Evaluation of Robust Constrained MFMVDR Filtering for Single-Channel Speech Enhancement

Dörte Fischer, Simon Doclo [1].

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany.
Email:{doerte.fischer, simon.doclo}@uni-oldenburg.de

Abstract

By considering the multi-frame signal model, speech correlation between different time-frames can be exploited. Based on this signal model, the multi-frame minimum variance distortionless response (MFMVDR) filter for single-channel speech enhancement has been derived, which minimizes the total signal output power while avoiding speech distortion. It has been shown that the MFMVDR filter is very sensitive to estimation errors in the speech correlation vector resulting in correlated speech components being mistakenly suppressed. Inspired by robust beamforming approaches, in this paper we propose a robust constrained MFMVDR filter for single-channel 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 signal-to-noise ratio (SNR). Experimental results for different noise types and SNRs show that the proposed robust approach yields a more accurate estimate of the speech correlation vector. A perceptual evaluation shows that the robust constrained MFMVDR filter leads to an improved speech quality but a lower noise reduction than the original non-robust MFMVDR filter, while still being preferred in overall quality.

1 Introduction

Using speech communication devices (e.g., hearing aids) in noisy environments may lead to a degraded speech quality and speech intelligibility. In these acoustic situations, speech enhancement algorithms are required to suppress the undesired noise while limiting speech distortion. Typically, single-channel 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 [2–5].

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-channel 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], the *multi-frame MVDR* (MFMVDR) filter for single-channel speech enhancement was derived in [3, 4], minimizing 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 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.

[1]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)

In the area of array processing several techniques have been proposed to increase the robustness of beamformers against estimation errors of the steering vector [10–13]. Inspired by the robust MVDR beamformer in [10], which estimates the steering vector that maximizes the total signal output power of the MVDR beamformer within a spherical uncertainty set, in this paper we propose a *robust constrained (RC) MFMVDR* filter by estimating the speech correlation vector that maximizes the total signal output power of the MFMVDR filter within a *spherical uncertainty set*. This spherical uncertainty 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) showed that the norm of the mismatch vector decreases with increasing SNR, we trained a mapping function to set the upper bound depending on the a-priori SNR for each time-frequency point. Evaluation results for different speech signals, noise types and SNRs show that the proposed RC speech correlation vector achieves a lower mean-square error than the ML estimate. Based on a pairwise preference listening experiment with 15 participants, we show that the proposed RC MFMVDR filter results in a significantly better speech quality than the ML MFMVDR filter although the amount of noise reduction is significantly lower. Nevertheless, the quality of the RC MFMVDR filter is preferred over the overall quality of the ML MFMVDR filter.

2 Multi-Frame Signal Model

We consider a single-channel 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

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

with $X(k,m)$ and $N(k,m)$ denoting the speech and the noise coefficients, respectively. For conciseness, in the remainder of the paper the frequency-bin k will be omitted where possible.

In multi-frame single-channel 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 $\boldsymbol{h}(m)$ contains L time-varying filter coefficients, i.e.,

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

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

Using (1), the noisy speech vector $\boldsymbol{y}(m)$ is given by

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

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

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

ISBN 978-3-8007-4767-2

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.,

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

with

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

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

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

where * denotes the complex-conjugate operator and $\boldsymbol{e} = [1, 0, ..., 0]^T$ is an L-dimensional selection vector. Through the normalization with $\boldsymbol{e}^T\boldsymbol{\Phi}_{\boldsymbol{xx}}(m)\boldsymbol{e}$, 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.,

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

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

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

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

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

with $\boldsymbol{e}^T\boldsymbol{\Phi}_{\boldsymbol{yy}}(m)\boldsymbol{e}$ and $\boldsymbol{e}^T\boldsymbol{\Phi}_{\boldsymbol{nn}}(m)\boldsymbol{e}$ corresponding to the noisy speech PSD $\phi_Y(m)$ and the noise PSD $\phi_N(m)$, respectively. Using (6), it can be easily shown that

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

such that

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

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

3 MFMVDR Filter

In this section, we briefly review the single-channel 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.,

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

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

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

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.,

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

As can be seen from (16), the MFMVDR filter requires an estimate 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.,

