SUBSPACE-BASED SPEECH CORRELATION VECTOR ESTIMATION FOR SINGLE-MICROPHONE MULTI-FRAME MVDR FILTERING

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

Aiming at exploiting the speech correlation across consecutive time-frames in the short-time Fourier transform domain, the multi-frame minimum variance distortionless response (MFMVDR) filter for single-microphone speech enhancement has been proposed. This filter is designed to avoid speech distortion while minimizing the total signal output power. To compute the MFMVDR filter, an estimate of the highly time-varying normalized speech correlation vector is required. In this paper, we propose a subspace-based estimator for the normalized speech correlation vector based on the Q largest eigenvalues and their corresponding eigenvectors of the prewhitened noisy speech correlation matrix. Experimental results for different speech signals, noise types and signal-to-noise ratios show that the proposed subspace-based estimator yields the best results in terms of speech quality and noise reduction compared to a state-of-the-art maximum-likelihood estimator.

Index Terms— MVDR, subspace estimation, interframe speech correlation, speech enhancement

1. INTRODUCTION

Speech enhancement algorithms for communication devices (e.g., hearing aids, mobile phones) are crucial to improve speech quality and speech intelligibility in noisy acoustic environments. Single-microphone speech enhancement algorithms are often implemented in the short-time Fourier transform (STFT) domain [1, 2].

To estimate the desired speech signal, on the one hand single-frame approaches such as the Wiener gain (WG) can be used, where a (real-valued) gain is applied to each noisy STFT coefficient [2]. On the other hand, multi-frame approaches such as the multi-frame minimum variance distortionless response (MFMVDR) filter [3], aim at exploiting speech correlation across consecutive time-frames by applying a (complex-valued) finite impulse response (FIR) filter to the noisy STFT coefficients [3, 4, 5, 6, 7].

The MFMVDR filter aims at minimizing the total signal output power while not distorting correlated speech components [3]. It requires estimates of the noisy speech correlation matrix and the highly time-varying *normalized speech correlation vector*, which contains the speech correlation between the current and previous time-frames. In [8] it was shown that the MFMVDR filter is more sensitive to estimation errors in the normalized speech correlation vector compared to estimation errors in the noisy speech correlation matrix. In [5], a maximum-likelihood (ML) estimator for the normalized speech correlation vector was proposed using a fixed (i.e., time-frequency-independent) mean normalized noise correlation vector. In [6], it was proposed to estimate the normalized speech correlation vector based on the noisy speech and speech periodograms in a high frequency-resolution filterbank and applying the Wiener-Khinchin theorem. In this paper, we propose a subspace-based estimator for the normalized speech correlation vector based on the Q largest eigenvalues and their corresponding eigenvectors of the prewhitened noisy speech correlation matrix. The prewhitening transform is performed using a (frequency-dependent) pretrained normalized noise correlation matrix. The dimension of the subspace Q is estimated per time-frequency point.

Experimental results for different speech signals, noise types, and signal-to-noise ratios (SNRs) show that the proposed subspace-based estimator keeps speech distortion as low as the ML estimator but improves the amount of noise reduction, leading to an increased speech quality. Moreover, the MFMVDR filter using the proposed subspace-based estimator yields a better speech quality than the traditional WG.

2. PROBLEM FORMULATION

Consider a single-microphone system, where a speech signal is degraded by additive noise. In the STFT domain, the (complex-valued) noisy speech STFT coefficient $Y(k,m)$ at frequency-bin k and time-frame m is given by

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

with $X(k,m)$ the speech STFT coefficient and $N(k,m)$ the noise STFT coefficient. The L-dimensional multi-frame noisy speech vector $\boldsymbol{y}(k,m)$ is defined as

$$\boldsymbol{y}(k,m) = \left[Y(k,m), Y(k,m-1), ..., Y(k,m-L+1)\right]^T, \tag{2}$$

where $[\cdot]^T$ denotes the transpose operator. Using (1), the noisy speech vector $\boldsymbol{y}(k,m)$ can be written as

$$\boldsymbol{y}(k,m) = \boldsymbol{x}(k,m) + \boldsymbol{n}(k,m), \tag{3}$$

where the speech vector $\boldsymbol{x}(k,m)$ and the noise vector $\boldsymbol{n}(k,m)$ are defined similarly as in (2).

