Sensitivity Analysis of the Multi-Frame MVDR Filter for Single-Microphone Speech Enhancement

Dörte Fischer and Simon Doclo
Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All,
University of Oldenburg, Germany
{doerte.fischer,simon.doclo}@uni-oldenburg.de

***Abstract*—Recently, a multi-frame minimum variance distortionless response (MFMVDR) filter for single-microphone noise reduction has been proposed, which exploits speech correlation across consecutive time frames. It has been shown that the MFMVDR filter achieves impressive results when the speech interframe correlation vector can be accurately estimated. In this paper, we analyze the influence of estimation errors for all required parameters, i.e., the speech interframe correlation vector and the undesired correlation matrix, on the performance of the MFMVDR filter. We compare the performance difference between oracle estimators and practically feasible blind estimators. Experimental results show that even small estimation errors substantially degrade the speech quality, where the most critical parameter is the speech interframe correlation vector.**

I. Introduction

In speech communication systems such as hearing aids or mobile phones, the target speech signal is often affected by ambient background noise, decreasing speech quality and intelligibility, especially at low signal-to-noise ratios (SNRs). Hence, noise reduction algorithms are required, which aim to suppress the background noise while preserving the speech components.

In many single-microphone noise reduction algorithms, the noisy signal is processed in the short-time Fourier transform (STFT) domain. To obtain an estimate of the speech signal, typically a multiplicative real-valued gain function is applied to the noisy speech signal at each time-frequency point [1]. These approaches intrinsically assume that consecutive time frames are uncorrelated, such that each time-frequency point can be processed independently. However, since it is well known that speech is highly correlated over time, in [2] [3] it was proposed to exploit this speech correlation across time frames, using a signal model that considers the current frame as well as previous frames. Based on this signal model, the multi-frame minimum variance distortionless response (MFMVDR) filter for single-microphone noise reduction was proposed. This filter minimizes the total signal output power while not distorting correlated speech components.

In [2] [3] it has been shown that the MFMVDR filter achieves a good noise reduction performance and impressive results in terms of speech distortion when using an oracle estimator for the speech interframe correlation vector, which however requires the noise signal to be available. Since in practice obviously only the noisy speech signal is available, in [4] [5] blind maximum-likelihood (ML) estimators for the speech interframe correlation vector have been proposed. In order to better understand the performance of the MFMVDR filter using either the oracle estimator [2] [3] or the blind estimators [4] [5], in this paper we analyze the sensitivity to estimation errors in all required parameters, i.e., the speech interframe correlation vector and the undesired correlation matrix. We consider different oracle estimators for both quantities depending on the availability of the different signal components. Experimental results for different noise types and SNRs show that even small estimation errors, especially in the speech interframe correlation vector, may strongly decrease the speech quality, explaining the difference between the performance of the oracle and the blind estimators.

The structure of this paper is as follows. In Sections II and III the multi-frame signal model is formulated and the MFMVDR filter is briefly reviewed. In Section IV several oracle and blind estimators for the required parameters of the MFMVDR filter are introduced and their performance is evaluated in Section V.

II. Multi-Frame Signal Model

We assume a single-microphone setup, where a speech signal is degraded by additive noise. In the STFT domain, the (complex-valued) noisy speech signal $Y(k,m)$ is given by

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

where $X(k,m)$ denotes the speech signal and $V(k,m)$ denotes the noise signal. The indices k and m denote the frequency bin and the time frame, respectively.

Typically, in single-microphone approaches the speech signal $X(k,m)$ is estimated by applying a (real-valued) gain to $Y(k,m)$ [1]. Alternatively, in [2] [3] it has been proposed to estimate $X(k,m)$ at each time-frequency point by applying a (complex-valued) FIR filter with the filter coefficients $H_l(k,m)$ to the noisy speech signal, i.e.,

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

where * indicates the complex-conjugate operator. In vector notation, this equation can be written as

$$\hat{X}(k,m) = \boldsymbol{h}^H(k,m)\ \boldsymbol{y}(k,m), \tag{3}$$

This work was supported in part by the joint Lower Saxony-Israeli Project ATHENA and by the DFG Cluster of Excellence EXC 1077/1 Hearing4all.

