Robust Constrained MFMVDR Filters for Single-Channel Speech Enhancement Based on Spherical Uncertainty Set

Dörte Fischer, *Member, IEEE*, and Simon Doclo, *Senior Member, IEEE*

***Abstract*—Aiming at exploiting speech correlation across consecutive time-frames in the short-time Fourier transform domain, the multi-frame minimum variance distortionless response (MFMVDR) filter for single-channel speech enhancement has been proposed. The MFMVDR filter requires an accurate estimate of the normalized speech correlation vector in order to avoid speech distortion and artifacts. In this paper we investigate the potential of using robust MVDR filtering techniques to estimate the normalized speech correlation vector as the vector maximizing the total signal output power within a spherical uncertainty set, which corresponds to imposing a quadratic inequality constraint. Whereas the singly-constrained (SC) MFMVDR filter only considers the quadratic inequality constraint to estimate the (non-normalized) speech correlation vector, the doubly-constrained (DC) MFMVDR filter integrates a linear normalization constraint into the optimization problem to directly estimate the normalized speech correlation vector. To set the upper bound of the quadratic inequality constraint for each time-frequency point, we propose to use a trained non-linear mapping function that depends on the a-priori signal-to-noise ratio (SNR). Experimental results for different speech signals, noise types and SNRs show that the proposed constrained approaches yield a more accurate estimate of the normalized speech correlation vector than a state-of-the-art maximum-likelihood (ML) estimator. An instrumental and a perceptual evaluation show that both constrained MFMVDR filters lead to less speech and noise distortion but a lower noise reduction than the ML-MFMVDR filter, where the DC-MFMVDR filter is preferred in terms of overall quality compared to the SC-MFMVDR and ML-MFMVDR filters.**

***Index Terms*—Multi-Frame MVDR Filter, single-microphone speech enhancement, speech interframe correlation.**

I. Introduction

SPEECH communication devices such as hearing aids or mobile phones are often used in acoustically challenging situations, where the desired speech signal is affected by ambient noise. In these situations, speech quality and speech intelligibility may be degraded, especially at low signal-to-noise ratios (SNRs), such that speech enhancement algorithms are required to suppress the undesired noise while keeping speech distortion low [1]–[4]. Depending on the number of available microphones, both single-channel as well as multi-channel speech enhancement algorithms were proposed. In this paper we focus on single-channel algorithms in the short-time Fourier transform (STFT) domain exploiting concepts proposed for multi-channel algorithms.

In single-channel speech enhancement algorithms, it is often assumed that neighboring STFT coefficients are uncorrelated over time and frequency, which is a valid assumption when using sufficiently long analysis frames (in the order of 20–30 ms) with a relatively small overlap (in the order of 50%) [3]. Hence, an estimate of the speech STFT coefficients can be obtained by applying a (real- or complex-valued) gain to the noisy speech STFT coefficients at each time-frequency point independently. Frequently used examples are the conventional Wiener gain [2], [3] and the minimum mean-square error logarithmic short-time spectral amplitude (LogSTSA) estimator [5]. In addition, numerous approaches based on deep learning have recently been proposed to estimate the spectro-temporal gain [6]–[9]. In this paper, we will however focus on conventional speech enhancement approaches.

Since, it is more realistic to assume that neighboring STFT coefficients are correlated over time and/or frequency, especially when using short analysis frames and/or a large overlap, speech correlation across time-frames was considered in [10]–[17], while speech correlation across frequency-bands was considered in [18], [19] and speech correlation across both time-frames as well as frequency-bands was considered in [20], [21]. Based on a multi-frame signal model, where the *normalized speech correlation vector* represents the speech correlation between the current and previous time-frames, it was proposed in [1], [11] to estimate the speech STFT coefficients by applying an multi-frame minimum variance distortionless response (MFMVDR) filter to the noisy speech STFT coefficients. The MFMVDR filter aims at minimizing the total signal output power while not distorting correlated speech components. Conceptually, the single-channel MFMVDR filter is similar to the multi-channel MVDR beamformer [22], [23] when interpreting time-frames as microphone inputs and the speech correlation vector as the steering vector of the desired speech source.

The MFMVDR filter requires estimates of the noisy speech correlation matrix and the (highly time-varying) normalized

Manuscript received April 12, 2020; revised September 23, 2020; accepted November 6, 2020. Date of publication December 7, 2020; date of current version January 14, 2021. This work was supported in part by Joint Lower Saxony-Israeli Project ATHENA and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2177/1 - Project ID 390895286. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding author: Dörte Fischer.)*

The authors are with the Department of Medical Physics and Acoustics and the Cluster of Excellence Hearing4all, University of Oldenburg, 26 129 Oldenburg, Germany (e-mail: doerte.fischer@uni-oldenburg.de; simon.doclo@uni-oldenburg.de).

Digital Object Identifier 10.1109/TASLP.2020.3042013

speech correlation vector. Several approaches were proposed to estimate the normalized speech correlation vector from the noisy speech STFT coefficients. In [12] a maximum-likelihood (ML) estimator was derived based on the assumption that the normalized speech and noise correlation vectors follow multivariate Gaussian distributions, while in [14] it was proposed to estimate the normalized speech correlation vector by applying the Wiener-Khinchin theorem on estimated periodograms in a high frequency-resolution filterbank. In [24], we showed that accurately estimating the normalized speech correlation vector is crucial since even small estimation errors may lead to a degraded performance of the MFMVDR filter, causing speech distortion and unpleasant artifacts in the background noise.

In the area of multi-channel processing, i.e., beamforming, several techniques were proposed to increase the robustness of MVDR beamformers against estimation errors in the steering vector. One of the most popular approaches is diagonal loading, imposing a quadratic inequality constraint on the filter vector [25]. However, since diagonal loading does not explicitly address uncertainty of the steering vector, several other approaches were proposed, e.g., by imposing (equality and/or inequality) constraints on the so-called mismatch vector, i.e., the difference between the steering vector and the presumed steering vector [26]–[34]. The robust MVDR beamformers in [28], [30] estimate the steering vector as the vector maximizing the total signal output power of the MVDR beamformer within a spherical uncertainty set. Inspired by these robust multi-channel approaches, in this paper we investigate the potential of estimating the normalized speech correlation vector as the vector maximizing the total signal output power of the MFMVDR filter within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector, i.e., the difference between the speech correlation vector and the presumed normalized speech correlation vector, e.g., the ML estimate in [12]. We propose two constrained MFMVDR filters. The singly-constrained (SC) MFMVDR filter only considers the quadratic inequality constraint on the mismatch vector to estimate the (non-normalized) speech correlation vector and applies normalization afterwards. On the other hand, the doubly-constrained (DC) MFMVDR filter integrates the (linear) normalization constraint into the optimization problem and directly estimates the normalized speech correlation vector by solving an optimization problem with two constraints. Since oracle simulations (i.e., assuming that the speech and noise signals are available) at different SNRs show that the norm of the mismatch vector decreases with increasing SNR, we propose to train a non-linear mapping function that depends on the a-priori SNR to set the upper bound of the spherical uncertainty set for each time-frequency point. Preliminary results for the SC-MFMVDR filter were already presented in [16].

The remainder of this paper is structured as follows. In Section II, the multi-frame signal model is presented. In Section III, the MFMVDR filter is reviewed and state-of-the-art methods to estimate the noisy speech correlation matrix and the normalized speech correlation vector are discussed. In Section IV, the proposed constrained MFMVDR filters based on a spherical uncertainty set as well as the proposed mapping function to set the upper bound of the spherical uncertainty set are presented. In Section V, an instrumental and perceptual performance comparison between both constrained MFMVDR filters, the state-of-the-art ML-MFMVDR filter [12], the oracle MFMVDR filter and the LogSTSA estimator [5] is provided for different speech signals, noise types and SNRs. The results show that the proposed constrained MFMVDR filters result in a more conservative noise reduction performance with a more natural speech quality and less noise distortion than the ML-MFMVDR filter, where the DC-MFMVDR filter is preferred in terms of overall quality.

II. Multi-Frame Signal Model

We consider a single-channel setup, where the speech signal $x(t)$ is degraded by additive noise $n(t)$, with t denoting the time index. In the STFT domain, the complex-valued noisy speech STFT coefficient $Y(k,m)$ at frequency-band k and time-frame m is given by

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

with $X(k,m)$ and $N(k,m)$ denoting the speech and noise STFT coefficients, respectively. For conciseness, in the remainder of the paper the frequency-band index k will be omitted if not required.

In single-frame speech enhancement approaches [2], [3], [5], [6], [8] the speech STFT coefficient $X(m)$ is typically estimated by applying a (real-valued) gain $G(m)$ to the noisy speech STFT coefficient $Y(m)$, i.e.,

$$\widehat{X}(m) = G(m)Y(m)\,. \tag{2}$$

In multi-frame speech enhancement approaches [1], the L-dimensional noisy speech vector $\boldsymbol{y}(m)$ is defined as

$$\boldsymbol{y}(m) = \Big[Y(m), Y(m-1), \ldots, Y(m-L+1)\Big]^T\,, \tag{3}$$

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

$$\boldsymbol{y}(m) = \boldsymbol{x}(m) + \boldsymbol{n}(m)\,, \tag{4}$$

where the speech vector $\boldsymbol{x}(m)$ and the noise vector $\boldsymbol{n}(m)$ are defined similarly as $\boldsymbol{y}(m)$ in (3). The speech STFT coefficient $X(m)$ is then estimated by applying a (complex-valued) FIR filter $\boldsymbol{h}(m)$ to the noisy speech vector $\boldsymbol{y}(m)$, i.e.,

$$\widehat{X}(m) = \boldsymbol{h}^H(m)\boldsymbol{y}(m)\,, \tag{5}$$

where $[\cdot]^H$ denotes the Hermitian operator. The filter $\boldsymbol{h}(m)$ contains the L time-varying filter coefficients, i.e.,

$$\boldsymbol{h}(m) = \Big[H_0(m), H_1(m), \ldots, H_{L-1}(m)\Big]^T\,. \tag{6}$$

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

$$\boldsymbol{\Phi}_{\boldsymbol{y}}(m) = \boldsymbol{\Phi}_{\boldsymbol{x}}(m) + \boldsymbol{\Phi}_{\boldsymbol{n}}(m)\,, \tag{7}$$

where $\mathbf{\Phi}_{\boldsymbol{x}}(m) = \mathbb{E}[\boldsymbol{x}(m)\boldsymbol{x}^H(m)]$ and $\mathbf{\Phi}_{\boldsymbol{n}}(m) = \mathbb{E}[\boldsymbol{n}(m)\boldsymbol{n}^H(m)]$ denote the speech and noise correlation matrices, respectively.