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

The speech STFT coefficient $X(k,m)$ is estimated by applying a (complex-valued) FIR filter $\boldsymbol{h}(k,m)$ to the noisy speech vector, i.e,

$$\hat{X}(k,m)=\boldsymbol{h}^H(k,m)\boldsymbol{y}(k,m), \tag{4}$$

where ^{H} denotes the Hermitian operator and $\boldsymbol{h}(k,\,m)$ contains the L time-varying filter coefficients, i.e., $\boldsymbol{h}(k,\,m) = \left[H_0(k,m),H_1(k,m),...,H_{L-1}(k,m)\right]^T$. For conciseness, in the remainder of the paper the indices k and m will be omitted.

Assuming that the speech and noise signals are uncorrelated, i.e., $\mathbb{E}\left[\boldsymbol{x}\boldsymbol{n}^H\right]=0$, with $\mathbb{E}[\cdot]$ the expectation operator, the $L\times L$-dimensional noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}=\mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^H\right]$ is given by

$$\boldsymbol{R}_{\boldsymbol{y}}=\boldsymbol{R}_{\boldsymbol{x}}+\boldsymbol{R}_{\boldsymbol{n}}, \tag{5}$$

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

To exploit the speech correlation across time-frames, it has been proposed in [3] to decompose the speech vector $\boldsymbol{x}$ into the temporally correlated speech component $\boldsymbol{s}$ and the temporally uncorrelated speech component $\boldsymbol{x}'$ with respect to the speech STFT coefficient X, i.e.,

$$\boldsymbol{x}=\boldsymbol{s}+\boldsymbol{x}'=\gamma_{\boldsymbol{x}}X+\boldsymbol{x}'. \tag{6}$$

The *normalized speech correlation vector* $\gamma_{\boldsymbol{x}}$ is defined as

$$\boxed{\gamma_{\boldsymbol{x}}=\frac{\mathbb{E}[\boldsymbol{x}X^*]}{\mathbb{E}[|X|^2]}=\frac{\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{e}}=\boldsymbol{\Gamma}_{\boldsymbol{x}}\boldsymbol{e}} \tag{7}$$

where * denotes the complex-conjugate operator and $\boldsymbol{e}=\left[1,\,0,\,...,\,0\right]^T$ is an L-dimensional selection vector. Due to the normalization term $\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{e}$, which corresponds to the speech power spectral density (PSD) $\phi_X=\mathbb{E}\left[|X|^2\right]$, the first element of $\gamma_{\boldsymbol{x}}$ is equal to 1.

Substituting (6) into (3) we obtain the *multi-frame signal model*

$$\boxed{\boldsymbol{y}=\gamma_{\boldsymbol{x}}X+\boldsymbol{x}'+\boldsymbol{n}} \tag{8}$$

where we consider the uncorrelated speech component $\boldsymbol{x}'$ as an interference.

The normalized speech correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ in (7) is defined as

$$\boldsymbol{\Gamma}_{\boldsymbol{x}}=\frac{\boldsymbol{R}_{\boldsymbol{x}}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{e}}. \tag{9}$$

Using (5) and (6), the speech correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$ can be decomposed as the rank-1 correlation matrix $\boldsymbol{R}_{\boldsymbol{s}}=\phi_X\gamma_{\boldsymbol{x}}\gamma_x^H$ and the correlation matrix $\boldsymbol{R}_{\boldsymbol{x}'}=\mathbb{E}\left[\boldsymbol{x}'\boldsymbol{x}'^H\right]$, whose first row and column are equal to 0. Hence, the normalized speech correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{x}}$ is equal to

$$\boldsymbol{\Gamma}_{\boldsymbol{x}}=\frac{\boldsymbol{R}_{\boldsymbol{s}}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{e}}+\frac{\boldsymbol{R}_{\boldsymbol{x}'}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{e}}=\gamma_{\boldsymbol{x}}\gamma_{\boldsymbol{x}}^H+\boldsymbol{\Gamma}_{\boldsymbol{x}'}. \tag{10}$$