ISBN 978-0-9928626-7-1

where ^{H} denotes the Hermitian operator and the L-dimensional vectors $\boldsymbol{h}(k,\,m)$ and $\boldsymbol{y}(k,\,m)$ correspond to the (time-varying) filter coefficients and L consecutive noisy speech coefficients, respectively, i.e.,

$$\boldsymbol{h}(k,m)=[H_0(k,m),\,H_1(k,m),\;\;...,\;\;H_{L-1}(k,m)]^T, \tag{4}$$

$$\boldsymbol{y}(k,m)=[Y(k,m),\,Y(k,m-1),\;\;...,\;\;Y(k,m-L+1)]^T. \tag{5}$$

The noisy speech vector $\boldsymbol{y}(k,m)$ can be decomposed as

$$\boldsymbol{y}(k,m)=\boldsymbol{x}(k,m)+\boldsymbol{v}(k,m), \tag{6}$$

where the speech vector $\boldsymbol{x}(k,m)$ and the noise vector $\boldsymbol{v}(k,m)$ are defined similarly as in (5). Assuming the speech and the noise signals are uncorrelated, the $L\times L$-dimensional noisy speech correlation matrix $\boldsymbol{\Phi_{yy}}(k,m)=\mathbb{E}\big[\boldsymbol{y}(k,m)\boldsymbol{y}^H(k,m)\big]$, with $\mathbb{E}[\cdot]$ the expectation operator, is given by

$$\boldsymbol{\Phi_{yy}}(k,m)=\boldsymbol{\Phi_{xx}}(k,m)+\boldsymbol{\Phi_{vv}}(k,m), \tag{7}$$

with $\boldsymbol{\Phi_{xx}}(k,m)$ the speech correlation matrix and $\boldsymbol{\Phi_{vv}}(k,m)$ the noise correlation matrix.

To exploit the speech correlation across frames it has been proposed in [2] to decompose the speech vector $\boldsymbol{x}(k,\,m)$ into correlated and uncorrelated components with respect to $X(k,m)$, i.e.,

$$\boldsymbol{x}(k,m)=\boldsymbol{\rho_x}(k,m)X(k,m)+\boldsymbol{x'}(k,m), \tag{8}$$

where the (time-varying) speech interframe correlation vector $\boldsymbol{\rho_x}(k,m)$ is defined as

$$\boldsymbol{\rho_x}(k,m)=\frac{\mathbb{E}[\boldsymbol{x}(k,m)X^*(k,m)]}{\mathbb{E}[|X(k,m)|^2]}=\frac{\boldsymbol{\Phi_{xx}}(k,m)\;\boldsymbol{e}}{\phi_X(k,m)}, \tag{9}$$

with $\phi_X(k,m)$ the speech power spectral density (PSD) and $\boldsymbol{e}$ an L-dimensional selection vector where the first element is equal to 1 and all other elements are equal to 0. Since the correlated speech component $\boldsymbol{\rho_x}(k,m)X(k,m)$ and the uncorrelated speech component $\boldsymbol{x'}(k,m)$ in (8) are uncorrelated by construction, the speech correlation matrix $\boldsymbol{\Phi_{xx}}(k,m)$ can be decomposed as

$$\boldsymbol{\Phi_{xx}}(k,m)=\phi_X(k,m)\boldsymbol{\rho_x}(k,m)\boldsymbol{\rho_x}^H(k,m)+\boldsymbol{\Phi_{x'x'}}(k,m), \tag{10}$$

with $\boldsymbol{\Phi_{x'x'}}(k,m)$ the correlation matrix of the uncorrelated speech component.

