ROBUST CONSTRAINED MFMVDR FILTERING FOR SINGLE-MICROPHONE SPEECH ENHANCEMENT

Dörte Fischer and Simon Doclo

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg, 26129 Oldenburg, Germany.

ABSTRACT

The multi-frame minimum variance distortionless response (MFMVDR) filter for single-microphone speech enhancement exploits speech correlation across consecutive time frames. This filter is designed to avoid speech distortion while minimizing the total signal output power. The MFMVDR filter is very sensitive to estimation errors in the speech correlation vector, since correlated speech components may be mistakenly suppressed. Inspired by robust beamforming approaches, in this paper we propose a robust constrained MFMVDR filter for single-microphone speech enhancement by estimating the speech correlation vector that maximizes the total signal output power within a spherical uncertainty set. For the upper bound of the spherical uncertainty set, we propose to use a trained mapping function that depends on the a-priori SNR. Experimental results for different noise types and SNRs show that the proposed robust approach achieves a more accurate estimate of the speech correlation vector resulting in low speech and noise distortion but a more conservative noise reduction.

Index Terms— Speech Enhancement, Robust MVDR, Noise Reduction, Multi-Frame, Speech Distortion

1. INTRODUCTION

Using speech communication devices (e.g., hearing aids) in noisy environments may lead to a degraded quality and intelligibility of the desired speech signal. In these acoustic situations, speech enhancement algorithms are required to suppress the undesired interference while limiting speech distortion. Typically, single-microphone speech enhancement algorithms are designed in the short-time Fourier transform (STFT) domain. When assuming that neighboring STFT coefficients are uncorrelated over time and frequency, the noisy STFT coefficients are processed by applying a (real-valued) gain to each time-frequency point [1]. When using the more realistic assumption that neighboring STFT coefficients are correlated over time, it has been proposed to process the noisy STFT coefficients by applying a (complex-valued) finite-impulse response (FIR) filter [2, 3, 4, 5].

In [3, 4], a multi-frame signal model has been proposed, where a *speech correlation vector* contains the speech correlation between the current and previous time-frames. Conceptually, this multi-frame signal model is similar to a multi-microphone signal model when interpreting time-frames as microphone inputs and the speech correlation vector as the steering vector. Similarly to the well-known minimum variance distortionless response (MVDR) beamformer [6, 7] in this way the *multi-frame MVDR* (MFMVDR) filter was derived in [3, 4], which minimizes the total signal output power while not distorting correlated speech components.

In practice, obviously only the noisy speech signal is available such that the speech correlation vector needs to be estimated from the noisy STFT coefficients. In [8] a purely data-driven maximum-likelihood (ML) estimator for the speech correlation vector was proposed. It was shown that when using this speech correlation vector, the MFMVDR filter introduces low speech distortion and achieves a good noise reduction. In [9], we showed that accurately estimating the speech correlation vector is crucial in that even a small mismatch between the optimal and the estimated speech correlation vector may lead to a degraded performance of the MFMVDR filter.

In the area of array processing several techniques have been proposed to increase the robustness of beamformers against estimation errors of the steering vector or with as little as possible information about the steering vector [10, 11, 12, 13]. Inspired by the robust MVDR beamformer in [10], which estimates the steering vector that maximizes the total signal output power of the MVDR within a spherical uncertainty set, in this paper we propose a *robust constrained (RC) MFMVDR* filter by estimating the speech correlation vector in the MFMVDR filter by adding a quadratic inequality constraint to the original problem formulation. The quadratic inequality constraint imposes an upper bound on the norm of the mismatch vector, i.e., the difference between the speech correlation vector and the presumed speech correlation vector, e.g., calculated using the ML method in [8]. Since oracle simulations using several speech and noise signals at different signal-to-noise ratios (SNR)s showed that the norm of the mismatch vector decreases with increasing SNR, we trained a linear mapping function to set the upper bound depending on the a-priori SNR for each time-frequency point. Evaluation results for different noise types and SNRs show that the proposed RC speech correlation vector achieves a lower mean-squared error than the ML estimate, resulting in considerably less speech and noise distortions with slightly less noise reduction.