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

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

3.3 Estimation of the Speech Correlation Vector

Based on (9) and (14), the optimal estimate of the speech correlation vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{Opt}}(m)$ is given by

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

$$= \frac{\hat{\xi}^{\text{Opt}}(m)+1}{\hat{\xi}^{\text{Opt}}(m)}\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m) - \frac{1}{\hat{\xi}^{\text{Opt}}(m)}\hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}(m), \tag{19b}$$

with $\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(m) = \hat{\boldsymbol{\Phi}}_{\boldsymbol{yy}}(m) - \hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(m)$ the estimated speech correlation matrix, where the noise correlation matrix $\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(m)$ is computed similarly as in (18). The estimates of the noisy speech correlation vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{y}}(m)$ and the noise correlation vector $\hat{\boldsymbol{\gamma}}_{\boldsymbol{n}}(m)$ can be computed similarly as in (19a) and the optimal estimate of the a-priori SNR is given by $\hat{\xi}^{\text{Opt}}(m) = (\boldsymbol{e}^T\hat{\boldsymbol{\Phi}}_{\boldsymbol{xx}}(m)\boldsymbol{e})/(\boldsymbol{e}^T\hat{\boldsymbol{\Phi}}_{\boldsymbol{nn}}(m)\boldsymbol{e})$.

Similarly to (19), the ML estimate of the speech correlation vector $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ proposed in [8] is given by

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

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.,

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

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.,

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

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 $\boldsymbol{\gamma}_{\boldsymbol{x}}(m)$ and $\hat{\boldsymbol{\gamma}}_{\boldsymbol{x}}^{\text{ML}}(m)$ 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].

ISBN 978-3-8007-4767-2

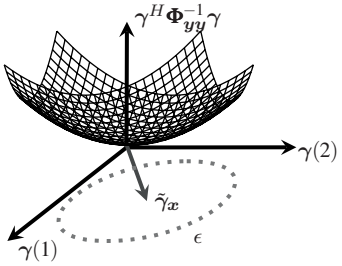

Figure 1: Quadratic cost function in (25) with exemplary presumed speech correlation vector $\tilde{\gamma}_{\boldsymbol{x}}$ and bound ϵ.

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 $\tilde{\gamma}_{\boldsymbol{x}}$, e.g., the ML estimate $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{ML}}$ in (20), the mismatch vector to the (unknown) speech correlation vector $\gamma_{\boldsymbol{x}}$ is defined as $\boldsymbol{\delta}_{\boldsymbol{x}} = \gamma_{\boldsymbol{x}} - \tilde{\gamma}_{\boldsymbol{x}}$, with $\epsilon_{\boldsymbol{x}} = \|\gamma_{\boldsymbol{x}} - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2$. We now define the spherical uncertainty set comprising all vectors whose squared distance to the presumed speech correlation vector $\tilde{\gamma}_{\boldsymbol{x}}$ is smaller than or equal to a bound ϵ, i.e.,

$$\boldsymbol{\Gamma} = \left\{ \gamma = \tilde{\gamma}_{\boldsymbol{x}} + \boldsymbol{\delta} \,|\, \|\boldsymbol{\delta}\|_2^2 \leq \epsilon \right\}. \tag{23}$$

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.,

$$\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}} = \underset{\gamma}{\operatorname{argmax}} \; \min_{\boldsymbol{h}} \; \boldsymbol{h}^H \boldsymbol{\Phi}_{\boldsymbol{yy}} \boldsymbol{h}, \;\; \text{s.t. } \boldsymbol{h}^H \gamma = 1, \;\; \|\gamma - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2 \leq \epsilon. \tag{24}$$

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

$$\boxed{\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}} = \underset{\gamma}{\operatorname{argmin}} \; \gamma^H \boldsymbol{\Phi}_{\boldsymbol{yy}}^{-1} \gamma, \;\; \text{s.t. } \|\gamma - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2 \leq \epsilon} \tag{25}$$

For $L = 2$, the quadratic cost function $\gamma^H \boldsymbol{\Phi}_{\boldsymbol{yy}}^{-1} \gamma$ in (25) is visualized in Figure 1 for an exemplary noisy speech correlation matrix $\boldsymbol{\Phi}_{\boldsymbol{yy}}$, together with an exemplary presumed speech correlation vector $\tilde{\gamma}_{\boldsymbol{x}}$ 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 $\tilde{\gamma}_{\boldsymbol{x}}$, i.e., if $\|\gamma_{\boldsymbol{x}} - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2$ is small, then ϵ should be small, whereas if $\|\gamma_{\boldsymbol{x}} - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2$ is large, then ϵ should be large.