To exploit the speech correlation across time-frames, it was proposed in [1], [11] to decompose the speech vector $\boldsymbol{x}(m)$ into the temporally correlated speech component $\boldsymbol{s}(m)$ and the temporally uncorrelated speech component $\boldsymbol{x}'(m)$ with respect to the speech STFT coefficient $X(m)$, i.e.,

$$\boldsymbol{x}(m) = \boldsymbol{s}(m) + \boldsymbol{x}'(m) = \boldsymbol{\gamma}_{\boldsymbol{x}}(m)X(m) + \boldsymbol{x}'(m)\,, \tag{8}$$

where the (highly time-varying) *normalized speech correlation vector* $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ is defined as

$$\boldsymbol{\gamma}_{\boldsymbol{x}}(m) = \frac{\mathbb{E}\left[\boldsymbol{x}(m)X^*(m)\right]}{\mathbb{E}\left[|X(m)|^2\right]} = \frac{\mathbf{\Phi}_{\boldsymbol{x}}(m)\boldsymbol{e}}{\boldsymbol{e}^T\mathbf{\Phi}_{\boldsymbol{x}}(m)\boldsymbol{e}}, \tag{9}$$

where * denotes the complex-conjugate operator and $\boldsymbol{e} = \begin{bmatrix}1,\ 0,\ \ldots,\ 0\end{bmatrix}^T$ is an L-dimensional selection vector. Due to the normalization term $\boldsymbol{e}^T\mathbf{\Phi}_{\boldsymbol{x}}(m)\boldsymbol{e}$, which corresponds to the speech power spectral density (PSD) $\phi_X(m) = \mathbb{E}[|X(m)|^2]$, the first element of the normalized speech correlation vector is equal to 1, i.e.,

$$\boldsymbol{e}^T\boldsymbol{\gamma}_{\boldsymbol{x}}(m) = 1\,, \tag{10}$$

which will be referred to as the normalization constraint.

Substituting (8) into (4), we obtain the *multi-frame signal model*

$$\boxed{\boldsymbol{y}(m) = \boldsymbol{\gamma}_{\boldsymbol{x}}(m)X(m) + \boldsymbol{x}'(m) + \boldsymbol{n}(m)} \tag{11}$$

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

Similarly to (9), the normalized noisy speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{y}}(m)$ and the normalized noise correlation vector $\boldsymbol{\gamma}_{\boldsymbol{n}}(m)$ are defined as

$$\boldsymbol{\gamma}_{\boldsymbol{y}}(m) = \frac{\mathbf{\Phi}_{\boldsymbol{y}}(m)\ \boldsymbol{e}}{\boldsymbol{e}^T\mathbf{\Phi}_{\boldsymbol{y}}(m)\boldsymbol{e}}, \qquad \boldsymbol{\gamma}_{\boldsymbol{n}}(m) = \frac{\mathbf{\Phi}_{\boldsymbol{n}}(m)\ \boldsymbol{e}}{\boldsymbol{e}^T\mathbf{\Phi}_{\boldsymbol{n}}(m)\boldsymbol{e}}, \tag{12}$$

with $\boldsymbol{e}^T\mathbf{\Phi}_{\boldsymbol{y}}(m)\boldsymbol{e}$ and $\boldsymbol{e}^T\mathbf{\Phi}_{\boldsymbol{n}}(m)\boldsymbol{e}$ corresponding to the noisy speech PSD $\phi_Y(m) = \mathbb{E}[|Y(m)|^2]$ and the noise PSD $\phi_N(m) = \mathbb{E}[|N(m)|^2]$, respectively. Using (7), (9) and (12), it can be easily shown that

$$\phi_Y(m)\boldsymbol{\gamma}_{\boldsymbol{y}}(m) = \phi_X(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m) + \phi_N(m)\boldsymbol{\gamma}_{\boldsymbol{n}}(m)\,, \tag{13}$$

such that the normalized speech correlation vector can be written as

$$\boldsymbol{\gamma}_{\boldsymbol{x}}(m) = \frac{\xi(m)+1}{\xi(m)}\boldsymbol{\gamma}_{\boldsymbol{y}}(m) - \frac{1}{\xi(m)}\boldsymbol{\gamma}_{\boldsymbol{n}}(m)\,, \tag{14}$$

where $\xi(m) = \phi_X(m)/\phi_N(m)$ denotes the a-priori SNR.

III. Multi-Frame MVDR Filter

In [1], [11], the MFMVDR filter for single-channel speech enhancement was proposed, which aims at minimizing the total signal output power while not distorting the correlated speech component[1], i.e.,

$$\min_{\boldsymbol{h}(m)}\ \boldsymbol{h}^H(m)\mathbf{\Phi}_{\boldsymbol{y}}(m)\boldsymbol{h}(m),\ \ \text{s.t. } \boldsymbol{h}^H(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m) = 1. \tag{15}$$

Solving this optimization problem yields the MFMVDR filter vector

$$\boxed{\boldsymbol{h}_{\text{MFMVDR}}(m) = \frac{\mathbf{\Phi}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}{\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\mathbf{\Phi}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}} \tag{16}$$

with the signal output power $\phi_Y^{\text{out}}(m)$ equal to

$$\begin{aligned}\phi_Y^{\text{out}}(m) &= \mathbb{E}\left[|\boldsymbol{h}_{\text{MFMVDR}}^H(m)\boldsymbol{y}(m)|^2\right] \\ &= \frac{1}{\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\mathbf{\Phi}_{\boldsymbol{y}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}.\end{aligned} \tag{17}$$

As can be seen from (16), the MFMVDR filter requires an estimate of the noisy speech correlation matrix $\mathbf{\Phi}_{\boldsymbol{y}}(m)$ and the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$, which need to be estimated from the noisy speech STFT coefficients. In [24] we showed 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. Hence, it is crucial to accurately estimate the normalized speech correlation vector and/or to make the MFMVDR filter more robust against estimation errors (as considered in this paper).

A. Noisy Speech Correlation Matrix

Estimating the noisy speech correlation matrix from the noisy speech STFT coefficients can be performed rather straightforwardly by applying first-order recursive smoothing, i.e.,

$$\widehat{\mathbf{\Phi}}_{\boldsymbol{y}}(m) = \alpha_y\widehat{\mathbf{\Phi}}_{\boldsymbol{y}}(m-1) + (1-\alpha_y)\boldsymbol{y}(m)\boldsymbol{y}^H(m)\,, \tag{18}$$

where $\widehat{\mathbf{\Phi}}_{\boldsymbol{y}}(m)$ denotes the estimated noisy speech correlation matrix and α_y is a smoothing parameter. As suggested in [12], to avoid numerical problems we apply diagonal loading before computing the inverse of $\widehat{\mathbf{\Phi}}_{\boldsymbol{y}}(m)$, i.e.,

$$\widehat{\mathbf{\Phi}}_{\boldsymbol{y}}^{-1}(m) = \left(\widehat{\mathbf{\Phi}}_{\boldsymbol{y}}(m) + \frac{\kappa\text{tr}\left[\widehat{\mathbf{\Phi}}_{\boldsymbol{y}}(m)\right]}{L}\boldsymbol{I}_{\text{L}}\right)^{-1}, \tag{19}$$

where κ denotes a small scaling parameter, the operator tr$[\cdot]$ denotes the trace of a matrix and $\boldsymbol{I}_{\text{L}}$ denotes the $L \times L$-dimensional identity matrix.

B. Normalized Speech Correlation Vector

When assuming the noise $\boldsymbol{n}(m)$ to be available (which is of course not the case in practice), an oracle estimate of the noise correlation matrix $\widehat{\mathbf{\Phi}}_{\boldsymbol{n}}^{\text{o}}(m)$ can be obtained similarly to (18) as

$$\widehat{\mathbf{\Phi}}_{\boldsymbol{n}}^{\text{o}}(m) = \alpha_n\widehat{\mathbf{\Phi}}_{\boldsymbol{n}}^{\text{o}}(m-1) + (1-\alpha_n)\boldsymbol{n}(m)\boldsymbol{n}^H(m)\,, \tag{20}$$

[1]Although a more appropriate name would be multi-frame minimum power distortionless response (MFMPDR) filter, we decided to keep the original terminology from [11].

where α_n is a smoothing parameter. Similarly to (12), an oracle estimate of the normalized noise correlation vector can hence be obtained as

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{n}}^{\mathrm{o}}(m) = \frac{\widehat{\boldsymbol{\Phi}}_{\boldsymbol{n}}^{\mathrm{o}}(m)\,\boldsymbol{e}}{\boldsymbol{e}^T\widehat{\boldsymbol{\Phi}}_{\boldsymbol{n}}^{\mathrm{o}}(m)\boldsymbol{e}}. \tag{21}$$

Using $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}}(m)$ in (18), an estimate of the normalized noisy speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m)$ can be obtained similarly to (12). Hence, using (14) and (21) an oracle estimate of the normalized speech correlation vector can be obtained as

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{o}}(m) = \frac{\widehat{\xi}^{\mathrm{o}}(m)+1}{\widehat{\xi}^{\mathrm{o}}(m)}\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m) - \frac{1}{\widehat{\xi}^{\mathrm{o}}(m)}\widehat{\boldsymbol{\gamma}}_{\boldsymbol{n}}^{\mathrm{o}}(m)\,, \tag{22}$$

with $\widehat{\xi}^{\mathrm{o}}(m)$ an oracle estimate of the a-priori SNR given by

$$\widehat{\xi}^{\mathrm{o}}(m) = \frac{\boldsymbol{e}^T\left(\widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}}(m) - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{n}}^{\mathrm{o}}(m)\right)\boldsymbol{e}}{\boldsymbol{e}^T\widehat{\boldsymbol{\Phi}}_{\boldsymbol{n}}^{\mathrm{o}}(m)\boldsymbol{e}}. \tag{23}$$

Since in practice the noise $\boldsymbol{n}(m)$ is obviously not available, both the normalized noise correlation vector as well as the a-priori SNR need to be estimated from the noisy speech STFT coefficients in order to be able to estimate the normalized speech correlation vector based on (14). Assuming that the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ and the normalized noise correlation vector $\boldsymbol{\gamma}_{\boldsymbol{n}}(m)$ follow multivariate complex Gaussian distributions, the ML estimate of the normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ was derived in [12] as

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}(m) = \frac{\widehat{\xi}(m)+1}{\widehat{\xi}(m)}\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m) - \frac{1}{\widehat{\xi}(m)}\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}} \tag{24}$$

with $\widehat{\xi}(m)$ an estimate of the a-priori SNR and $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$ the mean normalized noise correlation vector, which is solely determined by the frame overlap and the STFT analysis window and hence assumed constant for all time-frequency points [12].

To estimate the a-priori SNR, several estimators such as the decision-directed approach (DDA) [35] or cepstro-temporal smoothing [36] were proposed. In this paper we have used the DDA, i.e.,

$$\widehat{\xi}(m) = \beta_{\mathrm{DDA}}\frac{|\widehat{X}(m-1)|^2}{\widehat{\phi}_N(m-1)} + (1-\beta_{\mathrm{DDA}})\max\left[\frac{|Y(m)|^2}{\widehat{\phi}_N(m)} - 1, 0\right], \tag{25}$$

with β_{DDA} denoting a weighting parameter and $\widehat{\phi}_N(m)$ an estimate of the noise PSD for which we used the speech-presence-probability-based noise PSD estimator proposed in [37].

Simulation results in [12], [16], [24] showed that the accuracy of the ML estimate in (24) strongly depends on the a-priori SNR estimate $\widehat{\xi}(m)$. Especially at low a-priori SNRs, the ML estimate may become very large, such that the estimation error between $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}(m)$ and the oracle estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{o}}(m)$ becomes very large. This may cause correlated speech components to be mistakenly interpreted as uncorrelated, resulting in speech distortion and unpleasant artifacts in the background noise [12], [24], [38]. Examples of audio samples are available online (see https://uol.de/en/sigproc/research/audio-demos/multi-frame-speech-enhancement/constrained-mfmvdr-filters).