2. MULTI-FRAME SIGNAL MODEL

We consider a single-microphone setup, where a speech signal is degraded by additive noise. In the STFT domain, the complex-valued noisy speech coefficient $Y(k,m)$ at frequency-bin k and time-frame m is given by the complex-valued speech $X(k,m)$ and noise $N(k,m)$ coefficients, i.e.,

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

For conciseness, in the remainder of the paper the frequency-bin k will be omitted if not required.

In multi-frame single-microphone speech enhancement approaches [2] the speech coefficient $X(m)$ is estimated by applying a complex-valued FIR filter $\boldsymbol{h}(m)$ to the noisy speech vector $\boldsymbol{y}(m)$, i.e.,

$$\hat{X}(m)=\boldsymbol{h}^{H}(m)\boldsymbol{y}(m), \tag{2}$$

where ^{H} denotes the Hermitian operator. The noisy speech vector $\boldsymbol{y}(m)$ contains L consecutive noisy speech coefficients and the filter $\boldsymbol{h}(m)$

This work was supported in part by the joint Lower Saxony-Israeli Project ATHENA and by the Cluster of Excellence EXC 1077 Hearing4all, funded by the German Research Foundation (DFG).

978-1-5386-8151-0/18/$31.00

contains L time-varying filter coefficients, i.e.,

$$\boldsymbol{y}(m) = \big[Y(m), Y(m-1), ..., Y(m-L+1)\big]^T, \quad (3)$$

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

The noisy speech vector $\boldsymbol{y}(m)$ is given by

$$\boldsymbol{y}(m) = \boldsymbol{x}(m) + \boldsymbol{n}(m), \quad (5)$$

where the speech vector $\boldsymbol{x}(m)$ and the noise vector $\boldsymbol{n}(m)$ are defined similarly as in (3). Assuming that the speech and noise signals are uncorrelated, the $L \times L$-dimensional noisy speech correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{yy}}(m) = \mathbb{E}\big[\boldsymbol{y}(m)\boldsymbol{y}^H(m)\big]$, with $\mathbb{E}[\cdot]$ the expectation operator, is given by

$$\boldsymbol{\Phi}_{\boldsymbol{yy}}(m) = \boldsymbol{\Phi}_{\boldsymbol{xx}}(m) + \boldsymbol{\Phi}_{\boldsymbol{nn}}(m), \quad (6)$$

with $\boldsymbol{\Phi}_{\boldsymbol{xx}}(m)$ the speech correlation matrix and $\boldsymbol{\Phi}_{\boldsymbol{nn}}(m)$ the noise correlation matrix.

To exploit the speech correlation across time-frames, it has been proposed in [3, 4] to decompose the speech vector $\boldsymbol{x}(m)$ into a temporally correlated speech component $\boldsymbol{s}(m)$ and a temporally uncorrelated speech component $\boldsymbol{x}'(m)$ with respect to the speech coefficient $X(m)$, i.e.,

$$\boldsymbol{x}(m) = \boldsymbol{s}(m) + \boldsymbol{x}'(m), \quad (7)$$

with

$$\boldsymbol{s}(m) = \boldsymbol{\gamma}_{\boldsymbol{x}}(m) X(m). \quad (8)$$

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

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

where * denotes the complex-conjugate operator and $\boldsymbol{e} = \big[1, 0, ..., 0\big]^T$ is the L-dimensional selection vector. Based on the normalization term $\boldsymbol{e}^T\boldsymbol{\Phi}_{\boldsymbol{xx}}(m)\boldsymbol{e}$, which corresponds to the speech power spectral density (PSD) $\phi_X(m)$, the first element of the speech correlation vector is equal to 1, i.e.,

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

Substituting (7) and (8) into (5) and considering the uncorrelated speech component $\boldsymbol{x}'(m)$ as an interference, the *multi-frame signal model* is given by

$$\boxed{\boldsymbol{y}(m) = \boldsymbol{\gamma}_{\boldsymbol{x}}(m)X(m) + \boldsymbol{u}(m)} \quad (11)$$

with $\boldsymbol{u}(m) = \boldsymbol{x}'(m) + \boldsymbol{n}(m)$ the undesired signal vector.