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

$$\epsilon < \|\tilde{\gamma}_{\boldsymbol{x}}\|_2^2. \tag{26}$$

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.,

$$\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}} = \underset{\gamma}{\operatorname{argmin}} \; \gamma^H \boldsymbol{\Phi}_{\boldsymbol{yy}}^{-1} \gamma, \;\; \text{s.t. } \|\gamma - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2 = \epsilon. \tag{27}$$

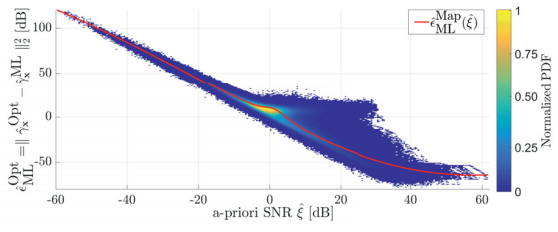

Figure 2: Normalized joint PDF of the optimal uncertainty parameter $\hat{\epsilon}_{\mathrm{ML}}^{\mathrm{Opt}}$ and the estimated a-priori SNR $\hat{\xi}$.

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

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

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* $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}}(\lambda)$ as

$$\boxed{\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}}(\lambda) = \tilde{\gamma}_{\boldsymbol{x}} - (\lambda \boldsymbol{\Phi}_{\boldsymbol{yy}} + \boldsymbol{I})^{-1} \tilde{\gamma}_{\boldsymbol{x}}} \tag{29}$$

with $\boldsymbol{I}$ 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

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

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

$$\boldsymbol{\Phi}_{\boldsymbol{yy}} = \boldsymbol{Q} \boldsymbol{U} \boldsymbol{Q}^H, \tag{31}$$

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

$$\boldsymbol{z}_{\boldsymbol{x}} = \boldsymbol{Q}^H \tilde{\gamma}_{\boldsymbol{x}}, \tag{32}$$

where $\boldsymbol{z}_{\boldsymbol{x}}(l)$ denotes the lth element of $\boldsymbol{z}_{\boldsymbol{x}}$. Using (31) and (32) in (30), we obtain

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

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 $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}}$. Since the condition in (9) may not be satisfied, possibly resulting in a scaling inaccuracy, a normalization is performed by dividing $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}}$ with its first element. Using the normalized RC speech correlation vector $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{RC}}$ 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 $\hat{\epsilon}^{\mathrm{Opt}} = \|\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Opt}} - \tilde{\gamma}_{\boldsymbol{x}}\|_2^2$ with the optimal estimate of the speech correlation vector defined in (19). When using the ML estimate $\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{ML}}$ as the presumed speech correlation vector $\tilde{\gamma}_{\boldsymbol{x}}$, 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 $\hat{\epsilon}_{\mathrm{ML}}^{\mathrm{Map}}(\hat{\xi})$ between the a-priori SNR estimate $\hat{\xi}$, computed using (21) and (22), and the optimal bound $\hat{\epsilon}_{\mathrm{ML}}^{\mathrm{Opt}} = \|\hat{\gamma}_{\boldsymbol{x}}^{\mathrm{Opt}} - \hat{\gamma}_{\boldsymbol{x}}^{\mathrm{ML}}\|_2^2$. Figure 2 shows the normalized joint probability density function (PDF) of the optimal bound $\hat{\epsilon}_{\mathrm{ML}}^{\mathrm{Opt}}$ 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 mapping function $\hat{\epsilon}_{\mathrm{ML}}^{\mathrm{Map}}(\hat{\xi})$ (shown in red in Figure 2) is based on the maximum value of the normalized PDF for each a-priori SNR estimate $\hat{\xi}$.

