply $\mathbf{\Phi}_{\mathbf{yy}} = \mathbf{\Phi}_{\mathbf{xx}} + \mathbf{\Phi}_{\mathbf{nn}}$ with $\mathbf{\Phi}_{\mathbf{xx}}$ and $\mathbf{\Phi}_{\mathbf{nn}}$ as the speech and undesired signal correlation matrices, as well as $\mathbf{\Phi}_{\mathbf{xx}} = \phi_X\gamma_{\mathbf{x}}\gamma_{\mathbf{x}}{}^H$. As a result, we obtain

$$\mathbf{h}_{\text{MFWF}} = \phi_X \left(\phi_X\gamma_{\mathbf{x}}\gamma_{\mathbf{x}}{}^H + \mathbf{\Phi}_{\mathbf{nn}}\right)^{-1}\gamma_{\mathbf{x}}. \quad (12)$$

Applying the Woodbury matrix identity $(A+UCD)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+DA^{-1}U\right)^{-1}DA^{-1}$ and after rearranging the terms, the final result is given by

$$\mathbf{h}_{\text{MFMVDR+WF}} = \underbrace{\frac{\mathbf{\Phi}_{\mathbf{yy}}^{-1}\gamma_{\mathbf{x}}}{\gamma_{\mathbf{x}}{}^H\mathbf{\Phi}_{\mathbf{yy}}^{-1}\gamma_{\mathbf{x}}}}_{\mathbf{h}_{\text{MFMVDR}}} \underbrace{\frac{\phi_X}{\phi_X + \left(\gamma_{\mathbf{x}}{}^H\mathbf{\Phi}_{\mathbf{nn}}{}^{-1}\gamma_{\mathbf{x}}\right)^{-1}}}_{h_{\text{WF}}}. \quad (13)$$

This shows that the MFWF as the MMSE solution can be factorized into the MFMVDR and the single-channel WF. The Wiener filter operates on the output of the MFMVDR, where $\left(\gamma_{\mathbf{x}}{}^H\mathbf{\Phi}_{\mathbf{nn}}{}^{-1}\gamma_{\mathbf{x}}\right)^{-1}$ denotes the undesired signal PSD $\phi_{N_{\text{out}}}$ at the output of the MFMVDR [2]. This quantity can be determined by filtering the undesired correlation matrix, i.e., $\phi_{N_{\text{out}}} = E[|\mathbf{h}_{\text{MFMVDR}}{}^H\mathbf{n}|^2] = \mathbf{h}_{\text{MFMVDR}}^H\mathbf{\Phi}_{\mathbf{nn}}\mathbf{h}_{\text{MFMVDR}}$. We obtain

$$\mathbf{h}_{\text{MFMVDR+WF}}(k,m) = \mathbf{h}_{\text{MFMVDR}}(k,m)\frac{\phi_X(k,m)}{\phi_{N_{\text{out}}}(k,m) + \phi_X(k,m)}. \quad (14)$$

The MFMVDR is designed to avoid speech distortions while the Wiener post-filter minimizes the mean-square error between the filtered single-channel output of the MFMVDR and the clean speech signal. Thus, the $\mathbf{h}_{\text{MFMVDR+WF}}(k,m)$ is capable to reduce the undesired signal components more strongly than the MFMVDR. However, speech distortions may be introduced, since the WF affects the speech signal as well. This effect can be mended by applying a lower limit as

$$\tilde{h}_{\text{WF}} = \max(h_{\text{min}}, h_{\text{WF}}). \quad (15)$$

4. PROPOSED PARAMETER ESTIMATION

We assume to have only access to the noisy speech signal such that we need to blindly estimate all required parameters.

Both the MFMVDR in (9) and the proposed MFMVDR+WF in (14) depend on the inverse of the noisy correlation matrix $\mathbf{\Phi}_{\mathbf{yy}}(k,m)$ and the clean-speech IFC $\gamma_{\mathbf{x}}(k,m)$. The quantity $\mathbf{\Phi}_{\mathbf{yy}}(k,m)$ is estimated by recursive smoothing, i.e.,

$$\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m) = \alpha\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m-1) + (1-\alpha)\mathbf{y}(k,m)\mathbf{y}^{\mathbf{H}}(k,m), \quad (16)$$

where the smoothing factor is experimentally set to $\alpha = 0.8$. The first element of the matrix $\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m)$ corresponds to the noisy speech PSD