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

$$\boldsymbol{\gamma}_{\boldsymbol{y}}(m) = \frac{\boldsymbol{\Phi}_{\boldsymbol{yy}}(m)\,\boldsymbol{e}}{\phi_Y(m)}, \qquad \boldsymbol{\gamma}_{\boldsymbol{n}}(m) = \frac{\boldsymbol{\Phi}_{\boldsymbol{nn}}(m)\,\boldsymbol{e}}{\phi_N(m)}, \quad (12)$$

with $\phi_Y(m)$ and $\phi_N(m)$ denoting the noisy speech PSD and the noise PSD, respectively. Using (6), it can be easily shown that

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

such that

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

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

3. MFMVDR FILTER

In this section, we review the single-microphone MFMVDR filter proposed in [3, 4] and the estimation of the noisy speech correlation matrix and the speech correlation vector proposed in [8].

3.1. Optimization Problem

The MFMVDR filter aims at minimizing the total signal output power while not distorting the correlated speech component, i.e.,

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

Solving this optimization problem yields the MFMVDR filter [3, 4]

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

with the signal output power $\phi_Y^{\text{out}}(m) = \mathbb{E}\big[|\boldsymbol{h}_{\text{MFMVDR}}^H(m)\boldsymbol{y}(m)|^2\big]$, i.e.,

$$\phi_Y^{\text{out}}(m) = \frac{1}{\boldsymbol{\gamma}_{\boldsymbol{x}}^H(m)\boldsymbol{\Phi}_{\boldsymbol{yy}}^{-1}(m)\boldsymbol{\gamma}_{\boldsymbol{x}}(m)}. \quad (17)$$

The MFMVDR filter in (16) is a function of the speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ and the noisy speech correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{yy}}(m)$. Typically, both quantities are highly time-varying, making it difficult to accurately estimate especially the speech correlation vector when only having access to the noisy speech signal.

3.2. Estimation of the Noisy Speech Correlation Matrix

Estimating the noisy speech correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{yy}}(m)$ can be performed by applying first-order recursive smoothing with smoothing parameter α_y, i.e.,

$$\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(m) = \alpha_y\hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(m-1) + (1-\alpha_y)\boldsymbol{y}(m)\boldsymbol{y}^H(m). \quad (18)$$

To improve the robustness of the MFMVDR filter, we apply diagonal loading with a regularization parameter of 0.04 as in [8], before computing the inverse matrix.

3.3. Estimation of the Speech Correlation Vector

The optimal speech correlation vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{Opt}}(m)$, can be estimated as

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{Opt}}(m) = \frac{\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(m)\boldsymbol{e}}{\boldsymbol{e}^T\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(m)\boldsymbol{e}} \quad (19)$$

where the speech correlation matrix is estimated as $\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(m) = \hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(m) - \hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(m)$. The noise correlation matrix $\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(m)$ is estimated similarly to (18). Based on (14), the ML estimate of the speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ was proposed in [8], which is given by

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{ML}}(m) = \frac{\hat{\xi}(m)+1}{\hat{\xi}(m)}\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m) - \frac{1}{\hat{\xi}(m)}\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}, \quad (20)$$

with $\hat{\xi}(m)$ an estimate of the a-priori SNR. In comparison to (19), the estimated noise correlation vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}(m)$ is assumed to be constant for all time-frequency points, such that it can be replaced by its mean value $\boldsymbol{\mu}_{\boldsymbol{\gamma}_{\boldsymbol{n}}}$, which is determined by the frame overlap and the STFT analysis window [8]. To estimate the a-priori SNR $\hat{\xi}(m)$ we use the noise PSD estimator proposed in [14], i.e.,

$$\hat{\phi}_N(m) = \min\Big[\hat{\phi}_Y(m), \hat{\phi}_N(m-1)\Big](1+\nu), \quad (21)$$

where the parameter ν is set to 5 dB/s as in [8]. For the speech PSD we use the ML estimator proposed in [15], i.e.,

$$\hat{\phi}_X(m) = \max\Big[\hat{\phi}_Y(m) - \hat{\phi}_N(m), 0\Big]. \quad (22)$$

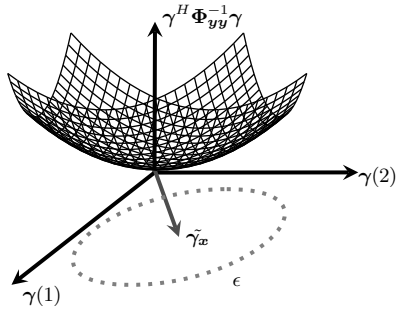

Fig. 1: Quadratic cost function in (25) with exemplary presumed speech correlation vector $\tilde{\gamma}_x$ and bound ϵ.