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

$$\gamma_{\boldsymbol{y}}=\frac{\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{e}}=\boldsymbol{\Gamma}_{\boldsymbol{y}}\boldsymbol{e}, \qquad \gamma_{\boldsymbol{n}}=\frac{\boldsymbol{R}_{\boldsymbol{n}}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{n}}\boldsymbol{e}}=\boldsymbol{\Gamma}_{\boldsymbol{n}}\boldsymbol{e} \tag{11}$$

where $\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{e}$ and $\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{n}}\boldsymbol{e}$ correspond to the noisy speech PSD $\phi_Y=\mathbb{E}\left[|Y|^2\right]$ and the noise PSD $\phi_N=\mathbb{E}\left[|N|^2\right]$, respectively. The normalized noisy speech correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{y}}$ and the normalized noise correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{n}}$ are defined similarly as in (9), i.e.,

$$\boldsymbol{\Gamma}_{\boldsymbol{y}}=\frac{\boldsymbol{R}_{\boldsymbol{y}}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{e}}, \qquad \boldsymbol{\Gamma}_{\boldsymbol{n}}=\frac{\boldsymbol{R}_{\boldsymbol{n}}}{\boldsymbol{e}^T\boldsymbol{R}_{\boldsymbol{n}}\boldsymbol{e}}. \tag{12}$$

The MFMVDR filter [3] is designed to minimize the total signal output power while not distorting the correlated speech component, i.e.,

$$\min_{\boldsymbol{h}}\ \boldsymbol{h}^H\boldsymbol{R}_{\boldsymbol{y}}\boldsymbol{h}, \quad \text{s.t. } \boldsymbol{h}^H\gamma_{\boldsymbol{x}}=1. \tag{13}$$

Solving this optimization problem yields the MFMVDR filter [3]

$$\boxed{\boldsymbol{h}_{\text{MFMVDR}}=\frac{\boldsymbol{R}_{\boldsymbol{y}}^{-1}\gamma_{\boldsymbol{x}}}{\gamma_{\boldsymbol{x}}^H\boldsymbol{R}_{\boldsymbol{y}}^{-1}\gamma_{\boldsymbol{x}}}} \tag{14}$$

To compute the MFMVDR filter, estimates of the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$ and the normalized speech correlation vector $\gamma_{\boldsymbol{x}}$ are required. While $\boldsymbol{R}_{\boldsymbol{y}}$ can be directly estimated from the noisy speech signal, e.g., using recursive smoothing, the highly time-varying $\gamma_{\boldsymbol{x}}$ is typically difficult to estimate accurately [8]. In this paper, we propose a new subspace-based method to estimate this vector.

3. NORMALIZED SPEECH CORRELATION VECTOR ESTIMATION

In this section, we describe two methods to estimate the normalized speech correlation vector. In Section 3.1, we review the state-of-the-art ML estimator [5] that uses a fixed (time-frequency-independent) mean normalized noise correlation vector. In Section 3.2, we propose a subspace-based estimator that uses a pretrained (frequency-dependent) estimate of the normalized noise correlation matrix.

3.1. Maximum-likelihood Estimator [5]

Using (5), (7) and (11) it can be easily shown that

$$\gamma_{\boldsymbol{y}}=\frac{\xi}{\xi+1}\gamma_{\boldsymbol{x}}+\frac{1}{\xi+1}\gamma_{\boldsymbol{n}}, \tag{15}$$

with $\xi=\phi_X/\phi_N$ the a-priori SNR.

Using (15), a ML estimator for the normalized speech correlation vector $\gamma_{\boldsymbol{x}}$ has been proposed in [5] by replacing the normalized noise correlation vector $\gamma_{\boldsymbol{n}}$ with its (time-frequency-independent) mean vector $\gamma_{\boldsymbol{n}}^{\text{mean}}$, i.e.,

$$\boxed{\hat{\gamma}_{\boldsymbol{x}}^{\text{ML}}=\frac{\hat{\xi}+1}{\hat{\xi}}\hat{\gamma}_{\boldsymbol{y}}-\frac{1}{\hat{\xi}}\gamma_{\boldsymbol{n}}^{\text{mean}}} \tag{16}$$

with $\hat{\xi}$ an estimate of the a-priori SNR and $\hat{\gamma}_{\boldsymbol{y}}$ an estimate of the normalized noisy speech correlation vector in (11). The constant vector $\gamma_{\boldsymbol{n}}^{\text{mean}}$ is determined by the frame overlap and the STFT analysis window [5].

