Comparison of Parameter Estimation Methods for Single-Microphone Multi-Frame Wiener Filtering

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

***Abstract*—The multi-frame Wiener filter (MFWF) for single-microphone speech enhancement is able to exploit speech correlation across consecutive time-frames in the short-time Fourier transform (STFT) domain. To achieve a high speech correlation, typically an STFT with a high time-resolution but a low frequency-resolution is applied. The MFWF can be decomposed into a multi-frame minimum power distortionless response (MFMPDR) filter and a single-frame Wiener postfilter. To implement the MFWF using this decomposition, estimates of several parameters are required, namely the speech correlation vector, the noisy speech correlation matrix, and the power spectral densities at the output of the MFMPDR filter. Correlations can be estimated either *directly* in the low frequency-resolution STFT filterbank, *indirectly* by estimating periodograms in a high frequency-resolution filterbank and applying the Wiener-Khinchin theorem, or in a *combined* way. In this paper, we compare the performance of different estimators for the required parameters. Experimental results for different speech material, noise conditions, and signal-to-noise ratios show that using a combined estimator for the speech correlation vector yields the best results in terms of speech quality compared to existing direct and indirect estimators.**

I. Introduction

Speech signals recorded in communication devices are frequently corrupted by undesired additive noise. To improve the speech quality, single-microphone noise reduction is often applied in the short-time Fourier transform (STFT) domain [1]. In contrast to single-frame approaches, where a (real-valued) gain is applied to each noisy STFT coefficient independently, multi-frame approaches aim to exploit speech correlation across consecutive time-frames [1]–[7]. To achieve a high speech correlation, typically an STFT with a high time-resolution but a low frequency-resolution is applied. In [1], [3], a complex-valued multi-frame Wiener filter (MFWF) was proposed which minimizes the mean square error (MSE) between the desired and estimated speech coefficients. In [4], [6], [8], it was reported that in practice the MFWF is very sensitive to estimation errors and more robust results can be obtained by decomposing the MFWF into a multi-frame minimum power distortionless response (MFMPDR) filter [3] and a single-frame Wiener postfilter [8].

The MFMPDR filter requires estimates of the speech correlation vector and noisy speech correlation matrix, where it was shown in [5] that the MFMPDR filter is more sensitive to estimation errors in the speech correlation vector. In [4], the noisy speech correlation matrix and the speech correlation vector were estimated *directly* in the low frequency-resolution STFT filterbank, where for the estimation of the speech correlation vector a fixed (i.e., time- and frequency-independent) average noise correlation vector was assumed. In [6], it was proposed to estimate the noisy speech correlation matrix and speech correlation vector *indirectly* by first estimating the noisy speech and speech periodograms in a high frequency-resolution filterbank and then applying the Wiener-Khinchin theorem. In this paper, we compare the performance of different estimators for the required parameters of the MFMPDR filter with single-frame Wiener postfilter, i.e., the speech correlation vector, the noisy speech correlation matrix, and the power spectral densities (PSDs) at the output of the MFMPDR filter. In addition to existing direct and indirect estimators, we also propose *combined* estimators using both the low and high frequency-resolution filterbank.

Experimental results for different speech signals, noise types, and signal-to-noise ratios (SNRs) show that using a combined estimator for the speech correlation vector can improve speech quality compared to using a direct or indirect estimator. Furthermore, the results show that a Wiener postfilter improves the speech quality compared to the MFMPDR filter, but independently estimating the PSDs at the output of the MFMPDR does not yield an improvement.

II. Multi-Frame Signal Model

By applying an STFT with frame length K and analysis window h_K to the noisy microphone signal $y[n]$ at time n, the noisy speech coefficient $Y[k,l]$ at time-frame l and frequency-bin k is obtained, i.e.,

$$Y[k,l]=S[k,l]+N[k,l], \quad k\in\left\{-\frac{K}{2}+1,-\frac{K}{2}+2,...,\frac{K}{2}\right\}, \tag{1}$$

where $S[k,l]$ and $N[k,l]$ denote the speech and the noise coefficients, respectively. The noisy speech vector $\boldsymbol{y}[k,l]$ is defined by considering M consecutive time-frames, i.e.,