IV. Constrained MFMVDR Filters Based on Spherical Uncertainty Set

Aiming at improving the robustness against estimation errors in the normalized speech correlation vector, in this section we propose two constrained MFMVDR filters. Inspired by the robust MVDR beamformers (also called robust Capon beamformers) in [28], [30], we propose to estimate the normalized speech correlation vector as the vector maximizing the total signal output power of the MFMVDR filter within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector with respect to the presumed normalized speech correlation vector. Section IV-A presents the singly-constrained MFMVDR filter, which only considers the quadratic inequality constraint, whereas Section IV-B presents the doubly-constrained MFMVDR filter, which jointly considers the quadratic inequality constraint as well as the (linear) normalization constraint. Section IV-C discusses a trained non-linear mapping function to set the upper bound of the spherical uncertainty set for each time-frequency point. For conciseness, the time-frame index m will be omitted in this section, although it should be realized that all calculations are performed for each time-frequency point.

A. Singly-Constrained (SC) MFMVDR Filter

Given a presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$, e.g., the ML estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$ in (24), the mismatch vector with respect to the (unknown) normalized speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}$ is defined as $\boldsymbol{\delta}_{\boldsymbol{x}} = \boldsymbol{\gamma}_{\boldsymbol{x}} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$, with $\epsilon_{\boldsymbol{x}} = ||\boldsymbol{\delta}_{\boldsymbol{x}}||_2^2$. We now define the spherical uncertainty set comprising all vectors whose squared distance to the presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$ is smaller than or equal to a bound $\epsilon \geq 0$, i.e.,

$$\boldsymbol{\Gamma} = \left\{ \boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}} + \boldsymbol{\delta} \mid \|\boldsymbol{\delta}\|_2^2 \leq \epsilon \right\}. \tag{26}$$

Similarly to the robust MVDR beamformer in [28], we proposed in [16] to compute the (non-normalized) speech correlation vector for the SC-MFMVDR filter as the vector maximizing the total signal output power of the MFMVDR filter in (17) within the spherical uncertainty set in (26), i.e.,

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}} \quad \frac{1}{\boldsymbol{\gamma}^H\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}}, \quad \text{s.t. } \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 \leq \epsilon\,, \tag{27}$$

which is equivalent to

$$\boxed{\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{SC}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \quad \boldsymbol{\gamma}^H\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}, \quad \text{s.t. } \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2 \leq \epsilon} \tag{28}$$

For an exemplary noisy speech correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}$ and $L = 2$, Fig. 1 visualizes the quadratic cost function $\boldsymbol{\gamma}^H\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\boldsymbol{\gamma}$ in (28), together with an exemplary presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$ and bound ϵ. Obviously, the bound ϵ in (28) plays an important role and should be chosen in accordance with the accuracy of the presumed normalized speech correlation

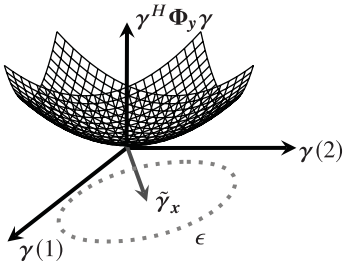


Fig. 1. Quadratic cost function in (28) with exemplary presumed normalized speech correlation vector $\tilde{\gamma}_{\boldsymbol{x}}$ and bound ϵ.

vector $\tilde{\gamma}_{\boldsymbol{x}}$, i.e., if $||\gamma_{\boldsymbol{x}} - \tilde{\gamma}_{\boldsymbol{x}}||_2^2$ is large, then ϵ should be large, whereas if $||\gamma_{\boldsymbol{x}} - \tilde{\gamma}_{\boldsymbol{x}}||_2^2$ is small, then ϵ should be small.

The minimum of the quadratic cost function $\gamma^H \boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\gamma$ is given by $\widehat{\gamma}_{\boldsymbol{x}} = 0$, which is obviously undesired. In order to avoid this solution, the bound ϵ should be chosen such that

$$\epsilon < \|\tilde{\gamma}_{\boldsymbol{x}}\|_2^2 \,. \tag{29}$$

Under this condition and considering the convex nature of the quadratic cost function in (28), the inequality constraint in (28) can be replaced by an equality constraint, i.e.,

$$\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{SC}} = \underset{\gamma}{\operatorname{argmin}} \;\; \gamma^H \boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\gamma, \;\; \text{s.t. } \|\gamma - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2 = \epsilon \,. \tag{30}$$

This constrained optimization problem can be solved using the method of Lagrange multipliers [39]. The Lagrangian function is given by

$$f^{\mathrm{SC}}(\gamma, \lambda) = \gamma^H \boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\gamma + \lambda \left(\|\gamma - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2 - \epsilon \right) , \tag{31}$$

with λ the Lagrange multiplier. Setting the gradient of $f^{\mathrm{SC}}(\gamma, \lambda)$ with respect to γ

$$\nabla_\gamma f^{\mathrm{SC}}(\gamma, \lambda) = 2\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\gamma + 2\lambda(\gamma - \tilde{\gamma}_{\boldsymbol{x}}) \tag{32}$$

equal to zero, yields

$$\gamma = \lambda \left(\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1} + \lambda \boldsymbol{I}_{\mathrm{L}} \right)^{-1} \tilde{\gamma}_{\boldsymbol{x}} \,. \tag{33}$$

Applying the matrix inversion lemma, we obtain the SC speech correlation vector $\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{SC}}(\lambda)$ as

$$\boxed{\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{SC}}(\lambda) = \tilde{\gamma}_{\boldsymbol{x}} - (\lambda \boldsymbol{\Phi}_{\boldsymbol{y}} + \boldsymbol{I}_{\mathrm{L}})^{-1} \tilde{\gamma}_{\boldsymbol{x}}} \tag{34}$$

Setting the partial derivative of $f^{\mathrm{SC}}(\gamma, \lambda)$ in (31) with respect to λ equal to zero and substituting (34) results in

$$g^{\mathrm{SC}}(\lambda) = \frac{\partial f^{\mathrm{SC}}(\gamma, \lambda)}{\partial \lambda} = \left\| \left(\lambda \boldsymbol{\Phi}_{\boldsymbol{y}} + \boldsymbol{I}_{\mathrm{L}} \right)^{-1} \tilde{\gamma}_{\boldsymbol{x}} \right\|_2^2 - \epsilon = 0, \tag{35}$$

which should be solved for the Lagrange multiplier λ.

Let the eigenvalue decomposition (EVD) of the noisy speech correlation matrix be given by

$$\boldsymbol{\Phi}_{\boldsymbol{y}} = \boldsymbol{U}\boldsymbol{Q}\boldsymbol{U}^H \,, \tag{36}$$

where the columns of $\boldsymbol{U}$ contain the orthogonal eigenvectors and the diagonal elements of the diagonal matrix $\boldsymbol{Q}$ are the corresponding eigenvalues, with $q_0 \geq q_1 \geq \cdots \geq q_{L-1}$. By defining

$$\boldsymbol{z}_{\tilde{\gamma}} = \boldsymbol{U}^H \tilde{\gamma}_{\boldsymbol{x}} \,, \tag{37}$$

and using (36) and (37) in (35), we obtain

$$\boxed{g^{\mathrm{SC}}(\lambda) = \sum_{l=0}^{L-1} \frac{|\boldsymbol{z}_{\tilde{\gamma}}(l)|^2}{(1 + \lambda q_l)^2} = \epsilon} \tag{38}$$

with $\boldsymbol{z}_{\tilde{\gamma}}(l)$ denoting the l-th element of $\boldsymbol{z}_{\tilde{\gamma}}$. This non-linear equation in the Lagrange multiplier λ can be solved, e.g., using Newton's method [39]. The solution is then used in (34), yielding the SC speech correlation vector $\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{SC}}$. Since the normalization constraint in (10) is typically not satisfied, resulting in a scaling inaccuracy, normalization is performed by dividing $\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{SC}}$ with its first element. However, there is no guarantee that the normalized SC speech correlation vector satisfies the quadratic inequality constraint in (26), i.e., lies within the spherical uncertainty set. Using the normalized SC speech correlation vector in (16) results in the *SC-MFMVDR filter*.

B. Doubly-Constrained (DC) MFMVDR Filter

Since it is not guaranteed that the normalized SC speech correlation vector satisfies both the quadratic inequality constraint in (26) as well as the (linear) normalization constraint in (10), in this section we propose to estimate the normalized speech correlation vector as the vector maximizing the total signal output power of the MFMVDR filter while satisfying both constraints, i.e.,

$$\boxed{\begin{aligned} \widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{DC}} = \underset{\gamma}{\operatorname{argmin}} \;\; & \gamma^H \boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}\gamma, \;\; \text{s.t. } \|\gamma - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2 \leq \epsilon, \\ & \boldsymbol{e}^T \gamma = 1 \end{aligned}} \tag{39}$$

This doubly-constrained optimization problem can be transformed into a singly-constrained optimization problem by decomposing the L-dimensional vector γ as

$$\gamma = \begin{bmatrix} 1 \\ -\boldsymbol{d} \end{bmatrix} = \boldsymbol{e} - \boldsymbol{E}\boldsymbol{d} \,, \tag{40}$$

with $\boldsymbol{d}$ an $(L-1)$-dimensional vector and the $L \times (L-1)$-dimensional matrix $\boldsymbol{E}$ defined as

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{0}_{1\times \mathrm{L}\text{-}1} \\ \boldsymbol{I}_{\mathrm{L}\text{-}1} \end{bmatrix} . \tag{41}$$

Similarly, the L-dimensional vectors $\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{DC}}$ and $\tilde{\gamma}_{\boldsymbol{x}}$ can be decomposed as

$$\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{DC}} = \begin{bmatrix} 1 \\ -\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}} \end{bmatrix} = \boldsymbol{e} - \boldsymbol{E}\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}} \tag{42}$$

$$\tilde{\gamma}_{\boldsymbol{x}} = \begin{bmatrix} 1 \\ -\tilde{\boldsymbol{d}}_{\boldsymbol{x}} \end{bmatrix} = \boldsymbol{e} - \boldsymbol{E}\tilde{\boldsymbol{d}}_{\boldsymbol{x}} \,. \tag{43}$$

Instead of estimating $\widehat{\gamma}_{\boldsymbol{x}}^{\mathrm{DC}}$, it is hence sufficient to estimate $\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}}$, which can be done by substituting (40), (42) and (43) into (39),

i.e.,

$$\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\text{DC}} = \underset{\boldsymbol{d}}{\text{argmin}} \, (\boldsymbol{e} - \boldsymbol{E}\boldsymbol{d})^H \, \boldsymbol{\Phi}_{\boldsymbol{y}}^{-1} \, (\boldsymbol{e} - \boldsymbol{E}\boldsymbol{d}) \,, \quad \text{s.t.} \; \left\| \boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|^2 \leq \epsilon, \tag{44}$$

transforming the doubly-constrained optimization problem in (39) into a singly-constrained optimization problem.