3.2. Proposed Subspace-based Estimator

In this section, we propose a subspace-based method to estimate the normalized speech correlation vector based on the eigenvalue decomposition (EVD) of the prewhitened noisy speech correlation matrix. The prewhitening transform is performed using a time-independent but frequency-dependent pretrained normalized noise correlation matrix.

Similarly to the vector formulation in (15), the normalized noisy speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{y}}$ can be written as

$$\mathbf{\Gamma}_{\boldsymbol{y}} = \frac{\xi}{\xi+1}\mathbf{\Gamma}_{\boldsymbol{x}} + \frac{1}{\xi+1}\mathbf{\Gamma}_{\boldsymbol{n}}. \tag{17}$$

Let us first decompose the normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}$ using the Cholesky decomposition, i.e.,

$$\mathbf{\Gamma}_{\boldsymbol{n}} = \boldsymbol{C}\boldsymbol{C}^H, \tag{18}$$

with $\boldsymbol{C}$ an $L \times L$-dimensional lower triangular matrix. Using (18), the prewhitened normalized noisy speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{y}}^w$ is defined as

$$\mathbf{\Gamma}_{\boldsymbol{y}}^w = \boldsymbol{C}^{-1}\mathbf{\Gamma}_{\boldsymbol{y}}\boldsymbol{C}^{-H}. \tag{19}$$

By substituting (17) in (19), we obtain

$$\mathbf{\Gamma}_{\boldsymbol{y}}^w = \frac{\xi}{\xi+1}\mathbf{\Gamma}_{\boldsymbol{x}}^w + \frac{1}{\xi+1}\boldsymbol{I}, \tag{20}$$

with $\mathbf{\Gamma}_{\boldsymbol{x}}^w$ the prewhitened normalized speech correlation matrix and $\boldsymbol{I}$ the $L \times L$-dimensional identity matrix. Let the EVD of $\mathbf{\Gamma}_{\boldsymbol{y}}^w$ be given by

$$\mathbf{\Gamma}_{\boldsymbol{y}}^w = \boldsymbol{V}\mathbf{\Lambda}_{\boldsymbol{y}}^w\boldsymbol{V}^H = \sum_{q=1}^{L}\lambda_{\boldsymbol{y},q}^w\boldsymbol{v}_q\boldsymbol{v}_q^H, \tag{21}$$

where the columns of $\boldsymbol{V}$ contain the orthogonal eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_L$, and the diagonal elements of $\mathbf{\Lambda}_{\boldsymbol{y}}^w$ are the corresponding noisy speech eigenvalues $\lambda_{\boldsymbol{y},1}^w \geq \lambda_{\boldsymbol{y},2}^w \geq \ldots \geq \lambda_{\boldsymbol{y},L}^w$. Due to (20), the EVD of $\mathbf{\Gamma}_{\boldsymbol{x}}^w$ is given by

$$\mathbf{\Gamma}_{\boldsymbol{x}}^w = \boldsymbol{V}\mathbf{\Lambda}_{\boldsymbol{x}}^w\boldsymbol{V}^H = \sum_{q=1}^{L}\lambda_{\boldsymbol{x},q}^w\boldsymbol{v}_q\boldsymbol{v}_q^H, \tag{22}$$

with the diagonal elements of $\mathbf{\Lambda}_{\boldsymbol{x}}^w$ are equal to the speech eigenvalues $\lambda_{\boldsymbol{x},1}^w \geq \lambda_{\boldsymbol{x},2}^w \geq \ldots \geq \lambda_{\boldsymbol{x},L}^w$. The speech eigenvalues are hence related to the noisy speech eigenvalues as

$$\lambda_{\boldsymbol{x},q}^w = \frac{\xi+1}{\xi}\lambda_{\boldsymbol{y},q}^w - \frac{1}{\xi}, \qquad q = 1, \ldots, L. \tag{23}$$

Hence, using (22) the normalized speech correlation matrix $\mathbf{\Gamma}_{\boldsymbol{x}}^w$ can be written using the eigenvalues and eigenvectors of $\mathbf{\Gamma}_{\boldsymbol{y}}^w$ as

$$\mathbf{\Gamma}_{\boldsymbol{x}} = \boldsymbol{C}\left(\sum_{q=1}^{L}\left(\frac{\xi+1}{\xi}\lambda_{\boldsymbol{y},q}^w - \frac{1}{\xi}\right)\boldsymbol{v}_q\boldsymbol{v}_q^H\right)\boldsymbol{C}^H \tag{24}$$