When substituting (8) into (6), the complete *multi-frame signal model* is defined as

$$\boxed{\boldsymbol{y}(k,m)=\boldsymbol{\rho_x}(k,m)X(k,m)+\boldsymbol{x'}(k,m)+\boldsymbol{v}(k,m)} \tag{11}$$

Since the uncorrelated speech component can be considered as an interference, we define the undesired signal vector $\boldsymbol{n}(k,m)$ as

$$\boldsymbol{n}(k,m)=\boldsymbol{x'}(k,m)+\boldsymbol{v}(k,m). \tag{12}$$

By substituting (10) into (7), the noisy speech correlation matrix can be written as

$$\boxed{\boldsymbol{\Phi_{yy}}(k,m)=\phi_X(k,m)\boldsymbol{\rho_x}(k,m)\boldsymbol{\rho_x}^H(k,m)+\boldsymbol{\Phi_{nn}}(k,m)} \tag{13}$$

with $\boldsymbol{\Phi_{nn}}(k,m)=\boldsymbol{\Phi_{x'x'}}(k,m)+\boldsymbol{\Phi_{vv}}(k,m)$ the undesired correlation matrix.

Similarly to (9), the noisy speech correlation vector $\boldsymbol{\rho_y}(k,m)$ and the noise correlation vector $\boldsymbol{\rho_v}(k,m)$ can be defined as

$$\boldsymbol{\rho_y}(k,m)=\frac{\boldsymbol{\Phi_{yy}}(k,m)\;\boldsymbol{e}}{\phi_Y(k,m)},\qquad \boldsymbol{\rho_v}(k,m)=\frac{\boldsymbol{\Phi_{vv}}(k,m)\;\boldsymbol{e}}{\phi_V(k,m)}, \tag{14}$$

with $\phi_Y(k,m)$ and $\phi_V(k,m)$ the noisy speech PSD and the noise PSD, respectively. Using (7), it can hence be easily shown that

$$\phi_Y(k,m)\boldsymbol{\rho_y}(k,m)=\phi_X(k,m)\boldsymbol{\rho_x}(k,m)+\phi_V(k,m)\boldsymbol{\rho_v}(k,m), \tag{15}$$

such that

$$\boldsymbol{\rho_x}(k,m)=\frac{\phi_Y(k,m)}{\phi_X(k,m)}\boldsymbol{\rho_y}(k,m)-\frac{\phi_V(k,m)}{\phi_X(k,m)}\boldsymbol{\rho_v}(k,m). \tag{16}$$

III. Multi-Frame MVDR Filter

In this section, we briefly review the MFMVDR filter proposed in [2] [3]. The objective of the MFMVDR filter is to minimize the PSD of the undesired component, subject to the constraint that the correlated speech component is not distorted, i.e.,

$$\min_{\boldsymbol{h}(k,m)}\;\;\boldsymbol{h}^H(k,m)\boldsymbol{\Phi_{nn}}(k,m)\boldsymbol{h}(k,m), \tag{17}$$

$$\text{subject to }\boldsymbol{h}^H(k,m)\boldsymbol{\rho_x}(k,m)=1.$$

Solving the optimization problem leads to the MFMVDR filter

$$\boxed{\boldsymbol{h}_{\text{MFMVDR}}(k,m)=\frac{\boldsymbol{\Phi}_{\boldsymbol{nn}}^{-1}(k,m)\boldsymbol{\rho_x}(k,m)}{\boldsymbol{\rho}_{\boldsymbol{x}}^H(k,m)\boldsymbol{\Phi}_{\boldsymbol{nn}}^{-1}(k,m)\boldsymbol{\rho_x}(k,m)}} \tag{18}$$

This formula is very similar to the well-known MVDR beamformer for multi-microphone noise reduction [6]. However, it should be noted that for the MFMVDR filter in (18) both the speech interframe correlation vector $\boldsymbol{\rho_x}(k,m)$ as well as the undesired correlation matrix $\boldsymbol{\Phi_{nn}}(k,m)$ (mainly due to the contribution of $\boldsymbol{\Phi_{x'x'}}(k,m)$) are typically highly time-varying, making it quite difficult to accurately estimate these quantities in practice.