$\hat{\phi}_Y(k,m)$, i.e., $\hat{\phi}_Y(k,m) = \left[\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m)\right]_{1,1}$. To compute the inverse of $\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m)$ we first perform a matrix regularization to improve the robustness of the filter computation as

$$\hat{\mathbf{\Phi}}_{\mathbf{yy}}^{-1}(k,m) = \left(\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m) + \frac{\delta_{\text{reg}}\text{tr}\left[\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m)\right]}{L}\mathbf{I}_{\mathbf{L}\times\mathbf{L}}\right)^{-1} \quad (17)$$

with a regularization parameter $\delta_{\text{reg}} = 0.04$ as in [8]. The operator tr$[\cdot]$ denotes the trace of a square matrix and $\mathbf{I}_{\mathbf{L}\times\mathbf{L}}$ is the identity matrix of size $L \times L$.

For the clean speech IFC we employ the proposed ML estimator for $\gamma_{\mathbf{x}}(k,m)$ in [8] based on the assumption that the noise and speech IFC vectors follow multivariate Gaussian distributions. This ML estimator is given by

$$\hat{\gamma}_{\mathbf{x}_{\mathbf{ML}}}(k,m) = \frac{\hat{\phi}_Y(k,m)}{\hat{\phi}_X(k,m)}\hat{\gamma}_{\mathbf{y}}(k,m) + \frac{\hat{\phi}_V(k,m)}{\hat{\phi}_X(k,m)}\mu_{\gamma_{\mathbf{v}}}(k,m). \quad (18)$$

Here, $\hat{\gamma}_{\mathbf{y}}(k,m)$ indicates the noisy IFC and is defined similar to the speech IFC in (8). Note that $\mathbf{\Phi}_{\mathbf{y}Y}(k,m) = \left[\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m)\right]_{:,1}$, where $[\cdot]_{:,1}$ denotes the first column of a matrix. The parameter $\mu_{\gamma_{\mathbf{v}}}(k,m)$ is the mean of the noise IFC. It is given by the frame overlap and the analysis window function [8] [12].

To estimate the speech PSD $\phi_X(k,m)$, it is assumed that $X(k,m)$ and $V(k,m)$ follow zero-mean Gaussian distributions. Thus, $\phi_X(k,m)$ can be estimated by the ML estimate for $\phi_X(k,m)$ [2] [8],

$$\hat{\phi}_X(k,m) = \max\left[\hat{\phi}_Y(k,m) - \hat{\phi}_V(k,m), 0\right]. \quad (19)$$

As in [8], the noise power $\phi_V(k,m)$ is obtained by the simple noise PSD estimator proposed in [13]

$$\hat{\phi}_V(k,m) = \min\left[\hat{\phi}_Y(k,m), \hat{\phi}_V(k,m-1)\right](1+\epsilon). \quad (20)$$

The parameter ϵ controls the maximum speed and is set to 5 dB/s as in [8].

In order to implement the proposed MFMVDR+WF, the power of the undesired signal at the output of the MFMVDR $\phi_{N_{\text{out}}}(k,m)$ is required. It is estimated by filtering the undesired correlation matrix $\mathbf{\Phi}_{\mathbf{nn}}(k,m)$, i.e.,

$$\phi_{N_{\text{out}}}(k,m) = \mathbf{h}_{\text{MFMVDR}}^H(k,m)\mathbf{\Phi}_{\mathbf{nn}}(k,m)\mathbf{h}_{\text{MFMVDR}}(k,m). \quad (21)$$

For this, we proposed to train the coherence $\mathbf{\Gamma}_{\mathbf{n}}(k,m) = \frac{\mathbf{\Phi}_{\mathbf{nn}}(k,m)}{[\mathbf{\Phi}_{\mathbf{nn}}(k,m)]_{1,1}}$ as the normalized IFC matrix of the undesired signal over a dataset. The matrix $\mathbf{\Phi}_{\mathbf{nn}}(k,m)$ is defined by the superposition of the correlation matrices of the noise $\mathbf{\Phi}_{\mathbf{vv}}(k,m)$ and the uncorrelated speech components $\mathbf{\Phi}_{\mathbf{x'x'}}(k,m)$, i.e., $\mathbf{\Phi}_{\mathbf{nn}} = \mathbf{\Phi}_{\mathbf{x'x'}} + \mathbf{\Phi}_{\mathbf{vv}}$. Since we know the speech and noise signals during the training perfectly, the interference signal can be calculated by