Assuming that speech signals can be described by a low-rank model [9, 10, 11, 12] of rank-Q, we propose to estimate $\mathbf{\Gamma}_{\boldsymbol{x}}$ as

$$\boxed{\hat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{Q}} = \hat{\boldsymbol{C}}\left(\sum_{q=1}^{\hat{Q}}\hat{\lambda}_{\boldsymbol{x},q}^w\hat{\boldsymbol{v}}_q\hat{\boldsymbol{v}}_q^H\right)\hat{\boldsymbol{C}}^H} \tag{25}$$

where $\hat{Q} \leq L$ is the estimated dimension of the speech subspace and $\hat{\lambda}_{\boldsymbol{x},q}^w$ is an estimate of the q-th speech eigenvalue according to (23), i.e.,

$$\hat{\lambda}_{\boldsymbol{x},q}^w = \frac{\hat{\xi}+1}{\hat{\xi}}\hat{\lambda}_{\boldsymbol{y},q}^w - \frac{1}{\hat{\xi}}, \quad q = 1, \ldots, \hat{Q}, \tag{26}$$

where $\hat{\lambda}_{\boldsymbol{y},q}^w$ and $\hat{\boldsymbol{v}}_q$ denote the q-th eigenvalue and eigenvector of the estimated prewhitened noisy speech correlation matrix $\hat{\mathbf{\Gamma}}_{\boldsymbol{y}}^w = \hat{\boldsymbol{C}}^{-1}\hat{\mathbf{\Gamma}}_{\boldsymbol{y}}\hat{\boldsymbol{C}}^{-H}$, with $\hat{\boldsymbol{C}}$ the Cholesky factor of the estimated noise correlation matrix. The normalized speech correlation vector can then be estimated from $\hat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{Q}}$ as

$$\boxed{\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Q}} = \frac{\hat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{Q}}\boldsymbol{e}}{\boldsymbol{e}^T\hat{\mathbf{\Gamma}}_{\boldsymbol{x}}^{\mathrm{Q}}\boldsymbol{e}}} \tag{27}$$

Assuming that $\mathbf{\Gamma}_{\boldsymbol{x}'} = 0$ in (10), which is of course not the case in practice, $\mathbf{\Gamma}_{\boldsymbol{x}}$ becomes a rank-1 matrix, i.e. $Q = 1$, such that (27) with (25) can be simplified to

$$\boxed{\hat{\gamma}_{\boldsymbol{x}}^{1} = \frac{\hat{\boldsymbol{C}}\hat{\boldsymbol{v}}_1}{\boldsymbol{e}^T\hat{\boldsymbol{C}}\hat{\boldsymbol{v}}_1}} \tag{28}$$

Note that (28) is similar to the so-called covariance whitening method proposed in [13] for estimating the relative transfer function vector of the desired speaker in multi-channel speech enhancement.

To implement the proposed subspace-based estimator, estimates of the normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}$ (cf. (18)) and the dimension of the speech subspace Q (cf. (25)) are required. Since in practice it is rather difficult to accurately estimate $\mathbf{\Gamma}_{\boldsymbol{n}}$, we propose to use a pretrained (frequency-dependent) $\mathbf{\Gamma}_{\boldsymbol{n}}$. During training, perfect knowledge of the noise signal is available and the noise correlation matrix $\boldsymbol{R}_{\boldsymbol{n}}$ can easily be obtained using recursive smoothing. The pretrained normalized noise correlation matrix $\mathbf{\Gamma}_{\boldsymbol{n}}^{\mathrm{tr}}$ is subsequently obtained by averaging $\boldsymbol{R}_{\boldsymbol{n}}$ over all training data and normalizing the resulting matrix to its first element, similarly as in (12). To estimate the dimension of the speech subspace Q, there are several estimators in the literature, e.g., see [14]. Since most estimators have a larger variance in Q [10] when using a limited amount of data, we used an estimator similar to the one proposed in [10]. Given a threshold $\hat{\delta} = -\frac{1}{\hat{\xi}+1}\log(P_f)$ [15], with P_f the false alarm rate, the estimated noisy speech eigenvalue $\hat{\lambda}_{\boldsymbol{y},l}^w$ is assigned to the speech subspace when $\hat{\lambda}_{\boldsymbol{y},l}^w \geq \hat{\delta}$, where $\hat{Q}$ is the number that satisfies this criterion.