Based on the definition of the normalized noisy speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{y}}$ in (12) and using the decomposition

$$\boldsymbol{\gamma}_{\boldsymbol{y}} = \begin{bmatrix} 1 \\ -\boldsymbol{d}_{\boldsymbol{y}} \end{bmatrix}, \tag{45}$$

the $L \times L$-dimensional noisy speech correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}$ can be decomposed as

$$\boldsymbol{\Phi}_{\boldsymbol{y}} = \left[\begin{array}{c|c} \phi_Y & -\phi_Y \boldsymbol{d}_{\boldsymbol{y}}^H \\ \hline -\phi_Y \boldsymbol{d}_{\boldsymbol{y}} & \boldsymbol{D}_{\boldsymbol{y}} \end{array}\right], \tag{46}$$

with $\boldsymbol{D}_{\boldsymbol{y}}$ an $(L-1)\times(L-1)$-dimensional matrix. Using blockwise inversion, the matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1}$ is equal to

$$\boldsymbol{\Phi}_{\boldsymbol{y}}^{-1} = \left[\begin{array}{c|c} a_Y & \boldsymbol{b}_{\boldsymbol{y}}^H \\ \hline \boldsymbol{b}_{\boldsymbol{y}} & \boldsymbol{S}_{\boldsymbol{y}}^{-1} \end{array}\right], \tag{47}$$

with $\boldsymbol{S}_{\boldsymbol{y}}$ the $(L-1)\times(L-1)$-dimensional Schur complement [39], i.e.,

$$\boldsymbol{S}_{\boldsymbol{y}} = \boldsymbol{D}_{\boldsymbol{y}} - \phi_Y \boldsymbol{d}_{\boldsymbol{y}} \boldsymbol{d}_{\boldsymbol{y}}^H, \tag{48}$$

and

$$a_Y = \phi_Y^{-1} + \boldsymbol{d}_{\boldsymbol{y}}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}_{\boldsymbol{y}}, \tag{49}$$

$$\boldsymbol{b}_{\boldsymbol{y}} = \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}_{\boldsymbol{y}}. \tag{50}$$

Using (47), the optimization problem in (44) can hence be reformulated as

$$\boxed{\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\text{DC}} = \underset{\boldsymbol{d}}{\text{argmin}} \; a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H \boldsymbol{d} - \boldsymbol{d}^H \boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}, \quad \text{s.t.} \; \left\| \boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|_2^2 \leq \epsilon} \tag{51}$$

which is similar but obviously not the same as the optimization problem in (28).

For an exemplary noisy speech correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}$ and $L = 3$, Fig. 2 visualizes the quadratic cost function $a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H \boldsymbol{d} - \boldsymbol{d}^H \boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}$, together with an exemplary presumed vector $\tilde{\boldsymbol{d}}_{\boldsymbol{x}}$ (part of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$) and bound ϵ. In comparison to Fig. 1, the quadratic cost function is shifted upwards by the scalar a_Y and shaped by the term $-\boldsymbol{b}_{\boldsymbol{y}}^H \boldsymbol{d} - \boldsymbol{d}^H \boldsymbol{b}_{\boldsymbol{y}}$. Similarly as in Section IV-A, for the optimization problem in (51) the bound ϵ plays an important role and should be chosen in accordance with the accuracy of the presumed vector $\tilde{\boldsymbol{d}}_{\boldsymbol{x}}$.

The minimum of the quadratic cost function $a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H \boldsymbol{d} - \boldsymbol{d}^H \boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}$ is given by $\widehat{\boldsymbol{d}}_{\boldsymbol{x}} = \boldsymbol{S}_{\boldsymbol{y}} \boldsymbol{b}_{\boldsymbol{y}}$, which using (50) is equal to $\boldsymbol{d}_{\boldsymbol{y}}$. Using this solution, or consequently $\boldsymbol{\gamma}_{\boldsymbol{y}}$, in (16) results in the MFMVDR filter being equal to the selection vector $\boldsymbol{e}$, which is obviously undesired. In order to avoid this solution,

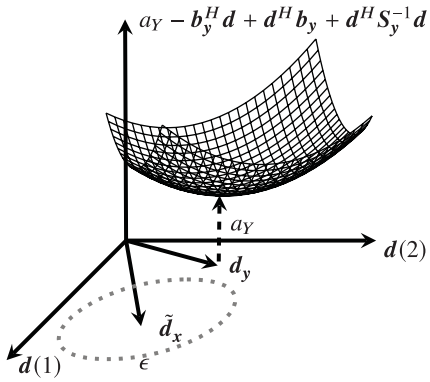


Fig. 2. Quadratic cost function in (51) with exemplary vector $\tilde{\boldsymbol{d}}_{\boldsymbol{x}}$ (part of $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$), vector $\boldsymbol{d}_{\boldsymbol{y}}$ (part of $\boldsymbol{\gamma}_{\boldsymbol{y}}$) and bound ϵ.

the bound ϵ should be chosen such that

$$\epsilon < \left\| \tilde{\boldsymbol{d}}_{\boldsymbol{x}} - \boldsymbol{d}_{\boldsymbol{y}} \right\|_2^2. \tag{52}$$

Under this condition and considering the convex nature of the quadratic cost function in (51), the inequality constraint in (51) can be replaced by an equality constraint, i.e.,

$$\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\text{DC}} = \underset{\boldsymbol{d}}{\text{argmin}} \; a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H \boldsymbol{d} - \boldsymbol{d}^H \boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d}, \quad \text{s.t.} \; \left\| \boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|_2^2 = \epsilon. \tag{53}$$

Similarly to (30), this constrained optimization problem can be solved using the method of Lagrange multipliers [39], where the Lagrangian function is now given by

$$f^{\text{DC}}(\boldsymbol{d}, \mu) = a_Y - \boldsymbol{b}_{\boldsymbol{y}}^H \boldsymbol{d} - \boldsymbol{d}^H \boldsymbol{b}_{\boldsymbol{y}} + \boldsymbol{d}^H \boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d} + \mu \left(\left\| \boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right\|_2^2 - \epsilon \right), \tag{54}$$

with μ the Lagrange multiplier. Setting the gradient of $f^{\text{DC}}(\boldsymbol{d}, \mu)$ with respect to $\boldsymbol{d}$

$$\nabla_{\boldsymbol{d}} f^{\text{DC}}(\boldsymbol{d}, \mu) = 2\boldsymbol{S}_{\boldsymbol{y}}^{-1} \boldsymbol{d} - 2\boldsymbol{b}_{\boldsymbol{y}} + 2\mu(\boldsymbol{d} - \tilde{\boldsymbol{d}}_{\boldsymbol{x}}) \tag{55}$$

equal to zero, yields

$$\boldsymbol{d} = \left(\boldsymbol{S}_{\boldsymbol{y}}^{-1} + \mu \boldsymbol{I}_{\text{L-1}} \right)^{-1} \left(\boldsymbol{b}_{\boldsymbol{y}} + \mu \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right). \tag{56}$$

Applying the matrix inversion lemma, we obtain the vector $\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\text{DC}}(\mu)$ as

$$\boxed{\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\text{DC}}(\mu) = \left(\boldsymbol{I}_{\text{L-1}} - \left(\mu \boldsymbol{S}_{\boldsymbol{y}} + \boldsymbol{I}_{\text{L-1}} \right)^{-1} \right) \left(\frac{1}{\mu} \boldsymbol{b}_{\boldsymbol{y}} + \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right)} \tag{57}$$

Setting the partial derivative of $f^{\text{DC}}(\boldsymbol{d}, \mu)$ in (54) with respect to μ equal to zero and substituting (57) results in

$$g^{\text{DC}}(\mu) = \frac{\partial f^{\text{DC}}(\boldsymbol{d}, \mu)}{\partial \mu} = \left\| \left(\mu \boldsymbol{S}_{\boldsymbol{y}} + \boldsymbol{I}_{\text{L-1}} \right)^{-1} \left(\frac{1}{\mu} \boldsymbol{b}_{\boldsymbol{y}} + \tilde{\boldsymbol{d}}_{\boldsymbol{x}} \right) - \frac{1}{\mu} \boldsymbol{b}_{\boldsymbol{y}} \right\|_2^2 - \epsilon = 0, \tag{58}$$

which should be solved for the Lagrange multiplier μ.

Let the EVD of the Schur complement $\boldsymbol{S_y}$ in (48) be given by

$$\boldsymbol{S_y} = \boldsymbol{V}\boldsymbol{C}\boldsymbol{V}^H, \tag{59}$$

where the columns of $\boldsymbol{V}$ contain the orthogonal eigenvectors and the diagonal elements of the diagonal matrix $\boldsymbol{C}$ are the corresponding eigenvalues, with $c_0 \geq c_1 \geq \cdots \geq c_{L-2}$. By defining

$$\boldsymbol{z}_{\tilde{\boldsymbol{d}}} = \boldsymbol{V}^H \tilde{\boldsymbol{d}}_{\boldsymbol{x}}, \tag{60}$$

$$\boldsymbol{z}_{\boldsymbol{b}} = \boldsymbol{V}^H \boldsymbol{b}_{\boldsymbol{y}}, \tag{61}$$

and using (59), (60) and (61) in (58), we obtain

$$\boxed{g^{\mathrm{DC}}(\mu) = \sum_{l=0}^{L-2} \frac{|\boldsymbol{z}_{\boldsymbol{b}}(l) c_l - \boldsymbol{z}_{\tilde{\boldsymbol{d}}}(l)|^2}{(1+\mu c_l)^2} = \epsilon} \tag{62}$$

with $\boldsymbol{z}_{\tilde{\boldsymbol{d}}}(l)$ and $\boldsymbol{z}_{\boldsymbol{b}}(l)$ denoting the l-th element of $\boldsymbol{z}_{\tilde{\boldsymbol{d}}}$ and $\boldsymbol{z}_{\boldsymbol{b}}$, respectively. This non-linear equation in the Lagrange multiplier μ can be solved similarly to (38), e.g., using Newton's method [39]. The solution is then used in (57), to obtain $\widehat{\boldsymbol{d}}_{\boldsymbol{x}}^{\mathrm{DC}}$, and subsequently in (42), yielding the normalized DC speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}}$. Using $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{DC}}$ in (16) results in the *DC-MFMVDR filter*.

It should be noted that due to the EVD in (36) and (59) and solving the non-linear equations in (38) and (62), the computational complexity for the constrained MFMVDR filters is obviously larger than for the ML-MFMVDR filter, where the computational complexity for the SC-MFMVDR filter and the DC-MFMVDR filter is similar.

C. Bound of the Spherical Uncertainty Set

As already mentioned, the bound ϵ of the spherical uncertainty set in (26) plays a crucial role for both constrained optimization problems in that it should be chosen in accordance with the accuracy of the presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$. In order to ensure that the oracle normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{o}}$ in (22) lies within the spherical uncertainty set (and hence can be found as a solution of the constrained optimization problems), the bound ϵ should be larger than or equal to the oracle bound $\widehat{\epsilon}^{\mathrm{o}} = \|\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{o}} - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}\|_2^2$.