It should be noted that the ML estimate of the speech correlation vector strongly depends on the a-priori SNR estimate $\hat{\xi}(m)$. Especially for low a-priori SNRs, the ML estimate may become very large, such that the estimation error between $\gamma_x(m)$ and $\hat{\gamma}_x^{\text{ML}}(m)$ may become very large. This may cause the ML-based MFMVDR filter to result in unpleasant artifacts in the background noise or introduce speech distortion [8, 16].

4. ROBUST CONSTRAINED MFMVDR FILTER

In [9] we showed that the performance of the MFMVDR filter is very sensitive to estimation errors of the speech correlation vector, such that correlated speech components may be mistakenly interpreted as uncorrelated and be suppressed rather than preserved. Inspired by the robust MVDR beamformer in [10], in this section we propose to estimate the speech correlation vector as the vector maximizing the total signal output power of the MFMVDR filter within a spherical uncertainty set. For conciseness, also the index m is omitted in this section. It should be noted that the calculations are performed for each frequency-bin k and time-frame m.

Given a presumed speech correlation vector $\tilde{\gamma}_x$, e.g., the ML estimate $\hat{\gamma}_x^{\text{ML}}$ in (20), the mismatch vector to the (unknown) speech correlation vector γ_x is defined as $\delta_x = \gamma_x - \tilde{\gamma}_x$, with $\epsilon_x = ||\gamma_x - \tilde{\gamma}_x||_2^2$. We now define the spherical uncertainty set comprising all vectors whose squared distance to the presumed speech correlation vector $\tilde{\gamma}_x$ is smaller than or equal to a bound ϵ, i.e.,

$$\mathbf{\Gamma} = \{\, \gamma = \tilde{\gamma}_x + \delta \mid \|\delta\|_2^2 \leq \epsilon \,\}. \tag{23}$$

Similarly to the method proposed in [10] to robustly estimate the steering vector for the MVDR beamformer, the RC speech correlation vector is computed as the vector maximizing the total signal output power of the MFMVDR filter in (17) within the spherical uncertainty set in (23), i.e.,

$$\hat{\gamma}_x^{\text{RC}} = \underset{\gamma}{\text{argmax}} \; \min_{h} \; h^H \Phi_{yy} h, \text{ s.t. } h^H \gamma = 1, \quad \|\gamma - \tilde{\gamma}_x\|_2^2 \leq \epsilon. \tag{24}$$

Using (17), this optimization problem can be reformulated as

$$\boxed{\hat{\gamma}_x^{\text{RC}} = \underset{\gamma}{\text{argmin}} \; \gamma^H \Phi_{yy}^{-1} \gamma, \text{ s.t. } \|\gamma - \tilde{\gamma}_x\|_2^2 \leq \epsilon} \tag{25}$$

For $L = 2$, the quadratic cost function $\gamma^H \Phi_{yy}^{-1} \gamma$ in (25) is visualized in Fig. 1 for an exemplary noisy speech correlation matrix Φ_{yy}, together with an exemplary presumed speech correlation vector $\tilde{\gamma}_x$ and bound ϵ. Obviously, the bound ϵ in (25) plays an important role and should be chosen in accordance with the accuracy of the presumed speech correlation vector $\tilde{\gamma}_x$, i.e., if $||\gamma_x - \tilde{\gamma}_x||_2^2$ is small, then ϵ should be small, whereas if $||\gamma_x - \tilde{\gamma}_x||_2^2$ is large, then ϵ should be large.