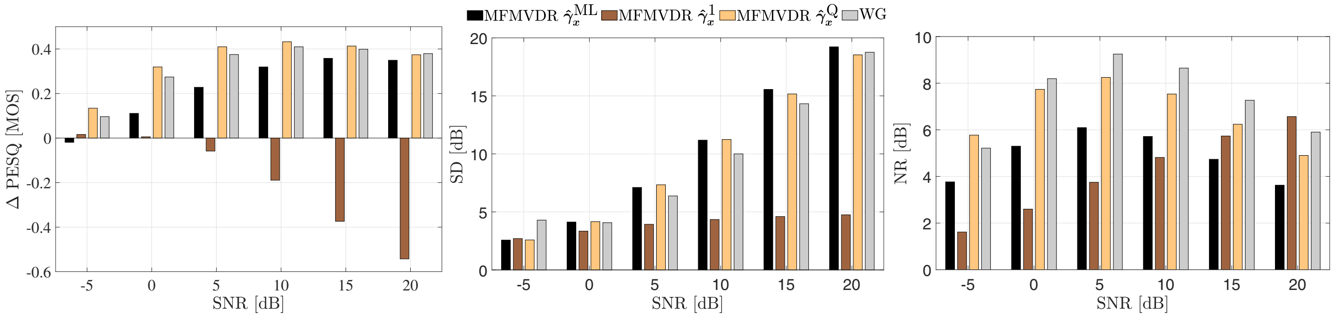

Fig. 1. Average PESQ improvement, segmental speech SNR (SD) and segmental noise reduction (NR) results of the MFMVDR filters using the state-of-the-art ML estimator $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{ML}}$ and using the proposed subspace-based estimators $\hat{\gamma}_{\boldsymbol{x}}^{1}$ and $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Q}}$ and of the WG ($L = 1$).

4. SIMULATION RESULTS

In this section, we compare the performance of the normalized speech correlation vector estimators for the MFMVDR filter either, using the ML estimator in (16) [5] or using the proposed subspace-based estimators $\hat{\gamma}_{\boldsymbol{x}}^{1}$ in (28) and $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Q}}$ in (27). As reference single-microphone speech enhancement algorithm, we use the traditional Wiener gain (WG) [2].

We used speech material from the TIMIT database [16] sampled at 16 kHz. The average performance is evaluated over 250 s of speech material (131 s female, 129 s male) under four different noise conditions (babble, modulated white Gaussian noise (WGN), traffic and speech-shaped noise) [17, 18], at input SNRs ranging from -5 dB to 20 dB. To train the normalized noise correlation matrix $\boldsymbol{\Gamma}_{\boldsymbol{n}}^{\mathrm{tr}}$ we used 5 noise types (WGN, 2 babble and 2 traffic noises) [17, 18], resulting in 230 s of noise material. We made sure that the training data differs from the evaluation data.

Similarly as in [5, 7], we used a STFT frame length of 4 ms and a frame shift of 1 ms to achieve a high speech correlation. As analysis and synthesis window we used a square-root Hann window. The number of consecutive time-frames is experimentally set to $L = 6$, resulting in 9 ms of analysis data. To estimate the noisy speech correlation matrix $\boldsymbol{R}_{\boldsymbol{y}}$, we applied recursive smoothing with a smoothing factor experimentally set to 0.92. To estimate the a-priori SNR ξ, we used the decision-directed approach (DDA) [19], where the weighting parameter is set to 0.97 and the noise PSD is estimated using the speech presence probability-based estimator in [20] with the same smoothing factor of 0.90 as for $\boldsymbol{R}_{\boldsymbol{n}}$ in the training. To reduce fluctuations in the estimation of the a-priori SNR, we only updated the estimate every 4 ms. For the WG, we used a frame length of 4 ms ($L = 1$) and an overlap of 50 $\%$. The estimation of the a-priori SNR is also performed by the DDA with a weighting parameter of 0.97 and the noise PSD estimator in [20]. To reduce the amount of speech distortion and to mask artifacts in the background noise, we apply a lower limit of -8 dB to all a-priori SNR estimates. To estimate the dimension of the speech subspace Q, we set the false-alarm rate $P_f = 0.05$.