In this paper, we will use the ML estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$ in (24) as the presumed normalized speech correlation vector $\tilde{\boldsymbol{\gamma}}_{\boldsymbol{x}}$. Since the accuracy of the ML estimate strongly depends on the a-priori SNR estimate $\widehat{\xi}$ in (25), we propose to train a (non-linear) mapping function between the oracle bound $\widehat{\epsilon}_{\mathrm{ML}}^{\mathrm{o}} = \|\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{o}} - \widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}\|_2^2$ and the a-priori SNR estimate $\widehat{\xi}$. For a wide range of speech and noise signals (30 TIMIT sentences [40], speech-shaped noise, two traffic and babble noise signals [41]) and a broadband SNR range of 0 dB to 15 dB in 5 dB steps, Fig. 3 shows the normalized joint probability density function (PDF) of the oracle bound $\widehat{\epsilon}_{\mathrm{ML}}^{\mathrm{o}}$ and the a-priori SNR estimate $\widehat{\xi}$ (for all time-frequency points). It can be clearly observed that the oracle bound decreases with increasing a-priori SNR. Fitting a linear function (in log-log scale) to the maximum value of the normalized PDF for each

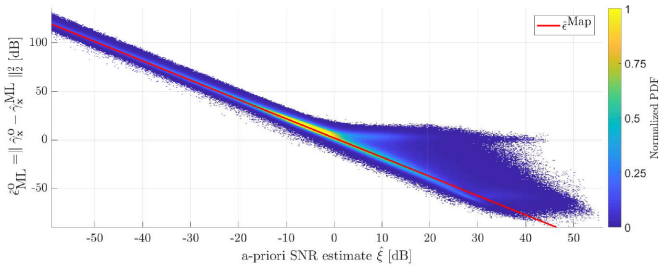


Fig. 3. Normalized joint PDF of the oracle bound $\widehat{\epsilon}_{\mathrm{ML}}^{\mathrm{o}}$ and the a-priori SNR estimate $\widehat{\xi}$ with the mapping function $\widehat{\epsilon}^{\mathrm{Map}}$ in red.

a-priori SNR estimate $\widehat{\xi}$ yields the mapping function

$$\boxed{\widehat{\epsilon}^{\mathrm{Map}}(\widehat{\xi}_{\mathrm{dB}}) = 10^{\left(-1.983\widehat{\xi}_{\mathrm{dB}}+2\right)/10}} \tag{63}$$

with $\widehat{\xi}_{\mathrm{dB}} = 10\log_{10}(\widehat{\xi})$. This mapping function is shown in red in Fig. 3.

V. Evaluation

In this section, we analyze the performance of the proposed constrained MFMVDR filters based on a spherical uncertainty set. After discussing the algorithm implementation framework in Section V-A and defining the instrumental performance measures in Section V-B, in Section V-C we compare the estimation accuracy of the proposed normalized SC and DC speech correlation vector estimates with the ML estimate. For different noise types and SNRs, in Section V-D, V-E and V-F we compare the instrumental and perceptual speech quality of the proposed SC-MFMVDR and DC-MFMVDR filters with the oracle MFMVDR filter, the state-of-the-art ML-MFMVDR filter [12] and the (single-frame) LogSTSA estimator [5] as a reference speech enhancement algorithm.

A. Implementation Framework

In order to exploit speech correlation across time-frames, for the MFMVDR filters we use a highly temporally resolved STFT framework at a sampling frequency of 16 kHz with a frame length of 4 ms, i.e., $K = 64$ frequency-bands, and an overlap of 75%, resulting in a frame shift of 1 ms. As the STFT analysis and synthesis window we use a square-root Hann window. Similarly as in [12], the number of consecutive time-frames is set to $L = 18$, resulting in 21 ms of data used in each filtering operation.

The smoothing parameters for the noisy speech correlation matrix in (18) and the oracle noise correlation matrix in (20) are experimentally set to $\alpha_y = \alpha_n = 0.9$, resulting in a smoothing window of 10 ms. The scaling parameter in (19) is set to $\kappa = 0.001$. The a-priori SNR estimate $\widehat{\xi}$ required for the ML estimate $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}$ in (24) and the bound $\widehat{\epsilon}^{\mathrm{Map}}$ in (63), is computed using the DDA in (25) with a weighting parameter of $\beta_{\mathrm{DDA}} = 0.70$.

Although the main objective is to compare the performance of the proposed constrained MFMVDR filters with the ML-MFMVDR filter, we will also consider the single-frame

LogSTSA estimator as a reference single-channel speech enhancement algorithm. For a fair comparison, the LogSTSA estimator is implemented using an equivalent frame length of 21 ms and an overlap of 50%. The a-priori SNR for the LogSTSA estimator is also estimated using the DDA in (25) with a weighting parameter of $\beta_{\mathrm{DDA}} = 0.98$ and the noise PSD estimator proposed in [37]. To reduce the amount of speech distortion and to mask artifacts in the background noise, a lower limit of -17 dB is applied to the LogSTSA estimator.

For the evaluation, we use 60 sentences from the TIMIT database [40], spoken by different speakers (10 male, 10 female) as speech signals. As noise signals, we use speech-shaped noise, traffic, babble and factory noise signals [41]. The considered broadband SNR range is -5 dB to 20 dB in 5 dB steps. We made sure that the evaluation data differs from the data used for training the mapping function in Section V-C.

B. Instrumental Performance Measures

The accuracy of the normalized speech correlation vector estimates is evaluated in terms of the mean-square error (MSE) between the oracle normalized speech correlation vector in (22) and the estimated normalized speech correlation vector, i.e.,

$$\mathrm{MSE} = \frac{1}{|\mathbb{F}_*|} \sum_{k,m \in \mathbb{F}_*} \frac{\|\widehat{\boldsymbol{\gamma}}^{\mathrm{o}}_{\boldsymbol{x}}(k,m) - \widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}(k,m)\|_2^2}{\|\widehat{\boldsymbol{\gamma}}^{\mathrm{o}}_{\boldsymbol{x}}(k,m)\|_2^2}, \tag{64}$$

where $|\mathbb{F}_*|$ denotes the cardinality of either the set of time-frequency points that contain noise-only ($|\mathbb{F}_N|$) or speech-and-noise ($|\mathbb{F}_Y|$), which are defined as time-frequency points with an oracle a-priori SNR estimate $\widehat{\xi}^{\mathrm{o}}(k,m)$ smaller or larger than -5 dB, respectively. Furthermore, we classify time-frequency points whose normalized squared error is larger than 200 as outliers and exclude them from the MSE calculation in (64).

To evaluate the performance of the MFMVDR filters and the LogSTSA estimator, several instrumental performance measures are used. Speech quality and speech intelligibility are evaluated using the perceptual evaluation of speech quality (PESQ) [3], [42] and the short-time objective intelligibility (STOI) [43] improvements, respectively, 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 (segSSNR) and the segmental noise reduction (segNR) [44], defined as

$$\mathrm{segSSNR} = \frac{10}{|\mathbb{S}|} \sum_{s \in \mathbb{S}} \log_{10} \frac{\sum_{t=1}^{T} x^2(sT+t)}{\sum_{t=1}^{T} \left[x(sT+t) - \tilde{x}(sT+t)\right]^2}, \tag{65}$$

$$\mathrm{segNR} = \frac{10}{|\mathbb{S}|} \sum_{s \in \mathbb{S}} \log_{10} \frac{\sum_{t=1}^{T} n^2(sT+t)}{\sum_{t=1}^{T} \tilde{n}^2(sT+t)}, \tag{66}$$

where T denotes the segment length ($T = 160$, corresponding to 10 ms) and $\mathbb{S}$ is the set of segments that contain speech-and-noise, defined as segments whose energy is larger than -45 dB with respect to the maximum segment energy. The signals $\tilde{x}(t)$ and $\tilde{n}(t)$ denote the processed speech and noise signals. Note that higher SegSSNR values indicate less speech distortion and higher SegNR values indicate more noise reduction. In addition, to evaluate the noise distortion, more in particular the presence of musical noise artifacts in the processed signal, we use the weighted log kurtosis ratio $\Delta\Psi_{\log}$ [45], which was shown to correlate well with perceptual listening results. This measure is defined as the natural logarithm of the ratio of the weighted kurtosis of the processed noise STFT coefficients $\tilde{N}(k,m)$ and the input noise STFT coefficients $N(k,m)$. Note that the perceived amount of noise distortion, especially musical noise, is lowest when $\Delta\Psi_{\log} = 0$ and higher $\Delta\Psi_{\log}$ values, i.e., $\Delta\Psi_{\log} > 0$, indicate more noise distortion.

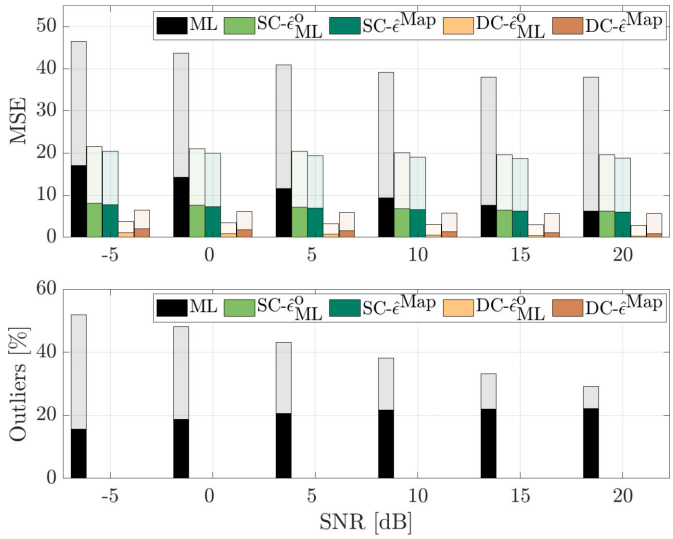
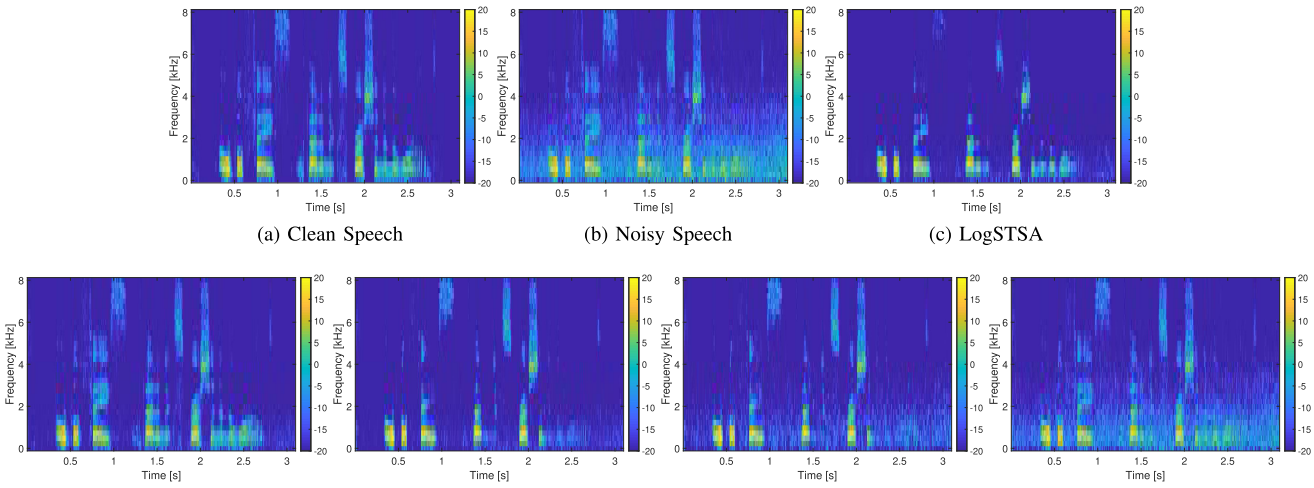


Fig. 4. Average MSE and percentage of outliers for the normalized ML, SC and DC speech correlation vector estimates using the oracle bound $\widehat{\epsilon}^{\mathrm{o}}_{\mathrm{ML}}$ and the mapping function $\widehat{\epsilon}^{\mathrm{Map}}$ for different SNRs. The lower and upper parts of the bars represent the performance in speech-and-noise and noise-only time-frequency points, respectively.