By applying the matrix inversion lemma to (13), it can be easily shown that

$$\boldsymbol{\Phi}_{\boldsymbol{yy}}^{-1}(k,m)\boldsymbol{\rho_x}(k,m)=\boldsymbol{\Phi}_{\boldsymbol{nn}}^{-1}(k,m)\boldsymbol{\rho_x}(k,m). \tag{19}$$

When substituting (19) into (18), the resulting filter is equal to

$$\boxed{\boldsymbol{h}_{\text{MFMPDR}}(k,m)=\frac{\boldsymbol{\Phi}_{\boldsymbol{yy}}^{-1}(k,m)\boldsymbol{\rho_x}(k,m)}{\boldsymbol{\rho}_{\boldsymbol{x}}^H(k,m)\boldsymbol{\Phi}_{\boldsymbol{yy}}^{-1}(k,m)\boldsymbol{\rho_x}(k,m)}} \tag{20}$$

which is known as the multi-frame minimum power distortionless response (MFMPDR) filter and has been used in [2] [3]. Although in practice it is obviously much easier to estimate the noisy speech correlation matrix $\boldsymbol{\Phi_{yy}}(k,m)$ in (20) instead of the undesired correlation matrix $\boldsymbol{\Phi_{nn}}(k,m)$ in (18), it should be realized that the MFMVDR filter and the MFMPDR filter

ISBN 978-0-9928626-7-1 © EURASIP 2017

are only equivalent when the speech interframe correlation vector $\boldsymbol{\rho_x}(k,m)$ can be perfectly estimated. In [2] [3] $\boldsymbol{\rho_x}(k,m)$ has been accurately estimated, showing that in this case the MFMPDR filter is able to achieve impressive results in terms of noise reduction and especially speech distortion. Similarly to the corresponding MVDR and MPDR beamformers for multi-microphone noise reduction [6], it is however to be expected that the MFMPDR filter is more sensitive to estimation errors of the speech interframe correlation vector than the MFMVDR filter, which will be investigated in the following sections.

IV. Oracle and Blind Estimators

The main objective of this paper is to investigate the sensitivity of the MFMVDR and MFMPDR filters against estimation errors in the speech interframe correlation vector $\boldsymbol{\rho_x}(k,m)$ and the undesired correlation matrix $\boldsymbol{\Phi_{nn}}(k,m)$ (for the MFMVDR filter only). Therefore, in this section we present several oracle estimators for the speech interframe correlation vector (Section IV-A) and the undesired correlation matrix (Section IV-B), in addition to blind estimators which can be used in practice (Section IV-C).

For the *oracle estimators*, we either make the (unrealistic) assumption that a perfect estimate of the speech correlation matrix $\boldsymbol{\Phi_{xx}}(k,m)$, the noise correlation matrix $\boldsymbol{\Phi_{vv}}(k,m)$ or the uncorrelated speech component $\boldsymbol{x'}(k,m)$ is available. The perfect speech and noise correlation matrix estimates are computed as

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(k,m) = \lambda \hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(k,m-1) + (1-\lambda)\boldsymbol{x}(k,m)\boldsymbol{x}^H(k,m), \quad (21)$$

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m) = \lambda \hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m-1) + (1-\lambda)\boldsymbol{v}(k,m)\boldsymbol{v}^H(k,m), \quad (22)$$

with λ the recursive averaging factor. Since in practice only the noisy speech signal is available, for the *blind estimators* we only assume that the noisy speech vector $\boldsymbol{y}(k,m)$ and the noisy speech correlation matrix estimate, computed as

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m) = \lambda \hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m-1) + (1-\lambda)\boldsymbol{y}(k,m)\boldsymbol{y}^H(k,m), \quad (23)$$

are available.

A. Oracle Estimators for Speech Interframe Correlation

For the first oracle estimator, we assume that a perfect estimate of the speech correlation matrix $\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(k,m)$ in (21) is available. Similarly to (9), the speech interframe correlation vector can then be estimated as

$$\boxed{\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}}(k,m) = \frac{\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(k,m)\ \boldsymbol{e}}{\hat{\phi}_X(k,m)}} \quad (24)$$

with the speech PSD estimate $\hat{\phi}_X(k,m) = \boldsymbol{e}^T \hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(k,m)\boldsymbol{e}$.