$$\boldsymbol{y}[k,l]=\left[Y[k,l],Y[k,l-1],...,Y[k,l-M+1]\right]^T, \tag{2}$$

where ^{T} denotes the transpose operator. Using (1), this vector can be written as

$$\boldsymbol{y}[k,l]=\boldsymbol{s}[k,l]+\boldsymbol{n}[k,l], \tag{3}$$

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 352015383 – SFB 1330 B2 and Cluster of Excellence 1077 Hearing4all.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

where the speech vector $\boldsymbol{s}[k,l]$ and the noise vector $\boldsymbol{n}[k,l]$ are defined similarly as in (2). In multi-frame approaches the speech coefficient $S[k,l]$ is estimated by applying an M-dimensional (complex-valued) finite impulse response filter $\boldsymbol{h}[k,l]$ to the noisy speech vector, i.e.,

$$\hat{S}[k,l] = \boldsymbol{h}^H[k,l]\ \boldsymbol{y}[k,l], \tag{4}$$

where ^{H} denotes the Hermitian operator. For conciseness, in the remainder of this paper the indices k and l will be omitted wherever possible.

Assuming that the speech and noise signals are uncorrelated, the $M \times M$-dimensional noisy speech correlation matrix $\boldsymbol{R_y} = \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^H\right]$, with $\mathbb{E}[\cdot]$ the expectation operator, is given by

$$\boldsymbol{R_y} = \boldsymbol{R_s} + \boldsymbol{R_n}, \tag{5}$$

where $\boldsymbol{R_s} = \mathbb{E}\left[\boldsymbol{s}\boldsymbol{s}^H\right]$ and $\boldsymbol{R_n} = \mathbb{E}\left[\boldsymbol{n}\boldsymbol{n}^H\right]$ denote the speech and noise correlation matrices, respectively.

Considering the speech correlation across time-frames, it was proposed in [3] to decompose the speech vector $\boldsymbol{s}$ into a temporally correlated speech component $\boldsymbol{x}$ and a temporally uncorrelated speech component $\boldsymbol{x'}$ with respect to the speech coefficient S, i.e.,

$$\boldsymbol{s} = \boldsymbol{x} + \boldsymbol{x'} = \boldsymbol{\gamma_s} S + \boldsymbol{x'}, \tag{6}$$

where $\boldsymbol{\gamma_s}$ denotes the normalized speech correlation vector, which is defined as

$$\boldsymbol{\gamma_s} = \frac{\mathbb{E}[\boldsymbol{s}S^*]}{\mathbb{E}[|S|^2]} = \frac{\boldsymbol{r_s}}{\phi_S}, \tag{7}$$

where * denotes the complex-conjugate operator and $\boldsymbol{r_s}$ denotes the speech correlation vector. Due to the normalization with the speech PSD $\phi_S = \mathbb{E}\left[|S|^2\right] = \boldsymbol{r_s}(1)$, the first element of $\boldsymbol{\gamma_s}$ is equal to 1.

Using (5), (6) and (7), the speech correlation matrix $\boldsymbol{R_s}$ can be decomposed into the rank-1 correlation matrix $\boldsymbol{R_x} = \phi_S \boldsymbol{\gamma_s}\boldsymbol{\gamma_s}^H$ and the correlation matrix $\boldsymbol{R_{x'}} = \mathbb{E}\left[\boldsymbol{x'}\boldsymbol{x'}^H\right]$. Hence, the speech correlation vector $\boldsymbol{r_s}$ and the speech PSD ϕ_S in (7) can be computed as

$$\boldsymbol{r_s} = \boldsymbol{R_x}\boldsymbol{e}, \qquad \phi_S = \boldsymbol{e}^T\boldsymbol{R_x}\boldsymbol{e}, \tag{8}$$

with $\boldsymbol{e} = \left[1,\ 0,\ ...,\ 0\right]^T$ an M-dimensional selection vector. Considering the uncorrelated speech component $\boldsymbol{x'}$ in (6) as an interference, the *multi-frame signal model* is given by