The performance of all considered algorithms is evaluated in terms of the perceptual evaluation of speech quality (PESQ) [21] improvement compared to the noisy speech signal, using the clean speech signal as the reference signal. Furthermore, the performance is evaluated in terms of speech distortion and noise reduction using the segmental speech SNR (SD) and the segmental noise reduction (NR) [22], where both measures have been computed only during time-frames where speech is active.

Fig. 1 depicts the results averaged over all speech and noise files. First, it can be observed that the performance of the MFMVDR filter using $\hat{\gamma}_{\boldsymbol{x}}^{1}$ results in the worst performance in terms of all performance measures and SNRs. This confirms the results in [8], where it has been shown that the influence of the uncorrelated speech component is crucial, especially at high SNRs. By assuming that $\boldsymbol{\Gamma}_{\boldsymbol{x}'} = 0$ in (10), i.e., using a fixed $\hat{Q} = 1$, the amount of speech distortion increases for higher SNRs such that the speech quality is reduced. Considering the PESQ improvement, the MFMVDR filter using the proposed subspace-based estimator $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Q}}$ outperforms all other filter for SNRs up to 15 dB. Regarding SD, it can be observed that the MFMVDR filter using $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Q}}$ leads to a similar performance as the MFMVDR filter using $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{ML}}$, and for SNRs larger than 0 dB it achieves less speech distortion than the WG. In terms of NR, it can be observed that the MFMVDR filter using $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Q}}$ achieves clearly better results than the MFMVDR filter using $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{ML}}$, but for SNRs larger than 0 dB it is worse than the WG. These results indicate that determining the normalized speech correlation vector based on the Q largest speech eigenvalues keeps the speech distortion as low as the ML estimator but clearly leads to more noise reduction, overall resulting in an increased objective speech quality. Moreover, the MFMVDR filter using the proposed subspace-based estimator yields less noise reduction and speech distortion than the traditional WG, resulting in a better speech quality for SNRs up to 15 dB.

5. CONCLUSIONS

In this paper, we proposed a subspace-based normalized speech correlation estimator for the single-microphone multi-frame minimum variance distortionless response (MFMVDR) filter. We proposed to estimate the normalized speech correlation vector based on the Q largest eigenvalues and their corresponding eigenvectors of the prewhitened noisy speech correlation matrix. Simulation results show that the MFMVDR filter using the proposed subspace-based estimator leads to a better speech quality and more noise reduction than the state-of-the-art ML approach, while speech distortion are kept low. Compared to the traditional Wiener gain (WG), the MFMVDR filter using the proposed estimator leads to less speech distortion and noise reduction, resulting in a slightly better speech quality.

6. REFERENCES

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

[2] Richard C. Hendriks, Timo Gerkmann, and Jesper Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, Morgan & Claypool, 2013.

[3] Yiteng Huang and Jacob 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] Thomas Esch and Peter Vary, "Modified Kalman Filter Exploiting Interframe Correlation of Speech and Noise Magnitudes," in *Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, WA, USA, Sept. 2008.

[5] Alexander Schasse and Rainer 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, Sept. 2014.

[6] Kristian T. Andersen and Marc 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] Dörte Fischer and Simon Doclo, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *Proc. of Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, Japan, Sept. 2018, pp. 41–45.

[8] Dörte Fischer and Simon 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.

[9] Robert McAulay and Thomas Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[10] Richard C. Hendriks, Jesper Jensen, and Richard Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

[11] Jesper R. Jensen, Jacob Benesty, and Mads Græsbøll Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 4, pp. 631–644, Apr. 2016.

[12] Simon Doclo and Marc Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

[13] Shmulik Markovich, Sharon Gannot, and Israel Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[14] Petre Stoica and Yngve Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[15] Steven M. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*, vol. 2, Prentice-Hall, 1998.

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

[17] Andrew Varga and Herman 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, July 1993.

[18] Hans-Günter Hirsch and David Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA workshop on automatic speech recognition*, Paris, France, Sept. 2000.

[19] Yariv Ephraim and David 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.

[20] Timo Gerkmann and Richard 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.

[21] Philipos Loizou, *Speech Enhancement: Theory and Practice*, CRC press, 2013.

[22] Thomas Lotter and Peter Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Applied Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.