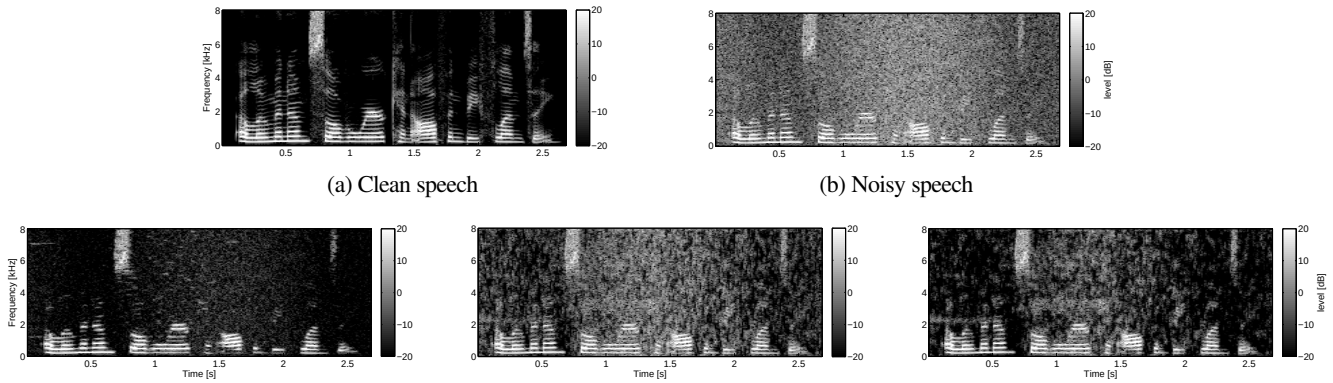
$$\mathbf{x}'(k,m) = \mathbf{x}(k,m) - \gamma_{\mathbf{x}}(k,m)X(k,m). \quad (22)$$

Thus, $\hat{\mathbf{\Phi}}_{\mathbf{x'x'}}(k,m)$ and $\hat{\mathbf{\Phi}}_{\mathbf{vv}}(k,m)$ can be estimated similar to $\hat{\mathbf{\Phi}}_{\mathbf{yy}}(k,m)$ in (16). We average the correlation matrices of the undesired signal over all data. To avoid scaling problems, the averaged $\mathbf{\Phi}_{\mathbf{nn}}(k,m)$ is normalized to the first element of the matrix and we obtain the coherence matrix $\mathbf{\Gamma}_{\mathbf{n}}(k,m)$. In the online processing, the trained $\mathbf{\Gamma}_{\mathbf{n}}(k,m)$ is applied and needs to be scaled by the signal power $\hat{\phi}_N(k,m) = \hat{\phi}_{X'}(k,m) + \hat{\phi}_V(k,m)$ to obtain an estimate of the current $\hat{\mathbf{\Phi}}_{\mathbf{nn}}(k,m)$. However, since the speech IFC in (8) implies that the first element of the vector is always 1, $[\mathbf{x}'(k,m)]_{1;1}$ in (22) is accordingly 0. As a result, the PSD $\phi_{X'}(k,m) = E\left[|[\mathbf{x}'(k,m)]_{1;1}|^2\right]$ is 0 as well and we can replace $\hat{\phi}_N(k,m)$ by $\hat{\phi}_V(k,m)$. Thus, the correlation matrix of the undesired signal vector is obtained by $\hat{\mathbf{\Phi}}_{\mathbf{nn}}(k,m) = \mathbf{\Gamma}_{\mathbf{n}}(k,m)\hat{\phi}_V(k,m)$ and $\phi_{N_{\text{out}}}(k,m)$ can be finally estimated by (21).

5. EVALUATION

In this section, we compare the proposed MFMVDR+WF to the conventional single-channel WF ($L = 1$) and the MFMVDR presented in [6] [7] with the proposed parameter estimation from Section 4. In our experiments, we also found the previously described sensitivity of the MFWF in (11) to estimation errors of the speech and noise PSD as in [8]. However, we now show that the proposed decomposition of the MFWF into MFMVDR and WF does not suffer from this sensitivity.