$$\boldsymbol{y} = \boldsymbol{\gamma_s} S + \boldsymbol{u}, \tag{9}$$

where the undesired signal vector $\boldsymbol{u} = \boldsymbol{x'} + \boldsymbol{n}$ and the correlated speech vector $\boldsymbol{\gamma_s} S$ are uncorrelated. Using (9), the noisy speech correlation matrix can therefore be written as

$$\boldsymbol{R_y} = \phi_S \boldsymbol{\gamma_s}\boldsymbol{\gamma_s}^H + \boldsymbol{R_u}, \tag{10}$$

with the undesired correlation matrix $\boldsymbol{R_u} = \boldsymbol{R_{x'}} + \boldsymbol{R_n}$. Since $\boldsymbol{e}^T\boldsymbol{R_{x'}}\boldsymbol{e} = 0$, the undesired PSD $\phi_U = \boldsymbol{e}^T\boldsymbol{R_u}\boldsymbol{e} = \phi_N$.

Similarly to (7), the normalized noisy speech correlation vector $\boldsymbol{\gamma_y}$ and the normalized noise correlation vector $\boldsymbol{\gamma_n}$ are defined as

$$\boldsymbol{\gamma_y} = \frac{\boldsymbol{r_y}}{\phi_Y} = \frac{\boldsymbol{R_y}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R_y}\boldsymbol{e}}, \qquad \boldsymbol{\gamma_n} = \frac{\boldsymbol{r_n}}{\phi_N} = \frac{\boldsymbol{R_n}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R_n}\boldsymbol{e}}. \tag{11}$$

Using (5) and (11), it can be easily shown that

$$\phi_Y \boldsymbol{\gamma_y} = \phi_S \boldsymbol{\gamma_s} + \phi_N \boldsymbol{\gamma_n}, \tag{12}$$

such that the normalized speech correlation vector can be obtained as

$$\boldsymbol{\gamma_s} = \frac{\phi_S + \phi_N}{\phi_S}\boldsymbol{\gamma_y} - \frac{\phi_N}{\phi_S}\boldsymbol{\gamma_n}. \tag{13}$$

III. Multi-Frame Wiener Filter

The aim of the MFWF is to minimize the MSE between the desired speech coefficient S and the estimated speech coefficient $\hat{S}$ in (4), i.e.,

$$\hat{\boldsymbol{h}} = \underset{\boldsymbol{h}}{\operatorname{argmin}}\ E\left\{|\boldsymbol{h}^H\boldsymbol{y} - S|^2\right\}. \tag{14}$$

Solving this optimization problem using the multi-frame signal model in (9) leads to the MFWF [3]

$$\boldsymbol{h}_{\text{MFWF}} = \boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}\phi_S. \tag{15}$$

In [4], [6], [8] it was reported that in practice the MFWF is very sensitive to estimation errors and more robust results can be obtained by decomposing the MFWF into an MFMPDR filter [3] and a single-frame Wiener postfilter [8], i.e.,

$$\boldsymbol{h}_{\text{MFMPDR-WG}} = \underbrace{\frac{\boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}}{\boldsymbol{\gamma_s}^H\boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}}}_{\boldsymbol{h}_{\text{MFMPDR}}} \underbrace{\frac{\phi_S}{\phi_Y^{\text{out}}}}_{G_{\text{WG}}}, \tag{16}$$

where $\phi_Y^{\text{out}} = \left(\boldsymbol{\gamma_s}^H\boldsymbol{R_y}^{-1}\boldsymbol{\gamma_s}\right)^{-1}$ denotes the signal PSD at the output of the MFMPDR filter.

As can be observed from (16), the MFWF requires estimates of the noisy speech correlation matrix $\boldsymbol{R_y}$, the normalized speech correlation vector $\boldsymbol{\gamma_s}$, the speech PSD ϕ_S and the signal output PSD ϕ_Y^{out}. In Section IV ,we present an indirect method to estimate the correlation vectors $\boldsymbol{r_y}$, $\boldsymbol{r_s}$, and $\boldsymbol{r_n}$. In Section V, we present different estimators (direct, indirect, and combined) for the noisy speech correlation matrix, the normalized speech correlation vector, and the signal output PSD.