For the second oracle estimator, we assume that a perfect estimate of the noise correlation matrix $\hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m)$ in (22) is available. Similarly to (14), the noisy speech interframe correlation vector and the noise interframe correlation vector can be estimated as

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}(k,m) = \frac{\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)\ \boldsymbol{e}}{\hat{\phi}_Y(k,m)}, \qquad \hat{\boldsymbol{\rho}}_{\boldsymbol{v}}(k,m) = \frac{\hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m)\ \boldsymbol{e}}{\hat{\phi}_V(k,m)}, \quad (25)$$

with $\hat{\phi}_Y(k,\ m) = \boldsymbol{e}^T \hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,\ m)\boldsymbol{e}$ and $\hat{\phi}_V(k,\ m) = \boldsymbol{e}^T \hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m)\boldsymbol{e}$. Based on (16), the speech interframe correlation vector can then be estimated as

$$\boxed{\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{II}}(k,m) = \frac{\hat{\phi}_Y(k,m)}{\hat{\phi}_X(k,m)}\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}(k,m) - \frac{\hat{\phi}_V(k,m)}{\hat{\phi}_X(k,m)}\hat{\boldsymbol{\rho}}_{\boldsymbol{v}}(k,m)} \quad (26)$$

with $\hat{\phi}_X(k,m) = \hat{\phi}_Y(k,m) - \hat{\phi}_V(k,m)$. It should be noted that since $\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)$ is not exactly equal to $\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(k,m) + \hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m)$, there will typically be a small difference between the "optimal" oracle estimate $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}}(k,m)$ and the oracle estimate $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{II}}(k,m)$.

B. Oracle Estimators for Undesired Correlation Matrix

Please recall that the undesired correlation matrix $\boldsymbol{\Phi_{nn}}(k,m)$ is equal to the sum of the correlation matrices of the uncorrelated speech component and the noise vector (cf. Section II), i.e.,

$$\boldsymbol{\Phi_{nn}}(k,m) = \boldsymbol{\Phi}_{\boldsymbol{x'x'}}(k,m) + \boldsymbol{\Phi_{vv}}(k,m). \quad (27)$$

For the first oracle estimator, we assume that perfect estimates of the uncorrelated speech component and the undesired signal vector, computed as

$$\widehat{\boldsymbol{x}}'(k,m) = \boldsymbol{x}(k,m) - \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}}(k,m)X(k,m), \quad (28)$$

$$\hat{\boldsymbol{n}}(k,m) = \widehat{\boldsymbol{x}}'(k,m) + \boldsymbol{v}(k,m), \quad (29)$$

are available. The undesired correlation matrix can then be estimated as

$$\boxed{\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}^{\mathrm{I}}(k,m) = \lambda \hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}^{\mathrm{I}}(k,m-1) + (1-\lambda)\hat{\boldsymbol{n}}(k,m)\hat{\boldsymbol{n}}^H(k,m)} \quad (30)$$

For the second oracle estimator, we assume that the uncorrelated speech component $\boldsymbol{x'}(k,m)$ can be neglected, such that the undesired correlation matrix can be approximated as

$$\boxed{\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}^{\mathrm{II}}(k,m) = \hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m)} \quad (31)$$

C. Blind estimators

In practice, only the noisy speech signal $Y(k,m)$ is available, such that all required quantities for the multi-frame filters need to be blindly estimated from this signal.

Estimating the undesired correlation matrix $\boldsymbol{\Phi_{nn}}(k,\ m)$ requires an estimate of the noise vector $\boldsymbol{v}(k,m)$ and possibly even the uncorrelated speech component $\boldsymbol{x'}(k,m)$ (cf. Section IV-B). Since both components are typically highly time-varying, it is hardly feasible to estimate these components at each time-frequency point in practice. Hence, for realistic scenarios we will only consider the MFMPDR in (20) using the noisy speech correlation matrix estimate in (23).