For the evaluation we employ 120 sentences from the TIMIT database [14] spoken by different speakers (6 male, 6 female). As noise



(a) Clean speech (b) Noisy speech

(c) Results obtained with the WF ($L = 1$) (d) Results obtained with the MFMVDR [8] (e) Results obtained with the MFMVDR + WF

Fig. 1. Spectrograms of the clean speech (a), noisy speech (b), and the resulting processed speech (c)-(e) with blindly estimated parameters from a female speaker corrupted by modulated white Gaussian noise at 5 dB SNR. The Wiener filter in (c) clearly decreases the background noise but also introduces speech distortions. The MFMVDR in (d) results in less speech distortions but also less noise reduction. In panel (e), the proposed approach combines the benefits of (c) and (d), speech distortions are less than in (c) and noise reduction is more than in (d).

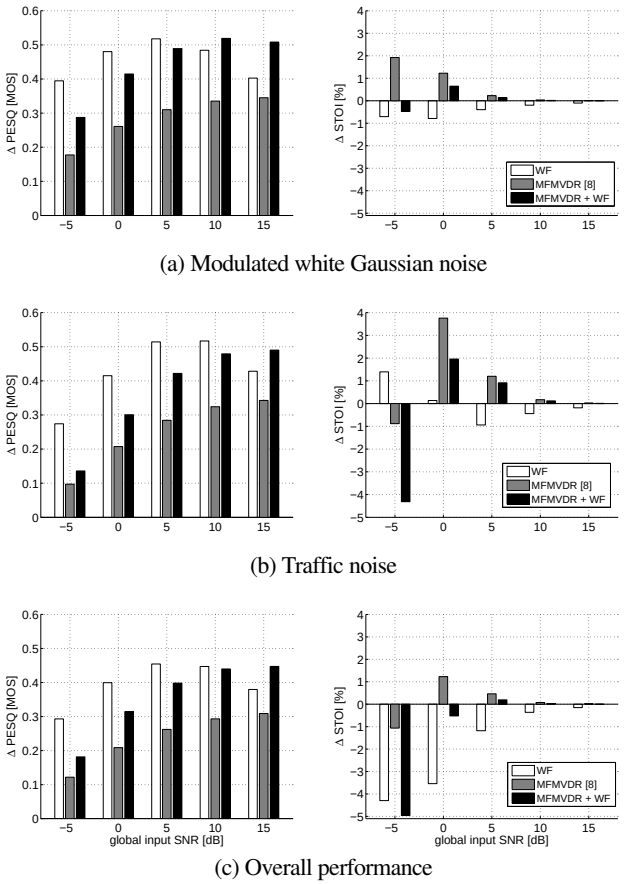
signals, we use white Gaussian noise, modulated white Gaussian noise with a modulation frequency of 0.5 Hz and babble and traffic noise. The sampling rate is set to $f_s = 16$ kHz. As the spectral analysis and synthesis window, we employ a square-root Hann window. Further, for the multi-frame approaches we use a high temporal resolution with a frame length of 4 ms and an overlap of 75 % to increase the exploitable IFC. The number of consecutive time-frames is set to $L = 18$ as in [8], resulting in 21 ms of data employed in each filtering operation. Also in our experiments this value is a good compromise between performance and computational complexity. In order to ensure a fair comparison, for the WF ($L = 1$) we apply a frame length of 21 ms. The overlap is set to 50 %. Moreover, the WF and Wiener post-filter were limited to $h_{\min} = -17$ dB in (15). For the WF ($L = 1$), an estimate of the speech PSD $\phi_x(k,m)$ is obtained using the decision-directed approach with a smoothing parameter of 0.98 [3].

For the trained coherence $\mathbf{\Gamma}_{\mathbf{n}}(k,m)$ we used a dataset of 60 speech files and 3 noise types at -5, 0 and 5 dB. We make sure that the training data differs from the evaluation data. As noise signals, we used pink noise, office noise and multi-talker babble noise in the training.

In Fig. 1, the resulting spectrograms of the processed speech signals are shown. It can be seen that the single-channel WF with $L = 1$ clearly decreases the background noise but speech components are also attenuated which results in speech distortions. The MFMVDR yields less speech distortions, but also less noise reduction. The proposed MFMVDR+WF combines the benefits of the WF and MFMVDR resulting in less speech distortions than the WF and the background noise is reduced more compared to the MFMVDR. Hence, we achieve a better trade-off between speech distortions and noise reduction with the proposed MFMVDR+WF. Informal listening tests confirm the results. But in terms of the background noise, more artifacts are audible with the MFMVDR+WF than with the WF ($L = 1$) but less than without the post-filter. This artifacts can be reduced for the MFMVDR approaches by using a different speech PSD estimator than in (19) like the decision-directed approach [3] or cepstral smoothing [15].