IV. Indirect Correlation Vector Estimation

The Wiener-Khinchin theorem states that the correlation of a wide-sense stationary process is given by the inverse discrete Fourier transform (IDFT) of the PSD. To accurately estimate the correlation based on the Wiener-Khinchin theorem, it was proposed in [6] to estimate the periodogram in a filterbank with a $2M/O$ higher frequency-resolution than the (processing) STFT filterbank, where O denotes the oversampling factor. In the following, the parameters in the high frequency-resolution filterbank will be denoted with a superscript $\{\}^F$. An overview of the indirect estimation of the correlation vectors is depicted in Fig. 1.

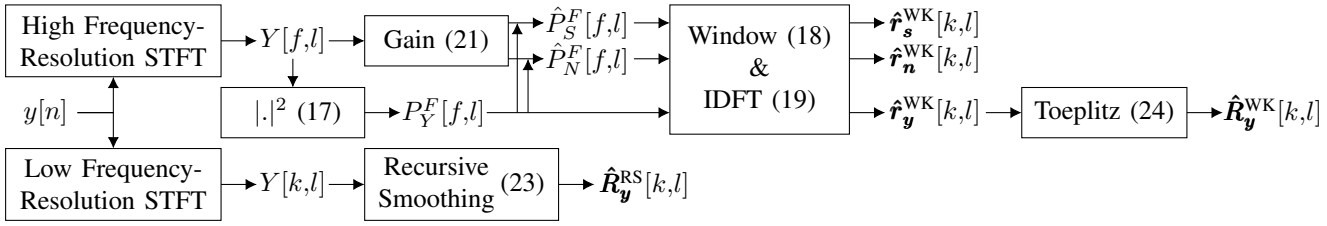

Fig. 1. Block diagram of parameter estimation from Sections IV and V-A.

In the high frequency-resolution filterbank, the noisy speech periodogram is given by

$$P_Y^F[f,l]=|Y^F[f,l]|^2,\quad f\in\left\{-\frac{F}{2}+1,-\frac{F}{2}+2,...,\frac{F}{2}\right\},\quad (17)$$

with $Y^F[f,l]$ the noisy speech coefficient at time-frame l and frequency-bin f. An estimate of the $2M$-dimensional noisy speech PSD vector $\hat{\boldsymbol{\phi}}_Y[k,l]$ in the low frequency-resolution filterbank is obtained by applying a windowing operation, i.e.,

$$\hat{\boldsymbol{\phi}}_Y[k,l](\tau)=\frac{1}{O}|H_F[\tau]|^2 P_Y^F\left[\frac{2Mk}{O}+\tau,l\right],\quad \tau=-M+1,-M+2,...,M,\quad (18)$$

where $\hat{\boldsymbol{\phi}}_Y[k,l](\tau)$ denotes the τ-th element of $\hat{\boldsymbol{\phi}}_Y[k,l]$ and H^F denotes the Fourier transform of the zero-padded analysis window h_K of length $F=2MK/O$. Using the Wiener-Khinchin theorem, an estimate of the M-dimensional noisy speech correlation vector $\hat{\boldsymbol{r}}_{\boldsymbol{y}}^{\mathrm{WK}}[k,l]$ is obtained by applying the IDFT to $\hat{\boldsymbol{\phi}}_Y[k,l]$ in (18), i.e.,

$$\hat{\boldsymbol{r}}_{\boldsymbol{y}}^{\mathrm{WK}}[k,l](m)=\frac{1}{2M}\sum_{\tau=-M+1}^{M}\hat{\boldsymbol{\phi}}_Y[k,l](\tau)e^{-j2\pi\tau m/2M},\quad m=0,1,...,M-1.\quad (19)$$

Similarly, it was proposed in [6] to estimate the speech correlation vector $\boldsymbol{r_s}$ using an estimate of the speech periodogram P_S^F in the high frequency-resolution filterbank. The speech periodogram can be estimated by applying power subtraction to the noisy speech periodogram, i.e.,

$$\hat{P}_S^F=\hat{G}^F P_Y^F,\quad (20)$$

where the Wiener gain $\hat{G}^F$ is computed as

$$\hat{G}^F=\frac{\hat{\xi}^F}{\hat{\xi}^F+1},\quad (21)$$

with $\hat{\xi}^F$ denoting an estimate of the a-priori SNR. The a-priori SNR is estimated using the decision-directed approach (DDA) [9] with the noise PSD estimator in [10]. Replacing P_Y^F with $\hat{P}_S^F$ in (18), followed by (19), yields an estimate of the speech correlation vector $\hat{\boldsymbol{r}}_{\boldsymbol{s}}^{\mathrm{WK}}$.