C. Accuracy of the Normalized Speech Correlation Vector Estimates

In this section, we compare the accuracy of the proposed normalized SC and DC speech correlation vector estimates $\widehat{\boldsymbol{\gamma}}^{\mathrm{SC}}_{\boldsymbol{x}}$ and $\widehat{\boldsymbol{\gamma}}^{\mathrm{DC}}_{\boldsymbol{x}}$ with the ML estimate $\widehat{\boldsymbol{\gamma}}^{\mathrm{ML}}_{\boldsymbol{x}}$. To evaluate the proposed mapping function $\widehat{\epsilon}^{\mathrm{Map}}$ for the bound of the spherical uncertainty set in (63), we compare the performance of the SC and DC estimates using the oracle bound $\widehat{\epsilon}^{\mathrm{o}}_{\mathrm{ML}}$ and using the mapping function $\widehat{\epsilon}^{\mathrm{Map}}$.

For different SNRs, Fig. 4 depicts the performance, averaged over all combinations of speech and noise signals, in terms of the MSE in (64) and the percentage of outliers in speech-and-noise time-frequency points (lower bar) and noise-only time-frequency points (upper bar). It can be clearly observed for all SNRs that the SC and DC estimates achieve a considerably lower MSE than the ML estimate, where the DC estimate achieves the lowest MSE of all considered estimates (both in speech-and-noise and noise-only time-frequency points). This shows that the accuracy of the normalized speech correlation vector



(a) Clean Speech (b) Noisy Speech (c) LogSTSA

(d) O-MFMVDR (e) ML-MFMVDR (f) SC-MFMVDR (g) DC-MFMVDR

Fig. 5. Spectrograms of (a) clean speech signal, (b) noisy speech signal, corrupted by traffic noise at 5 dB SNR, and the processed signals using (c) LogSTSA estimator, (d) oracle MFMVDR filter, (e) ML-MFMVDR filter, (f) SC-MFMVDR filter, and (g) DC-MFMVDR filter.

estimate can be substantially improved by jointly considering the quadratic inequality constraint and the normalization constraint (see Section IV-B).

Furthermore, it can be observed for both the SC and DC estimates that the MSE obtained using the proposed mapping function $\widehat{\epsilon}^{\text{Map}}$ is similar to the MSE obtained using the oracle bound $\widehat{\epsilon}^{\text{o}}_{\text{ML}}$, showing that the proposed mapping function is a good approximation. In addition, whereas the ML estimate causes a large amount of outliers, resulting in speech distortion and unpleasant artifacts in the background noise, it can be observed that no outliers occur for the SC and DC estimates.

In conclusion, these results show that the SC and DC estimates are more accurate than the ML estimate, with the DC estimate achieving the highest estimation accuracy.

D. Instrumental Evaluation

In this section, the performance of the proposed SC-MFMVDR and DC-MFMVDR filters using the mapping function $\widehat{\epsilon}^{\text{Map}}$ is evaluated and compared with the oracle MFMVDR (O-MFMVDR) filter using the oracle normalized speech correlation vector $\widehat{\boldsymbol{\gamma}}^{\text{o}}_{\boldsymbol{x}}(m)$ in (22) and the state-of-the-art ML-MFMVDR filter [12]. As already mentioned, although the main objective is to compare the proposed constrained MFMVDR filters with the ML-MFMVDR filter, we also consider the single-frame LogSTSA estimator as reference algorithm.

For a speech signal from the TIMIT database corrupted by traffic noise at 5 dB SNR, Fig. 5 depicts the spectrograms of the clean speech and noisy speech signals and the processed signals using the LogSTSA estimator and the MFMVDR filters. First, it can be observed that the spectrogram of the oracle MFMVDR filter in Fig. 5(d) is very similar to the spectrogram of the clean speech signal in Fig. 5(a). Second, it can be observed that the LogSTSA estimator in Fig. 5(c) clearly reduces the background noise, but also suppresses the speech signal. Third, it can be observed that all blind MFMVDR filters in Fig. 5(e)-(g) reduce less noise than the LogSTSA estimator, but clearly preserve the speech signal better (especially at high frequencies). Among the blind MFMVDR filters there is a trade-off between speech distortion and noise reduction, which will be investigated in more detail in Section V-E.

For different SNRs, Fig. 6 depicts the results, averaged over all combinations of speech and noise signals, in terms of the considered instrumental performance measures, i.e., segSSNR (speech distortion), segNR (noise reduction), $\Delta\Psi_{\log}$ (noise distortion), ΔPESQ (speech quality) and ΔSTOI (speech intelligibility). First, it can be observed that the oracle MFMVDR filter clearly outperforms all other filters in terms of all instrumental performance measures. Second, it can be observed that the constrained MFMVDR filters yield larger segSSNR values (i.e., less speech distortion) but smaller segNR values (i.e., less noise reduction) than the ML-MFMVDR filter and the LogSTSA estimator. Among the blind MFMVDR filters, the DC-MFMVDR filter yields the largest segSSNR values, which are close to the oracle MFMVDR results. This can be explained by the high estimation accuracy of the normalized DC speech correlation vector (see Fig. 4). The more conservative noise reduction performance of the SC-MFMVDR and DC-MFMVDR filters compared to the ML-MFMVDR filter (especially at low SNRs) can be explained by the additional robustness constraints. Please note that although the estimation accuracy of the SC and DC normalized speech correlation vector estimates is quite good, it is not good enough for the SC-MFMVDR and DC-MFMVDR filters to reach the performance of the oracle MFMVDR filter. Third, it can be observed that the constrained MFMVDR filters yield lower $\Delta\Psi_{\log}$ values (i.e., less noise distortion) than the ML-MFMVDR filter and the LogSTSA estimator, and that among the blind MFMVDR filters, the DC-MFMVDR filter yields the lowest $\Delta\Psi_{\log}$ values. Fourth, the ΔPESQ results indicate that at low SNRs a better overall quality is obtained by the constrained MFMVDR filters than the ML-MFMVDR filter, whereas at high SNRs a better overall quality is obtained by the ML-MFMVDR

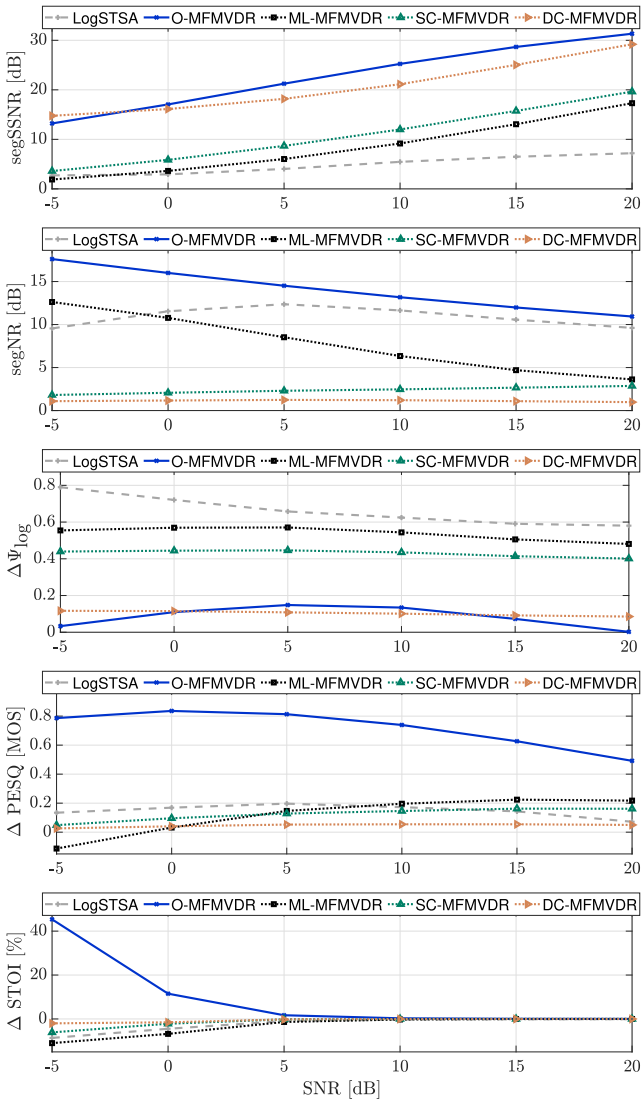


Fig. 6. Average segmental speech SNR (segSSNR), segmental noise reduction (segNR), weighted log kurtosis ratio ($\Delta\Psi_{\log}$), PESQ improvement (ΔPESQ) and STOI improvement (ΔSTOI) obtained using the LogSTSA estimator, the oracle MFMVDR (O-MFMVDR) filter, the ML-MFMVDR filter and the proposed SC-MFMVDR, and DC-MFMVDR filters for different SNRs.

filter than the constrained MFMVDR filters. However, since in [12], [38] is was reported that the ML-MFMVDR filter introduces unpleasant artifacts, e.g., musical noise in the background noise, and informal listening experiments suggest that for the constrained MFMVDR filters the speech sounds more natural and less musical noise is present than for the ML-MFMVDR filter for all SNRs, we decided to conduct a formal listening test in Section V-F. Finally, the ΔSTOI results indicate no speech intelligibility improvement at low SNRs for all blind filters, where the ML-MFMVDR filter and the LogSTSA estimator yield lower ΔSTOI values than the constrained MFMVDR filters.

E. Trade-off Between Speech Distortion and Noise Reduction

As already mentioned, for the blind MFMVDR filters a trade-off exists between speech distortion (segSSNR values) and noise reduction (segNR values). To further investigate this trade-off, in this section we consider the trade-off MFMVDR (T-MFMVDR) filter, which uses a normalized speech correlation vector defined as

$$\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{T}}(m) = \beta_{\mathrm{T}}\,\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}(m) + (1-\beta_{\mathrm{T}})\,\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m)\,, \tag{67}$$

with $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m)$ the normalized noisy speech correlation vector and β_{T} a trade-off parameter. When β_{T} is equal to zero, $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{T}}(m) = \widehat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m)$ and the T-MFMVDR filter is equal to the selection vector $\boldsymbol{e}$, i.e., no speech distortion is introduced but also no noise reduction is obtained. When β_{T} is equal to one, $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{T}}(m) = \widehat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\mathrm{ML}}(m)$ and the T-MFMVDR filter is equal to the ML-MFMVDR filter, i.e., leading to the highest noise reduction but also highest speech distortion of the considered blind MFMVDR filters.