Further, we evaluate the MFMVDR, WF and MFMVDR+WF in terms of PESQ [10] and STOI [11] improvements compared to the noisy speech signal. PESQ and STOI are instrumental measures for speech quality and intelligibility respectively. In Fig. 2, the results for modulated white Gaussian noise, traffic noise and an average over all evaluated speech and noise files are given. It can be seen that the proposed noise reduction algorithm exhibits considerably higher PESQ improvements than the MFMVDR for all noise and SNR conditions. However, the conventional WF ($L = 1$) performs better than the MFMVDR and even better than the MFMVDR+WF for SNRs up to 10 dB. Considering the overall performance at 5 dB SNR, the proposed approach performs 0.14 MOS better than the MFMVDR and 0.05 MOS worse than the WF. In terms of STOI, the multi-frame algorithms achieve mainly improvements for SNRs over 0 dB in comparison to the noisy input. The improvements with post-filter are smaller than without. In contrast, the WF ($L = 1$) achieves no predicted speech intelligibility for all SNRs. At 5 dB SNR, the overall performance of the proposed approach yields a STOI score 0.3 % worse than the MFMVDR and 1.4 % better than the WF.

The results indicate that the WF ($L = 1$) reduces the background noise well, which can be seen in the PESQ performance and in Fig. 1. However, the speech components are also affected such that PESQ gets worse with increasing SNR and speech intelligibility improvements are not be predicted by STOI. The MFMVDR is designed to prevent speech distortions and leads to improved STOI scores for SNRs larger than 0 dB. However, in Fig. 1 it can be seen that the noise reduction is less compared to the WF which is why the instrumentally predicted quality improvements are less. The proposed algorithm is a combination of the MFMVDR and WF. The MFMVDR ensures that speech distortions are kept low whereas the Wiener post-filter is designed to minimize the



(a) Modulated white Gaussian noise

(b) Traffic noise

(c) Overall performance

Fig. 2. Noise reduction performance of the WF, MFMVDR and the proposed MFMVDR+WF. The plots show the PESQ and STOI improvements compared to the noisy speech signal. From top to bottom, the results for modulated white Gaussian noise (a), traffic noise (b) and an averaged performance (c) over 120 speech signals and 4 noise types are shown.

mean-squared error between the clean speech signal and the estimated speech. Thus, the background noise is reduced much more such that the PESQ scores are better compared to the MFMVDR, while STOI is slightly reduced but remains positive for SNRs larger 0 dB.

6. CONCLUSION

In this paper, we proposed to add a Wiener post-filter to the single-microphone multi-frame minimum variance distortionless response (MFMVDR) filter [6] [7] as the minimum-mean-squared error optimal decomposition of the multi-frame Wiener filter (MFWF) [7] in the short-time Fourier domain for single-microphone noise reduction. While in an oracle setup the MFWF leads to a similar noise reduction performance as the MFMVDR [7], in a blind setup the MFWF is very sensitive to estimation errors [8]. In this paper we showed that robust results can also be achieved in a blind setup by decomposing the MFWF into a MFMVDR and a single-channel Wiener post-filter. For the parameter estimation, we propose to pre-train typical interframe coherence matrices of the interferences over a large database. The proposed approach has been shown to be a good compromise between the performance of a single-channel Wiener filter and the MFMVDR in terms of speech quality and intelligibility predicted by PESQ and STOI respectively.

7. REFERENCES

[1] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 47–56, Mar. 2011.

[2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, 2006.

[3] 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, Dec. 1984.

[4] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[5] H. Huang, L. Zhao, J. Chen, and J. Benesty, "A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction," *ELSEVIER Digital Signal Process.*, vol. 33, pp. 169–179, Oct. 2014.

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

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

[8] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 1355–1365, Sep. 2014.

[9] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*. Springer, 2001, pp. 39–60.

[10] "ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[11] 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, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[12] A. Schasse and R. Martin, "Online inter-frame correlation estimation methods for speech enhancement in frequency subbands," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, May 2013, pp. 7482–7486.

[13] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*. John Wiley & Sons, 2005, vol. 40.

[14] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," in *National Institute of Standards and Technology (NIST)*, 1988.

[15] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, Mar. 2008, pp. 4897–4900.