In order to avoid the (undesired) solution $\hat{\gamma}_x^{\text{RC}} = \mathbf{0}$, the bound ϵ should be chosen such that

$$\epsilon < \|\tilde{\gamma}_x\|_2^2. \tag{26}$$

Under this condition and considering the convex nature of the quadratic cost function in (25), the inequality constraint in (25) can be replaced by an equality constraint, i.e.,

$$\hat{\gamma}_x^{\text{RC}} = \underset{\gamma}{\text{argmin}} \; \gamma^H \Phi_{yy}^{-1} \gamma, \text{ s.t. } \|\gamma - \tilde{\gamma}_x\|_2^2 = \epsilon. \tag{27}$$

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

$$f(\gamma, \lambda) = \gamma^H \Phi_{yy}^{-1} \gamma + \lambda \left(\|\gamma - \tilde{\gamma}_x\|_2^2 - \epsilon \right), \tag{28}$$

with λ the Lagrange multiplier. Setting the gradient of $f(\gamma, \lambda)$ with respect to γ equal to zero and applying the matrix inversion lemma, we obtain the *RC speech correlation vector* $\hat{\gamma}_x^{\text{RC}}(\lambda)$ as

$$\boxed{\hat{\gamma}_x^{\text{RC}}(\lambda) = \tilde{\gamma}_x - (\lambda \Phi_{yy} + I)^{-1} \tilde{\gamma}_x} \tag{29}$$

with I denoting the $L \times L$-dimensional identity matrix. Setting the derivative of $f(\gamma, \lambda)$ with respect to λ equal to zero and substituting (29) results in

$$\frac{\partial f(\gamma, \lambda)}{\partial \lambda} = \|(\lambda \Phi_{yy} + I)^{-1} \tilde{\gamma}_x\|_2^2 - \epsilon = 0. \tag{30}$$

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

$$\Phi_{yy} = Q U Q^H, \tag{31}$$

where the columns of Q contain the orthogonal eigenvectors of Φ_{yy} and the diagonal elements of U are the corresponding eigenvalues, with $u_0 \geq u_1 \geq \ldots \geq u_{L-1}$. In addition, let

$$z_x = Q^H \tilde{\gamma}_x, \tag{32}$$

where $z_x(l)$ denotes the lth element of z_x. Using (31) and (32) in (30), we obtain

$$\boxed{g(\lambda) = \sum_{l=0}^{L-1} \frac{|z_x(l)|^2}{(1 + \lambda u_l)^2} = \epsilon} \tag{33}$$

This non-linear equation in the Lagrange multiplier λ can be solved, e.g., using Newton's method. The solution is then used in (29), yielding the RC speech correlation vector $\hat{\gamma}_x^{\text{RC}}$. Since the condition in (9) may not be satisfied, possibly resulting in a scaling inaccuracy, a normalization is performed by dividing $\hat{\gamma}_x^{\text{RC}}$ with its first element. Using the normalized RC speech correlation vector $\hat{\gamma}_x^{\text{RC}}$ in (16) results in the *RC MFMVDR filter*.

As already mentioned, the bound ϵ should be chosen in accordance with the accuracy of the presumed speech correlation vector, e.g., assuming that the optimal bound is equal to $\hat{\epsilon}^{\text{Opt}} = \|\hat{\gamma}_x^{\text{Opt}} - \tilde{\gamma}_x\|_2^2$ with the optimal estimate of the speech correlation vector defined in (19). When using the ML estimate $\hat{\gamma}_x^{\text{ML}}$ as the presumed speech correlation vector $\tilde{\gamma}_x$, simulations have shown that the accuracy of the ML estimate (not unexpectedly) depends on the a-priori SNR estimate $\hat{\xi}$. Hence, we propose to train a mapping function $\hat{\epsilon}_{\text{ML}}^{\text{Map}}(\hat{\xi})$ between the a-priori SNR estimate $\hat{\xi}$, computed using (21) and (22), and the optimal bound $\hat{\epsilon}_{\text{ML}}^{\text{Opt}} = \|\hat{\gamma}_x^{\text{Opt}} - \hat{\gamma}_x^{\text{ML}}\|_2^2$. Fig. 2 shows the normalized joint probability density function (PDF) of the optimal bound $\hat{\epsilon}_{\text{ML}}^{\text{Opt}}$ and the a-priori SNR estimate $\hat{\xi}$ for a wide range of speech and noise signals (30 TIMIT sentences [17], speech-shaped noise, two traffic and babble noise signals), for a broadband SNR range of 0 to 15 dB. It can be observed that with increasing a-priori SNR the optimal bound decreases. The linear mapping function $\hat{\epsilon}_{\text{ML}}^{\text{Map}}(\hat{\xi})$ (shown in red in Fig. 2) is based on the maximum value of the normalized PDF for each a-priori SNR estimate $\hat{\xi}$.