To estimate the noise correlation vector $\boldsymbol{r_n}$, we propose a similar approach, namely by estimating the noise periodogram P_N^F in the high frequency-resolution filterbank as

$$\hat{P}_N^F=(1-\hat{G}^F)P_Y^F.\quad (22)$$

Replacing P_Y^F with $\hat{P}_N^F$ in (18), followed by (19), yields an estimate of the noise correlation vector $\hat{\boldsymbol{r}}_{\boldsymbol{n}}^{\mathrm{WK}}$.

V. Parameter Estimation

In this section, we present several existing and novel estimators for the required parameters of the MFWF, specifically, estimating the parameters *directly* in the low frequency-resolution filterbank, *indirectly* using the estimated correlation vectors from Section IV or in a *combined* approach.

A. Noisy Speech Correlation Matrix

The noisy speech correlation matrix $\boldsymbol{R_y}$ can be estimated *directly* using first-order recursive smoothing as

$$\boxed{\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\mathrm{RS}}[k,l]=\lambda\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\mathrm{RS}}[k,l-1]+(1-\lambda)\boldsymbol{y}[k,l]\boldsymbol{y}^H[k,l]}\quad (23)$$

with λ a forgetting factor.

Alternatively, since $\boldsymbol{R_y}$ can be assumed to be a Hermitian Toeplitz matrix, this matrix can be fully defined by the noisy speech correlation vector $\boldsymbol{r_y}$. Hence, using the estimate in (19), the noisy speech correlation matrix can be estimated *indirectly* as in [6]

$$\boxed{\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\mathrm{WK}}=\mathrm{Toeplitz}\big(\hat{\boldsymbol{r}}_{\boldsymbol{y}}^{\mathrm{WK}}\big)}\quad (24)$$

B. Normalized Speech Correlation Vector

Based on (13), a maximum-likelihood estimator has been proposed in [4] to *directly* estimate the normalized speech correlation vector $\boldsymbol{\gamma_s}$ by replacing the normalized noise correlation vector $\boldsymbol{\gamma_n}$ with a long-term estimate $\hat{\mu}_{\boldsymbol{\gamma_n}}$, i.e.,

$$\boxed{\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\diamond\text{-ML}}=\frac{\hat{\phi}_S+\hat{\phi}_N}{\hat{\phi}_S}\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}^{\diamond}-\frac{\hat{\phi}_N}{\hat{\phi}_S}\hat{\mu}_{\boldsymbol{\gamma_n}}}\quad (25)$$

where $\hat{\mu}_{\boldsymbol{\gamma_n}}$ is defined by the STFT frame overlap and the analysis window h_K [4]. The parameter $\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}^{\diamond}$ denotes an estimate of $\boldsymbol{\gamma_y}$, computed similarly to (11) using $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\diamond}$, where $\diamond$ denotes either (23) or (24). The parameters $\hat{\phi}_N$ and $\hat{\phi}_S$ are estimates of the noise and speech PSDs in the low frequency-resolution filterbank, respectively. The noise PSD is estimated using [10] and the speech PSD is estimated similarly as the speech periodogram in the high frequency-resolution filterbank in (20), by applying a Wiener gain G to the noisy PSD, i.e., $\hat{\phi}_S=G\,\boldsymbol{e}^T\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\diamond}\boldsymbol{e}$.