To blindly estimate the speech interframe correlation vector $\boldsymbol{\rho_y}(k,\ m)$ from the noisy speech signal $Y(k,\ m)$, different estimators have been proposed in [4] [5]. Similarly to (26), the ML estimator for $\boldsymbol{\rho_x}(k,m)$ proposed in [4] is given by

$$\boxed{\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{ML}}(k,m) = \frac{\hat{\xi}(k,m)+1}{\hat{\xi}(k,m)}\hat{\boldsymbol{\rho}}_{\boldsymbol{y}}(k,m) - \frac{1}{\hat{\xi}(k,m)}\boldsymbol{\mu}_{\boldsymbol{\rho_v}}} \quad (32)$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

with $\hat{\xi}(k,m)$ an estimate of the *a-priori* SNR $\xi(k,m) = \frac{\phi_X(k,m)}{\phi_V(k,m)}$. The main difference between (26) and (32) is the fact that the estimated noise interframe correlation vector $\hat{\boldsymbol{\rho}}_{\boldsymbol{v}}(k,m)$ is assumed to be constant for all time-frequency points, such that it can be replaced by its mean value $\boldsymbol{\mu}_{\boldsymbol{\rho}_{\boldsymbol{v}}}$. This mean value is determined by the frame overlap and the STFT analysis window [4].

In [4] and [5] different estimation procedures have been proposed for computing the *a-priori* SNR estimate $\hat{\xi}(k,m)$. In both procedures the noise PSD $\phi_V(k,m)$ has been estimated using the noise PSD estimator proposed in [7]. Assuming that the speech signal $X(k,m)$ and the noise signal $V(k,m)$ follow complex-valued, zero-mean Gaussian distributions, in [4] the speech PSD $\phi_X(k,m)$ has been estimated using the ML estimator [8]. However, due to the use of short analysis frames in the multi-frame filter, it has been shown that the ML estimate for $\phi_X(k,m)$ leads to outliers in $\hat{\xi}(k,m)$ and hence $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{ML}}(k,m)$, which may result in unpleasant artifacts in the processed speech [4] [5]. Since it is well known that the decision directed approach (DDA) [9] provides smoother estimates of $\xi(k,m)$ than when using the ML estimate of $\phi_X(k,m)$ [10], in [5] it has been proposed to estimate $\xi(k,m)$ by applying a modified DDA for short analysis frames based on temporally smoothed observations.

V. Experimental Results

In Section V-A we evaluate the oracle performance of the MFMVDR and MFMPDR filters for different oracle estimators of the speech interframe correlation vector (cf. Section IV-A) and the undesired correlation matrix (cf. Section IV-B), thereby analyzing the sensitivity of these filters to estimation errors. In Section V-B we compare the oracle performance with the realistic performance of the MFMPDR filter using blind estimators of the speech interframe correlation vector (cf. Section IV-C).

Label	Description
MFMVDR$_p$	- "Perfect" estimate - Compute $\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(k,m)$ as in (30)
MFMVDR$_a$	- $\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(k,m) \approx \hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m)$ - Compute $\hat{\boldsymbol{\Phi}}_{\boldsymbol{vv}}(k,m)$ as in (22)
MFMPDR	- Use $\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)$ instead of $\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(k,m)$ - Compute $\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)$ as in (23)

TABLE I
Overview of the applied correlation matrices

For all considered techniques we have used 60 sentences from the TIMIT database [11], spoken by different speakers (5 male, 5 female). As noise signals we have used white Gaussian noise and traffic noise. The sampling frequency is equal to 16 kHz and an SNR range of 0 dB to 15 dB has been considered. As the STFT analysis and synthesis window we have used a square-root Hann window. To increase the exploitable interframe correlation we have used a high temporal resolution with a frame length of 4 ms and an overlap of 75 %. Similarly as in [4] [5], the number of consecutive time-frames is set to $L = 18$, resulting in 21 ms of data used in each filtering operation and the recursive smoothing factor λ is set to 0.88, allowing for tracking fast changes. The performance is evaluated in terms of PESQ [12] and segmental SNR (segSNR) [13] improvement compared to the noisy speech signal, averaged over all considered sentences and noise types. For both measures the clean speech signal has been used as reference signal.