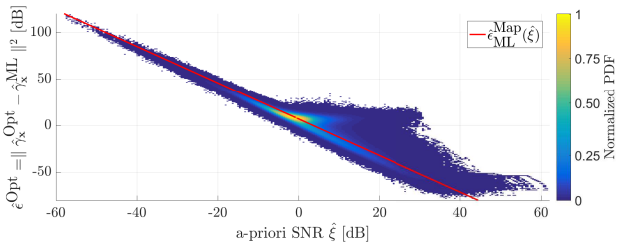

Fig. 2: Normalized joint PDF of the optimal uncertainty parameter $\hat{\epsilon}_{\text{ML}}^{\text{Opt}}$ and the estimated a-priori SNR $\hat{\xi}$.

5. EXPERIMENTAL EVALUATION

In this section, we compare the estimation accuracy error of the proposed RC speech correlation vector in (29) with the ML speech correlation vector in (20) and the performance of the MFMVDR filter when using these speech correlation vectors.

Similarly as in [8, 9], to increase the exploitable speech correlation across time-frames, we use a highly temporally resolved STFT framework with a frame length of 4 ms 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. The number of consecutive time-frames is set to $L = 18$, resulting in 21 ms of data used in each filtering operation. The smoothing parameter in (18) is set to $\alpha_y = 0.9$. As clean speech signals, we use 60 sentences from the TIMIT database [17], spoken by different speakers (10 male, 10 female). As noise signals we use modulated white Gaussian noise, traffic and two babble noise signals. The sampling frequency is 16 kHz and the considered SNR range is -5 dB to 20 dB. We make sure that the evaluation data differs from the training set.

The both estimated speech correlation vectors is evaluated using the mean-square error (MSE) between the optimal speech correlation vector in (19) and the estimated speech correlation vector, i.e.,

$$\text{MSE} = \frac{1}{\mathbb{F}} \sum_{k,m \in \mathbb{F}} \| \hat{\gamma}_{\boldsymbol{x}}^{\text{Opt}}(k,m) - \hat{\gamma}_{\boldsymbol{x}}(k,m) \|_2^2, \tag{34}$$

where $\mathbb{F}$ is the set of time-frequency points that contain either noise-only or speech-and-noise points (with a-priori SNR $\xi(m)$ larger than -5 dB). Furthermore, we classify time-frequency points whose squared error is larger than 200 as outliers and exclude them from the MSE calculation. In Fig. 3, the performance averaged over all speech and noise files in terms of the MSE and the percentage of outliers in speech-and-noise and noise-only points are shown for different SNRs, separately. It can be observed that compared to the ML estimate the RC estimate yields a lower MSE. Moreover, the percentage of outliers is removed, indicating that the RC estimate is more accurate and much stabler as the ML estimate.

The performance of the MFMVDR filter using both estimates is evaluated in terms speech distortion and noise reduction using the segmental speech SNR (segSSNR) and the segmental noise reduction (segNR) [18], where both measures have only been computed during time-frames where speech is active. In addition, to evaluate the noise distortion, more in particular the presence of musical noise, we use the weighted log kurtosis ratio $\Delta\Psi_{\log}$ [19], where the musical noise is lowest at $\Delta\Psi_{\log} = 0$.