Alternatively, it has been proposed in [6] to *indirectly* estimate the normalized speech correlation vector based on

the estimated speech correlation vector $\hat{\boldsymbol{r}}_{\boldsymbol{s}}^{\text{WK}}$ from Section IV, i.e.,

$$\boxed{\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{WK}} = \frac{\hat{\boldsymbol{r}}_{\boldsymbol{s}}^{\text{WK}}}{\hat{\boldsymbol{r}}_{\boldsymbol{s}}^{\text{WK}}(1)}} \tag{26}$$

In addition, in this paper we propose a *combined* estimator by replacing the (time- and frequency-independent) long-term estimate $\hat{\mu}_{\boldsymbol{\gamma_n}}$ by the (time- and frequency-dependent) estimated noise correlation vector $\hat{\boldsymbol{r}}_{\boldsymbol{n}}^{\text{WK}}$ from Section IV, normalized similarly as in (26), i.e.,

$$\boxed{\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\diamond\text{-WK}} = \frac{\hat{\phi}_S + \hat{\phi}_N}{\hat{\phi}_S}\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}^{\diamond} - \frac{\hat{\phi}_N}{\hat{\phi}_S}\frac{\hat{\boldsymbol{r}}_{\boldsymbol{n}}^{\text{WK}}}{\hat{\boldsymbol{r}}_{\boldsymbol{n}}^{\text{WK}}(1)}} \tag{27}$$

C. Signal Output PSD

The MFMPDR filter is designed to prevent speech distortion such that in theory, the speech PSD at the output of the MFMPDR filter is equal to the input speech PSD. Using this assumption, the speech PSD estimate $\hat{\phi}_S$ can be applied at the output of the MFMPDR filter.

The signal PSD ϕ_Y^{out} at the output of the MFMPDR filter can be estimated in several ways. The most straightforward way is to estimate this PSD as

$$\boxed{\hat{\phi}_Y^{\text{out},\boldsymbol{R_y}} = \left(\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^H \hat{\boldsymbol{R}}_{\boldsymbol{y}}^{-1} \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}\right)^{-1}} \tag{28}$$

using the previously presented estimators for $\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}$ and $\hat{\boldsymbol{R}}_{\boldsymbol{y}}$.

Alternatively, by substituting (10) into (28) and applying the matrix inversion lemma, the signal output PSD can be estimated as

$$\boxed{\hat{\phi}_Y^{\text{out},\boldsymbol{R_u}} = \hat{\phi}_S + \left(\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^H \hat{\boldsymbol{R}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}\right)^{-1}} \tag{29}$$

with $\left(\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^H \hat{\boldsymbol{R}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}\right)^{-1}$ an estimate of the undesired output PSD ϕ_U^{out}. According to (10), the correlation matrix $\hat{\boldsymbol{R}}_{\boldsymbol{u}}$ can be determined as $\hat{\boldsymbol{R}}_{\boldsymbol{u}} = \hat{\boldsymbol{R}}_{\boldsymbol{y}} - \hat{\phi}_S \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}} \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^H$. Since $\hat{\boldsymbol{R}}_{\boldsymbol{u}}$ may not be positive semi-definite due to estimation errors, we set the negative eigenvalues of $\hat{\boldsymbol{R}}_{\boldsymbol{u}}$ to zero.

The undesired output PSD can also be estimated independently of the previously estimated parameters. Since $\phi_U = \phi_N$, we can replace $\hat{\phi}_U^{\text{out}}$ in (29) with an estimate of the noise output PSD $\hat{\phi}_N^{\text{out}}$, i.e,

$$\boxed{\hat{\phi}_Y^{\text{out},\phi_N} = \hat{\phi}_S + \hat{\phi}_N^{\text{out}}} \tag{30}$$

where $\hat{\phi}_N^{\text{out}}$ is determined at the output of the MFMPDR filter using [10].

VI. Experimental Results

In this section, we compare the performance of the different presented estimators for the parameters of the MFWF. In Section VI-A we describe the used speech and noise material and the algorithmic implementation details. In Section VI-B we compare the performance of different MFMPDR filters. Using the best MFMPDR filter, in Section VI-C we compare the performance of different Wiener postfilters.

TABLE I
Overview of the evaluated parameter estimators in the MFMPDR filter.