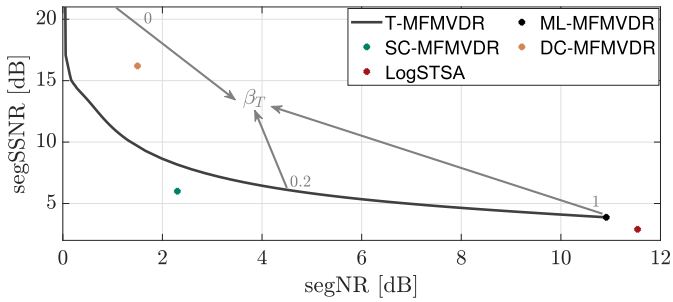


Fig. 7. Average segSSNR vs. average segNR for the T-MFMVDR filter for different values of β_{T}, the ML-MFMVDR filter, the SC-MFMVDR filter, the DC-MFMVDR filter, and the LogSTSA estimator (SNR = 0 dB).

For an SNR of 0 dB, Fig. 7 depicts the segSSNR results vs. the segNR results, averaged over all combinations of speech and noise signals, for the T-MFMVDR filter for different values of the trade-off parameter β_{T}, the ML-MFMVDR filter ($\beta_{\mathrm{T}} = 1$), the SC-MFMVDR and DC-MFMVDR filters and the LogSTSA estimator. For the T-MFMVDR filter, it can be observed that with increasing β_{T} the segSSNR value decreases while the segNR value increases. Note that for $\beta_{\mathrm{T}} = 0$, the segSSNR value is equal to infinity, while the segNR value is equal to zero. It can be observed that compared to the T-MFMVDR filter the proposed DC-MFMVDR filter achieves a segSSNR value that is 6.5 dB higher at the same segNR value and a segNR value that is 1.5 dB higher at the same segSSNR value. In contrast compared to the T-MFMVDR filter, the SC-MFMVDR filter leads to a segSSNR value that is 2 dB lower at the same segNR values and a segNR value that 3 dB lower at the same segSSNR value. The LogSTSA estimator results in a higher segNR value than the T-MFMVDR filter for $\beta_{\mathrm{T}} = 1$ (i.e., the ML-MFMVDR filter) but a lower segSSNR value.

These results show that the DC-MFMVDR filter achieves a better trade-off between noise reduction and speech distortion than a comparable trade-off MFMVDR filter.

F. Perceptual Evaluation

In this section, we perceptually compare the speech enhancement performance of the SC-MFMVDR filter, the DC-MFMVDR filter, the ML-MFMVDR filter, the oracle MFMVDR filter and the LogSTSA estimator using a subjective

listening test. For two speech signals and two acoustic scenarios (noise types), we conducted a procedure similar to the multi stimulus test with hidden reference and anchor (MUSHRA) [46], evaluating three attributes: (a) overall quality, (b) speech distortion and (c) noise reduction. For attribute (a), the participants were asked to rate the overall signal quality of the test signals with respect to a reference signal. For attribute (b), the participants were asked to rate how distorted the speech component of the test signals sounds with respect to a reference signal. For attribute (c), the participants were asked to rate how noticeable the amount of noise reduction of the test signals is with respect to a reference signal. As speech signals, we used two sentences from the TIMIT database [40], spoken by a male and a female speaker. To generate both acoustic scenarios, we mixed the speech signals with traffic noise and babble noise taken from the NOISEX-92 database [41] at 5 dB SNR. For each attribute, acoustic scenario and speech signal, in addition to the processed signals a noisy speech signal, a hidden reference and an anchor were presented to the participants. For the attributes (a) and (b), the (hidden) reference was the noisy speech signal at 20 dB SNR. The anchor was the speech signal low-pass filtered at 3 kHz, corrupted with the noise at -5 dB and processed with an aggressive Wiener gain using a weighting parameter of $\beta_{\mathrm{DDA}} = 0.97$ and a lower limit of -20 dB. For the attribute (c), the (hidden) reference was the unprocessed noisy speech signal at 5 dB SNR, while the anchor was a noisy speech signal at 20 dB SNR. For the sake of completeness, for attribute (c) we also presented the anchor for the attributes (a) and (b) as a test signal. Hence, for each attribute, acoustic scenario and speech signal, the participants compared eight test signals with a reference signal, i.e., either a noisy speech signal at 20 dB SNR for the attributes (a) and (b) or the unprocessed noisy speech signal at 5 dB SNR for the attribute (c). Examples of audio samples for all test signals are available online (see https://uol.de/en/sigproc/research/audio-demos/multi-frame-speech-enhancement/constrained-mfmvdr-filters).

A total of 11 self-reported normal-hearing participants in the range of 22 to 39 years participated in the subjective listening test. Due to the COVID-19 pandemic, the experiment took place in quiet rooms at the participants' home. The signals were presented diotically to the participants, using their own sound cards and over-the-ear headphones.

The listening test consisted of two phases. First, the participants were trained to familiarize themselves with the presented signals and to adjust the volume to a comfortable level. Second, the participants were instructed to rate the test signals according to the three aforementioned attributes on a continuous scale from 0 to 100 using sliders in a graphical user interface. For the attribute (a) overall quality, 0 was labeled with "bad" and 100 with "excellent," while for the attribute (b) speech distortion, 0 was labeled with "extremely distorted" and 100 with "not distorted," and for the attribute (c) noise reduction, 0 was labeled with "extremely noticeable" and 100 with "not noticeable". The participants were allowed to listen to the reference signal and all test signals as often as they wanted. The participants were instructed to rate at least one test signal with a score of 100, which should correspond to the hidden reference. The order of the presentation of the test signals and acoustic scenarios were randomized between all participants.

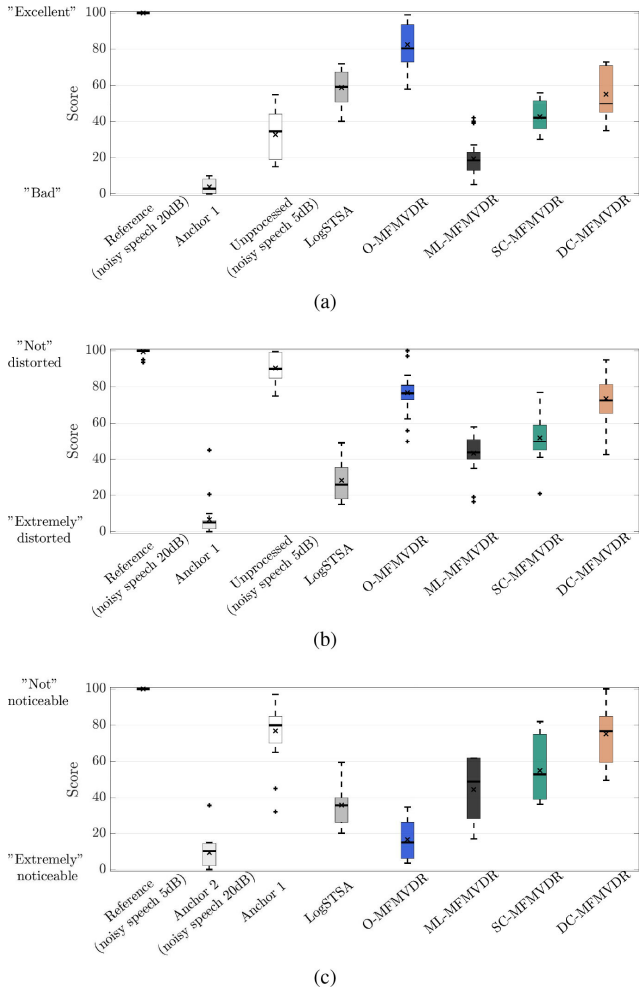


Fig. 8. Averaged MUSHRA scores for the attributes (a) overall quality, (b) speech distortion and (c) noise reduction, for a hidden reference, an anchor, a noisy speech signal and the processed signals using the LogSTSA estimator, the oracle MFMVDR (O-MFMVDR) filter, the ML-MFMVDR filter and the proposed SC-MFMVDR and DC-MFMVDR filters for SNR $= 5$ dB. On each box, the central horizontal line is the median, the edges of the box are the 25-th and 75-th percentiles and the whiskers extend to 1.5 times the interquartile range from the median. The means are indicated by $\times$ markers. Outliers are indicated by $+$ markers.

For each attribute, a statistical analysis was conducted using the resulting MUSHRA scores of both speech signals and both acoustic scenarios. Since the data are normally distributed, as shown by the Shapiro-Wilk test, a repeated-measures analysis of variance (ANOVA) [47] was performed with factors "acoustic scenario" and "algorithm". Since the statistical analysis showed no significant influence of the factor "acoustic scenario" for all attributes, we averaged the MUSHRA scores over both acoustic scenarios. Since the statistical analysis showed a significant influence of the factor "algorithm" for all attributes, we tested for statistically significant differences between the algorithm mean values by conducting a post-hoc pairwise comparison t-test with Bonferroni correction. Fig. 8 depicts the averaged MUSHRA scores for all three attributes using boxplots. The t-test results

TABLE I
OVERVIEW OF THE T-TEST RESULTS FOR THE ATTRIBUTE OVERALL QUALITY. THE ASTERISKS DENOTE RESULTS THAT ARE STATISTICALLY SIGNIFICANT (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) AND O DENOTES RESULTS THAT ARE NOT STATISTICALLY SIGNIFICANT ($p > 0.05$)

	Reference	Anchor 1	Unprocessed	LogSTSA	O-MFMVDR	ML-MFMVDR	SC-MFMVDR	DC-MFMVDR
Reference		***	***	***	o	***	***	***
Anchor 1	***		**	***	***	o	**	***
Unprocessed	***	**		*	***	o	o	*
LogSTSA	***	***	*		*	***	o	o
O-MFMVDR	o	***	***	*		***	***	*
ML-MFMVDR	***	o	o	***	***		*	**
SC-MFMVDR	***	**	o	o	***	*		o
DC-MFMVDR	***	***	*	o	*	**	o	

TABLE II
OVERVIEW OF THE T-TEST RESULTS FOR THE ATTRIBUTE SPEECH DISTORTION. THE ASTERISKS DENOTE RESULTS THAT ARE STATISTICALLY SIGNIFICANT (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) AND O DENOTES RESULTS THAT ARE NOT STATISTICALLY SIGNIFICANT ($p > 0.05$)

	Reference	Anchor 1	Unprocessed	LogSTSA	O-MFMVDR	ML-MFMVDR	SC-MFMVDR	DC-MFMVDR
Reference		***	o	***	o	***	***	*
Anchor 1	***		***	o	***	**	***	***
Unprocessed	o	***		***	o	***	**	o
LogSTSA	***	o	***		***	o	*	***
O-MFMVDR	o	***	o	***		**	*	o
ML-MFMVDR	***	**	***	o	**		o	**
SC-MFMVDR	***	***	**	*	*	o		o
DC-MFMVDR	*	***	o	***	o	**	o	

TABLE III
OVERVIEW OF THE T-TEST RESULTS FOR THE ATTRIBUTE NOISE REDUCTION. THE ASTERISKS DENOTE RESULTS THAT ARE STATISTICALLY SIGNIFICANT (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$) AND O DENOTES RESULTS THAT ARE NOT STATISTICALLY SIGNIFICANT ($p > 0.05$)

	Unprocessed	Anchor 2	Anchor 1	LogSTSA	O-MFMVDR	ML-MFMVDR	SC-MFMVDR	DC-MFMVDR
Unprocessed		***	*	***	***	***	***	**
Anchor 2	***		***	**	o	***	***	***
Anchor 1	*	***		***	***	**	*	o
LogSTSA	***	**	***		*	o	*	***
O-MFMVDR	***	o	***	*		**	***	***
ML-MFMVDR	***	***	**	o	**		o	**
SC-MFMVDR	***	***	*	*	***	o		*
DC-MFMVDR	**	***	o	***	***	**	*	

are presented in Tables I– III, with asterisks denoting statistically significant differences and o denoting not statistically significant differences.