Fig. 4 depicts the results averaged over all speech and noise files. On the one hand, it can be observed that for all SNRs the proposed RC MFMVDR filter achieves a larger segSSNR and lower $\Delta\Psi_{\log}$ than the ML MFMVDR filter. On the other hand, in terms of segNR it can be seen that for SNRs up to 10 dB the proposed RC MFMVDR filter achieves a lower noise reduction than the ML MFMVDR filter. These results indicate that the RC MFMVDR filter leads to clearly less speech and noise distortion than the ML MFMVDR filter but is more conservative in suppressing

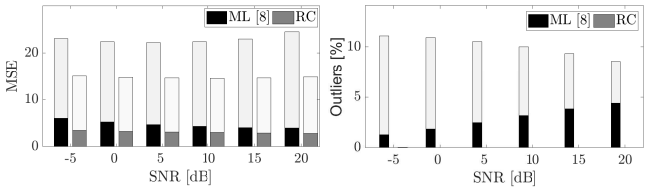

Fig. 3: Average MSE and percentage of outliers for the ML and RC speech correlation vectors. The lower and upper parts of the bars represent the performance in speech-and-noise and noise-only, respectively.

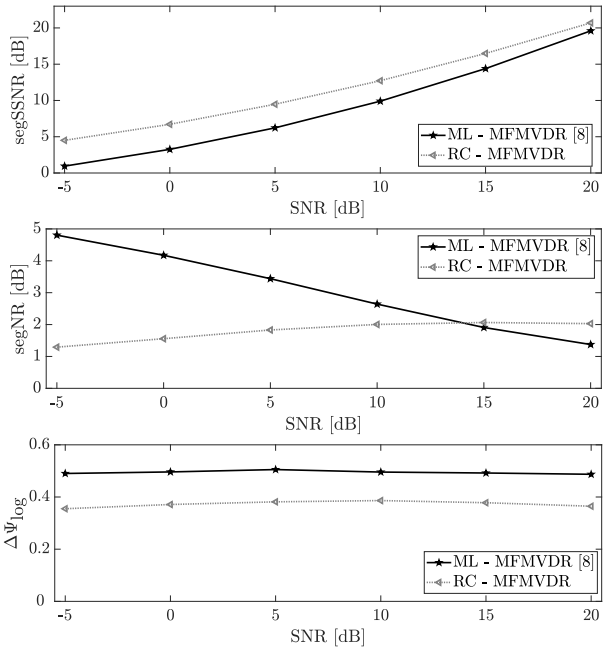

Fig. 4: Average segNR, segSSNR and $\Delta\Psi_{\log}$ using the ML estimate $\hat{\gamma}_{\boldsymbol{x}}^{\text{ML}}(m)$ and the RC estimate $\hat{\gamma}_{\boldsymbol{x}}^{\text{RC}}(m)$ for different SNRs.

the background noise. Informal listening tests confirm that for the RC MFMVDR filter the speech sounds clearly less distorted and more natural and less musical noise is present than for the ML MFMVDR filter.

6. CONCLUSIONS

In this paper, we proposed a robust constrained (RC) multi-frame minimum variance distortionless response (MFMVDR) filter for single-channel speech enhancement. Inspired by robust beamforming approaches, we proposed to estimate the speech correlation vector as the vector maximizing the total signal output power within a spherical uncertainty set. The spherical uncertainty set imposes an upper bound on the norm of the mismatch vector between the speech correlation vector and the presumed speech correlation vector. We proposed to set this bound by training a mapping function depending on the a-priori SNR estimate. Simulation results show that the proposed RC approach leads to a more accurate and stable estimate of the speech correlation vector than the ML approach. The RC MFMVDR filter produces less speech and noise distortions than the ML MFMVDR filter such that the speech sounds more natural and less musical noise is present but the RC MFMVDR filter leads to a more conservative noise reduction performance than the ML MFMVDR filter.

7. REFERENCES

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

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

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

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

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

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

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

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

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

[10] 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.

[11] 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.

[12] S. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *IEEE Trans. Signal Process.*, vol. 93, no. 12, pp. 3264–3277, Dec. 2013.

[13] 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. (WASPAA)*, New Paltz, NY, USA, 2017, pp. 86–90.

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

[15] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

[16] 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 Conference on Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 292–296.

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

[18] T. Lotter and P. 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.

[19] H. Yu and T. Fingscheidt, "A weighted log kurtosis ratio measure for instrumental musical tones assessment in wideband speech," in *Proc. ITG Conference on Speech Commun.*, Braunschweig, Germany, Sep. 2012, pp. 1–4.