Label	Estimation of $\boldsymbol{R}_y$	Estimation of $\boldsymbol{\gamma_s}$
$\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{RS-ML}}$	(23)	$\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\diamond\text{-ML}}$: (25) using $\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}^{\text{RS}}$: (23)
$\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{WK}}$	(23)	$\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{WK}}$: (26)
$\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{RS-WK}}$	(23)	$\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\diamond\text{-WK}}$: (27) using $\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}^{\text{RS}}$: (23)
$\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{WK}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{WK-ML}}$	(24)	$\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\diamond\text{-ML}}$: (25) using $\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}^{\text{WK}}$: (24)
$\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{WK}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{WK}}$	(24)	$\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{WK}}$: (26)
$\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{WK}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{WK-WK}}$	(24)	$\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\diamond\text{-WK}}$: (27) using $\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}^{\text{WK}}$: (24)

A. Algorithm Implementation and Performance Measures

The performance is evaluated in terms of the perceptual evaluation of speech quality (PESQ) [11] improvement over the noisy speech signal, using the clean speech signal as the reference signal. We used audio material from [12] sampled at 16 kHz and we evaluated the average performance over 176 s of speech material (92 s female, 84 s male) under five different noise conditions (babble, white Gaussian noise (WGN), traffic, modulated WGN, crossroad) at three different input SNRs (0, 5, and 10 dB).

To achieve a high speech correlation, we use a frame length of $K = 64$ samples (4 ms) and a frame shift of 16 samples (1 ms) in the low frequency-resolution STFT filterbank. As analysis and synthesis window h_K we use a Hann window. In the high frequency-resolution STFT filterbank, we use a four-times higher frequency-resolution, i.e. a frame length of $F = 256$ samples (16 ms), a frame shift of 16 samples (1 ms), and apply an asymmetric analysis window similarly to [6]. The number of the consecutive time-frames is $M = 8$ as in [6], resulting in 11 ms of analysis data in the low frequency-resolution filterbank. In the high and low frequency-resolution filterbanks, the weighting parameter for the DDA [9] is set to 0.97 and to reduce the amount of musical noise the Wiener gain is limited to -17 dB. The forgetting factor in (23) is experimentally set to $\lambda = 0.9$, resulting in a smoothing window of 10 ms. Before computing the inverse of $\hat{\boldsymbol{R}}_{\boldsymbol{y}}$ estimated using (23) or (24) regularization based on diagonal loading is performed with a regularization parameter of 0.04 as in [4].

B. Comparison of MFMPDR Filters

In this section, we compare the performance of the MFMPDR filter for different direct, indirect, and combined estimators of the noisy speech correlation matrix $\boldsymbol{R_y}$ and the normalized speech correlation vector $\boldsymbol{\gamma_s}$ (see Table I).

For different SNRs, Fig. 2 depicts the average PESQ improvements of the considered MFMPDR filters. Using the direct estimate $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}}$ in (23) leads to a higher PESQ improvement than using the indirect estimate $\hat{\boldsymbol{R}}_{\boldsymbol{s}}^{\text{WK}}$ in (24). In combination with either $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}}$ or $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{WK}}$, using the indirect estimate of the normalized speech correlation vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\diamond\text{-WK}}$ in (26) leads to a degradation (or minor improvement) of the PESQ scores. The best PESQ improvements are obtained by combining the direct estimate $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}}$ in (23) with the proposed combined estimate $\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{RS-WK}}$ in (27). This suggest that using the (time- and

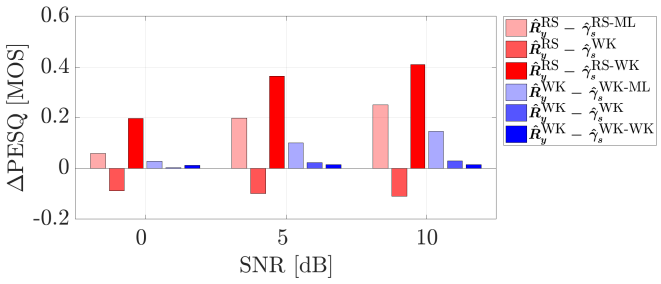

Fig. 2. Average PESQ improvement of the MFMPDR filters using the parameters in Table I at different SNRs.