A. Oracle Performance

In this section we analyze the sensitivity of the MFMVDR and MFMPDR filters for different oracle estimators of the speech interframe correlation vector and the undesired correlation matrix. As the optimal filter we consider the MFMVDR using perfect estimates, i.e., $\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}^{\mathrm{I}}(k,m)$ in (30) and $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}}(k,m)$ in (24).

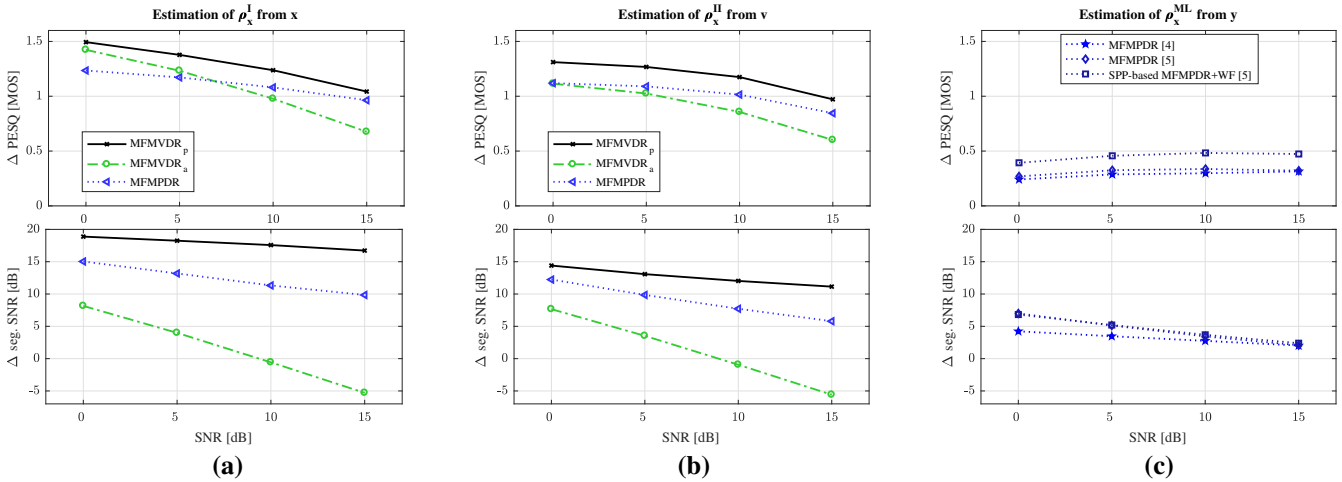

Fig. 1. Influence of different estimators for the speech interframe correlation vector $\boldsymbol{\rho}_{\boldsymbol{x}}$ on the MFMVDR$_p$, MFMVDR$_a$ and MFMPDR filter: using in **(a)** the perfect estimate $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}}(k,m)$, in **(b)** the accurate estimate $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{II}}(k,m)$ and in **(c)** the blind estimate $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{ML}}(k,m)$. The plots show the average performance in terms of PESQ and segmental SNR improvements.

ISBN 978-0-9928626-7-1 © EURASIP 2017

Fig. 1(a) depicts the average performance of the considered filters MFMVDR$_p$, MFMVDR$_a$, MFMPDR (cf. Table V) for different SNRs, using the (perfect) oracle estimate $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}}(k,m)$ in (24). First, it can be observed that the MFMVDR$_p$ achieves the highest PESQ and segSNR improvements for all SNRs, which are quite impressive for a single-microphone noise reduction approach. Although the MFMPDR filter should theoretically be equivalent to the MFMVDR filter when using a perfect estimate of $\boldsymbol{\rho}_{\boldsymbol{x}}(k,m)$, it can be observed that the performance is lower. This can be explained due to the fact that in practice $\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(k,m)$ is not exactly equal to $\hat{\phi}_X(k,m)\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}}(k,m)\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{I}\,H}(k,m) + \hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}^{\mathrm{I}}(k,m)$, cf. (13). The performance of the MFMVDR$_a$ filter is the worst of all considered filters, especially in terms of segSNR and for high SNRs. This implies that the influence of the uncorrelated speech component is crucial, especially at high SNRs, and neglecting these component increases the amount of speech distortion, leading to a reduced speech quality.