In terms of the attribute (a) overall quality (cf. Fig. 8(a) and Table I), the mean score for the hidden reference was equal to 100 and the anchor was rated with the lowest mean score of 3.9, as desired. The mean score for the unprocessed noisy speech signal at 5 dB SNR was equal to 32.7, which is significantly lower than for all processed signals, except for the ML-MFMVDR filter and the SC-MFMVDR filter with mean scores of 19.3 and 42.6, respectively. For the LogSTSA estimator, the oracle MFMVDR filter and the DC-MFMVDR filter, the mean score was equal to 58.8, 82.6 and 55.2, respectively. The statistical analysis showed that the oracle MFMVDR filter was rated significantly higher than all blind algorithms. While the differences between the LogSTSA estimator, the SC-MFMVDR filter and the DC-MFMVDR filter are not statistically significant, these mean scores are significantly higher than the mean score of the ML-MFMVDR filter, while only the LogSTSA estimator and the DC-MFMVDR filter were rated significantly higher than the unprocessed noisy speech signal.

In terms of the attribute (b) speech distortion (cf. Fig. 8(b) and Table II), the mean score of the hidden reference was equal to 100 and the anchor was rated with the lowest mean score of 6.7, as desired. The unprocessed noisy speech signal at 5 dB SNR was rated with the highest mean score of 90.4, followed by the oracle MFMVDR filter with a mean score of 76.5 and the DC-MFMVDR filter with a mean score of 73.6. These differences are not statistically significant. The DC-MFMVDR filter was rated significantly higher than the ML-MFMVDR filter with a mean-score of 43.2, but there was no statistically significant difference between the DC-MFMVDR filter and the SC-MFMVDR filter with a mean score of 51.9. Except for the ML-MFMVDR filter, the MFMVDR filters were rated significantly higher than the LogSTSA estimator, with a mean score of 28.2.

In terms of the attribute (c) noise reduction (cf. Fig. 8(c) and Table III), the mean score of the hidden reference (in this case the unprocessed noisy speech signal at 5 dB) was equal to 100 and the anchor (in this case the noisy speech signal at 20 dB SNR) was rated with the lowest mean score of 9.3, as desired. The oracle MFMVDR filter was rated with a mean score of 16.6. For the LogSTSA estimator, the ML-MFMVDR filter, the SC-MFMVDR filter and the DC-MFMVDR filter, the mean score was equal to 35.6, 44.8, 55.5 and 75.1, respectively. All blind algorithms were rated significantly worse than the oracle MFMVDR filter. Although the difference between the LogSTSA estimator and the ML-MFMVDR filter is not statistically significant, the LogSTSA estimator was rated significantly better than the SC-MFMVDR and DC-MFMVDR filters. Although the difference between the SC-MFMVDR filter and the ML-MFMVDR filter is not statistically significant, the ML-MFMVDR filter was rated significantly better than the DC-MFMVDR filter.

In conclusion, these results show that the perceived overall quality for the SC-MFMVDR filter and the DC-MFMVDR filter is significantly better than for the ML-MFMVDR filter and

shows no statistically significant difference to the LogSTSA estimator. This can presumably be explained by the fact that all considered algorithms produce different artifacts and distortions in the speech and noise signals, which may be perceived and rated differently by the listeners. Although the perceived amount of noise reduction for the DC-MFMVDR filter is clearly lower than for the ML-MFMVDR filter, the SC-MFMVDR filter and the LogSTSA estimator, this is apparently compensated by the extremely low perceived speech distortion. Hence, the DC-MFMVDR is most suitable for applications where low speech distortion is considered to be more important than much noise reduction.

VI. Conclusion

In this paper we investigated the potential of using concepts proposed for robust MVDR beamforming in the context of single-channel multi-frame speech enhancement. We proposed two constrained MFMVDR filters that estimate the normalized speech correlation vector as the vector maximizing the total signal output power within a spherical uncertainty set. This corresponds to imposing a quadratic inequality constraint on the mismatch vector with respect to the presumed normalized speech correlation vector. While the SC-MFMVDR filter only considers the quadratic inequality constraint and applies the required normalization afterwards, the DC-MFMVDR jointly considers the quadratic inequality constraint and the linear normalization constraint in the optimization problem. To set the upper bound of the spherical uncertainty set, we proposed to use a trained non-linear mapping function that depends on the a-priori SNR.

Simulation results show that the proposed approaches to estimate the normalized speech correlation vector clearly lead to a more accurate estimate than the ML estimate, with the DC estimate achieving the highest estimation accuracy. An instrumental evaluation for different noise types and SNRs indicates that although the proposed constrained MFMVDR filters lead to a more conservative noise reduction than the ML-MFMVDR filter and the logarithmic short-time spectral amplitude estimator, especially the DC-MFMVDR filter produces less speech and noise distortions than the ML-MFMVDR filter. Moreover, the DC-MFMVDR filter achieves a better trade-off between noise reduction and speech distortion than a comparable trade-off MFMVDR filter. The results of a perceptual listening test show that the perceived overall quality for the proposed constrained MFMVDR filters is significantly better than for the ML-MFMVDR filter.

References

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

[2] 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.

[3] P. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. CRC press, 2013.

[4] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

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

[6] M. Kolbk, Z. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[7] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.

[8] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2019.

[9] J. Lee and H. Kang, "A joint learning algorithm for complex-valued T-F masks in deep learning-based single-channel speech Enhancement Systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1098–1108, Jun. 2019.

[10] T. Esch and P. Vary, "Modified kalman filter exploiting interframe correlation of speech and noise magnitudes," in *Proc. Int. Workshop Acoustic Echo Noise Control*, Seattle, WA, USA, Sep. 2008, pp. 1–4.

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

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

[13] M. Parchami, W. P. Zhu, and B. Champagne, "Speech dereverberation using weighted prediction error with correlated inter-frame speech components," *Speech Commun.*, vol. 87, pp. 49–57, Mar. 2017.

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

[15] R. Ranjbaryan, H. R. Abutalebi, and S. Doclo, "Reduced-complexity semi-distributed multi-channel multi-frame MVDR filter," in *Proc. Europ. Signal Process. Conf.*, Rome, Italy, 2018, pp. 2095–2099.

[16] D. Fischer and S. Doclo, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *Proc. Int. Workshop Acoust. Signal Enhancement*, Tokyo, Japan, Sep. 2018, pp. 41–45.

[17] J. Stahl and P. Mowlaee, "Exploiting temporal correlation in pitch-adaptive speech enhancement," *Speech Commun.*, vol. 111, pp. 1–13, Aug. 2019.

[18] E. Plourde, "Multidimensional STSA estimators for speech enhancement with correlated spectral components," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3013–3024, Jul. 2011.

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

[20] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 305–308.

[21] H. Momeni, H. R. Abutalebi, and E. A. P. Habets, "Conditional MMSE-based single-channel speech enhancement using inter-frame and inter-band correlations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 5215–5219.

[22] 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, Apr. 1988.

[23] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[24] 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.*, Kos, Greece, Aug. 2017, pp. 633–637.

[25] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.

[26] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, Feb. 2003.

[27] P. Stoica, Z. Wang, and J. Li, "Robust Capon beamforming," *IEEE Signal Process. Lett.*, vol. 10, no. 6, Jun. 2003.

[28] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, Jul. 2003.

[29] S. Vorobyov, A. Gershman, and Z.-Q. Luo, "Adaptive beamforming with joint robustness against mismatched signal steering vector and interference nonstationarity," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 108–111, Feb. 2004.
[30] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust Capon beamformer," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2407–2423, Sep. 2004.
[31] S. Vorobyov, H. Chen, and A. Gershman, "On the relationship between robust minimum variance beamformers with probabilistic and worst-case distortionless response constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5719–5724, Nov. 2008.
[32] A. Khabbazibasmenj, S. A. Vorobyov, and A. Hassanien, "Robust adaptive beamforming based on steering vector estimation with as little as possible prior information," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2974–2987, Feb. 2012.
[33] S. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *IEEE Trans. Signal Process.*, vol. 93, no. 12, pp. 3264–3277, Dec. 2013.
[34] Y. Zhao, J. R. Jensen, M. G. Christensen, S. Doclo, and J. Chen, "Experimental study of robust beamforming techniques for acoustic applications," in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, New Paltz, NY, USA, 2017, pp. 86–90.
[35] 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.
[36] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, USA, Mar. 2008, pp. 4897–4900.
[37] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[38] 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 Conf. Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 292–296.
[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.
[40] J. S. Garofolo *et al.*, "DARPA TIMIT acoustic phonetic continuous speech database," in *Nat. Inst. Standards Technol.*, 1993.
[41] A. Varga and H. J. 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 Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
[42] "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.
[43] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[44] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Jan. 2005.
[45] H. Yu and T. Fingscheidt, "A weighted log kurtosis ratio measure for instrumental musical tones assessment in wideband speech," in *Proc. ITG Conf. Speech Commun.*, Braunschweig, Germany, Sep. 2012, pp. 1–4.
[46] "Recommendation ITU-R BS.1534-3.: Method for the subjective assessment of intermediate quality level of audio systems," Oct. 2015.
[47] A. Field, *Discovering statistics using IBM SPSS statistics*, 3rd ed. SAGE, 2009.

Dörte Fischer (Member, IEEE) received the B.Eng. degree from the Jade University of Applied Sciences Oldenburg, Wilhelmshaven, Germany, in 2013 and the M.Sc. degree from the University of Oldenburg, Oldenburg, Germany, in 2014 both in hearing technology and audiology. From 2014 to 2020, she was a Doctoral Researcher with the Signal Processing Division, Department of Medical Physics and Acoustics, University of Oldenburg. Her research interests include digital speech and audio processing, including speech enhancement and hearing devices. Since 2018, she is the Head of Audiology with the ENT Clinic, Güstrow, Germany.

Simon Doclo (Senior Member, IEEE) received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003, respectively. From 2003 to 2007, he was a Postdoctoral Fellow with the Research Foundation Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Cognitive Systems Laboratory (McMaster University, Canada). From 2007 to 2009, he was a Principal Scientist with NXP Semiconductors, Leuven, Belgium. Since 2009, he is a Full Professor with the University of Oldenburg, Germany, and scientific advisor for the Division Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks and hearing aid processing. He was the recipient of the several best paper awards (International Workshop on Acoustic Echo and Noise Control 2001, *EURASIP Signal Processing* 2003, IEEE Signal Processing Society 2008, VDE Information Technology Society 2019). He is member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing and the EAA Technical Committee on Audio Signal Processing. He was Technical Program Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in 2013 and Chair of the ITG Conference on Speech Communication in 2018. In addition, he served as a Guest Editor for several special issues (*IEEE Signal Processing Magazine, Elsevier Signal Processing*) and was an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and *EURASIP Journal on Advances in Signal Processing*.