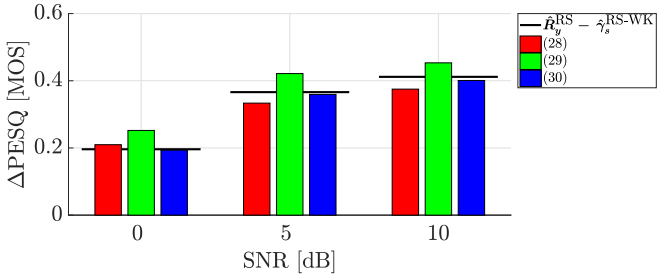

Fig. 3. Average PESQ improvement of the postfilters using the parameters from Section V-C at different SNRs, in reference to the MFMPDR filter $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{RS-WK}}$.

frequency-dependent) indirect estimate $\hat{\boldsymbol{r}}_{\boldsymbol{n}}^{\text{WK}}$ is more effective to estimate the normalized speech correlation vector than using the (time- and frequency-independent) long-term estimate $\hat{\mu}_{\boldsymbol{\gamma_n}}$.

C. Comparison of Single-Frame Wiener Postfilters

In this section we compare the performance of different Wiener postfilters using the signal output PSD estimators from Section V-C at the output of the best MFMPDR filter, i.e. using $\hat{\boldsymbol{R}}_{\boldsymbol{y}}^{\text{RS}}$ and $\hat{\boldsymbol{\gamma}}_{\boldsymbol{s}}^{\text{RS-WK}}$. For different SNRs, Fig. 3 depicts the average PESQ improvements of the different parameter estimators in the Wiener postfilter, in reference to the MFMPDR filter. From these results it can be observed that the best PESQ improvements are obtained using the signal output PSD estimate $\hat{\phi}_{Y}^{\text{out},\boldsymbol{R_u}}$ in (29). It can also be observed that independently estimating the undesired PSD at the output of the MFMPDR filter, i.e., using $\hat{\phi}_{Y}^{\text{out},\phi_N}$ in (30), does not yield a PESQ improvement compared to the MFMPDR filter. This can presumably be explained by the used noise PSD estimator [10], which assumes that the noise is more stationary than the speech, whereas the output of the MFMPDR filter may contain highly fluctuating residual noise leading to an underestimation of the noise output PSD.

VII. CONCLUSION

In this paper, we compare the noise reduction performance of different estimators for the required parameters of the single-microphone MFMPDR filter with single-frame Wiener postfilter, i.e., the noisy speech correlation matrix, the speech correlation vector and the signal output PSD. Existing estimation methods estimate the required parameters either directly in the low frequency-resolution STFT filterbank or indirectly by estimating periodograms in a high frequency-resolution filterbank and applying the Wiener-Khinchin theorem. We compare these existing estimators with proposed combined estimators using both the low and the high frequency-resolution filterbank. The results show that combining the direct estimate of the noisy speech correlation matrix with the proposed combined estimate of the speech correlation vector achieves the best speech quality improvement. Furthermore, the results show that with the Wiener postfilter the speech quality can be improved in reference to the MFMPDR filter. However, when estimating the signal output PSD independently to the output of the MFMPDR filter, no speech quality improvement can be predicted.

REFERENCES

[1] J. Benesty, J. Chen, and E. A. P. Habets, *Speech enhancement in the STFT domain*. Springer Science & Business Media, 2011.

[2] T. Esch and P. Vary, "Modified Kalman Filter Exploiting Interframe Correlation of Speech and Noise Magnitudes," in *Proc. of Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Seattle, WA, USA, Sep. 2008.

[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 and S. Doclo, "Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement," in *Proc. of Europ. Signal Process. Conf. (EUSIPCO)*, Kos, Greece, Aug. 2017, pp. 603–607.

[6] K. T. Andersen and M. Moonen, "Robust speech-distortion weighted interframe Wiener filters for single-channel noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 97–107, Jan. 2018.

[7] D. Fischer and S. Doclo, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *Proc. of Int. Workshop Acoustic Echo, Noise Control (IWAENC)*, Tokyo, Japan, Sep. 2018, pp. 41–45.

[8] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 201–205.

[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] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[11] "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.

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