Fig. 1(b) depicts the average performance of the considered filters using oracle estimate $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{II}}(k,m)$ in (26), i.e., considering small estimation errors. Compared to the results in Fig. 1(a) using the perfectly estimated $\boldsymbol{\rho}_{\boldsymbol{x}}(k,m)$, it can be observed that the performance for all filters decreases. E.g., for an input SNR of 5 dB, the PESQ improvements are reduced by 0.1 MOS for the MFMVDR$_p$ and the MFMPDR, respectively, and by 0.2 MOS for the MFMVDR$_a$. However, the MFMVDR$_p$ still outperforms both the MFMVDR$_a$ and the MFMPDR. These results show that even small estimation errors in $\boldsymbol{\rho}_{\boldsymbol{x}}$ decrease the overall performance. In addition, informal listening tests revealed slight artifacts in the background noise, which is due to the fact that $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}^{\mathrm{II}}(k,m)$ in (26) is not exactly zero during speech pauses.

B. Realistic Performance Using Blind Estimators

In this section, we evaluate the realistic performance of the MFMPDR filter using blind estimators of the speech interframe correlation vector. Fig. 1(c) depicts the average performance of the MFMPDR filter, either using the ML estimate of the speech PSD [4] or using a modified DDA [5] to estimate the *a-priori* SNR (cf. Section IV-C). In addition, this figure shows the performance of a combined MFMPDR filter with a single-channel Wiener filter (WF) (using speech presence probability weighting), which has been proposed in [5] to reduce artifacts caused by estimation errors of $\boldsymbol{\rho}_{\boldsymbol{x}}$ during speech pauses.

As has already been shown in [5], these results show that the MFMPDR with modified DDA and combined with a WF leads to slightly larger PESQ and segSNR improvements than the MFMPDR in [4]. More importantly, when comparing the performance of the MFMPDR filter using blind estimators for the speech interframe correlation vector (Fig. 1(c)) with the related oracle estimator (Fig. 1(b)), it can be observed that the performance is substantially degraded. E.g., for an input SNR of 5 dB, the PESQ improvement is reduced by 0.55 - 0.82 MOS. These results indicate that as expected estimation errors in $\boldsymbol{\rho}_{\boldsymbol{x}}$ lead to a strongly reduced performance for the MFMPDR filter. Hence, further work is required to either improve the accuracy of blind estimators, which is a non-trival task since $\boldsymbol{\rho}_{\boldsymbol{x}}$ is a highly time-varying quantity, or to improve the robustness of the MFMPDR filter against estimation errors.

VI. Conclusion

In this paper, we analyzed the sensitivity of the MFMVDR filter for single-microphone noise reduction to estimation errors in the required parameters, i.e., the speech interframe correlation vector and the undesired correlation matrix. In [2] [3] it has been shown that the MFMVDR filter achieves good noise reduction performance while keeping speech distortion low when using an oracle estimator for the speech interframe correlation vector. To quantify the performance of the MFMVDR filter in the presence of estimation errors in all required parameters we compared different (unrealistic) oracle and practically feasible blind estimators. The results show that, as expected, the optimal MFMVDR filter achieves the best results. However, even small estimation errors in all required parameters lead to a reduced performance. Furthermore, the results show that the MFMPDR filter and the perfect MFMVDR filter perform similar for small estimation errors in the speech interframe correlation vector. However, when using existing blind estimators the performance of the practically feasible MFMPDR is strongly reduced due to large estimations errors in the speech interframe correlation vector.

References

[1] 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, 2013.

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

[3] 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.

[4] 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.

[5] 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 Symposium on Speech Communication*, Paderborn, Germany, Oct. 2016, pp. 292–296.

[6] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

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

[8] R. McAulay and M. 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.

[9] 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.

[10] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, 1994.

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

[12] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2007.

[13] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*. Prentice Hall, 1988.

ISBN 978-0-9928626-7-1 © EURASIP 2017