ISBN 978-3-8007-4767-2

5 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). Furthermore, we evaluate the perceptual quality of both MFMVDR filters, i.e., the proposed RC MFMVDR filter and the ML MFMVDR filter.

5.1 STFT Framework

Similarly as in [8], 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 sampling frequency is 16 kHz. 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 $\alpha_y = 0.9$, resulting in a smoothing window of 10 ms.

5.2 Evaluation of the Speech Correlation Vector Estimation

The performance of both estimated speech correlation vectors, i.e., the RC estimator in (29) and the ML estimator in (20), is evaluated in terms of the mean-square error (MSE) between the optimal speech correlation vector in (19) and the estimated speech correlation vector, i.e.,

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

where $\mathbb{F}$ is the set of time-frequency points that contain either noise-only or speech-and-noise (i.e., with a-priori SNR $\hat{\xi}^{\text{Opt}}(m)$ 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. 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 speech-shaped noise, traffic and babble noise signals. The considered SNR range is -5 dB to 20 dB. We make sure that the evaluation data differs from the training set.

For different SNRs, Figure 3 depicts the performance (averaged over all combinations of speech and noise signals) in terms of the MSE and the percentage of outliers in speech-and-noise and noise-only points. It can be observed that compared to the ML estimate the RC estimate yields a considerably lower MSE. Moreover, the outliers are completely removed, indicating that the RC estimate is more accurate and much stabler than the ML estimate.

5.3 Perceptual Evaluation

The perceptual evaluation is performed using a pairwise preference test, where the quality of both MFMVDR filters, i.e., the proposed RC MFMVDR filter and the ML MFMVDR filter, and the unprocessed noisy speech signal is compared to each other. As speech signals, we used two TIMIT sentences, spoken by a female and a male speaker. As noise signals, we used speech-shaped noise, traffic and babble noise at an SNR of 0 dB. For each of the three noise types 15 self-reported normal hearing listeners aged between 21 and 38 were asked to judge which of the two presented signals is preferred in terms of speech quality (SQ), noise reduction (NR) and overall quality (OQ). The order of the signal pairs as well as the order of the noise types was randomized. As a reference, the clean speech signal was also available. The signals were presented diotically over Sennheiser HDA 200 headphones using an RME Babyface sound card in a quiet office. Similarly as in [18], the ratings are expressed in the percentage of times each algorithm was preferred, i.e., a value of 100% indicates that the algorithm won every pairwise comparison it was involved in. To determine statistical significance, we used the non-parametric Friedman test followed by a Wilcoxon signed-rank test with Bonferroni correction. We assume statistical significance for $p<0.05$.

Since the ratings between the different noise types were not statistically significant, Figure 4 shows the preference results averaged over all three noise types for the three evaluation criteria. In terms of the speech quality, the proposed RC MFMVDR filter and the unprocessed noisy speech signal are preferred over the ML MFMVDR filter, while there is no statistically significant difference the RC MFMVDR filter and the noisy speech signal. In terms of the noise reduction, the ML MFMVDR filter is preferred over the RC MFMVDR filter but the RC MFMVDR filter is still preferred over the noisy speech signal. In terms of the overall quality, the RC MFMVDR filter is preferred over the ML MFMVDR filter and the noisy signal, while there is no statistically significant difference between the ML MFMVDR filter and the noisy speech signal.

The results indicate that the RC MFMVDR filter leads to less speech distortion and a more natural speech quality than the ML MFMVDR filter. Although the RC MFMVDR filter is more conservative in suppressing the background noise than the ML MFMVDR filter, the proposed RC approach leads to less musical noise such that the overall signal quality is significantly preferred over the ML MFMVDR filter.

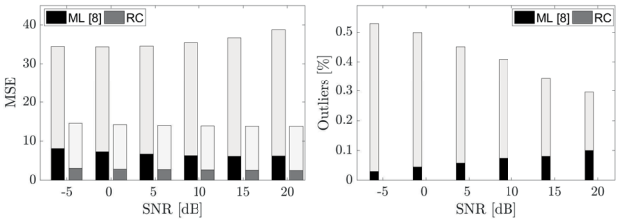

Figure 3: Average MSE and percentage of outliers for the ML and RC speech correlation vectors for different SNRs. The lower and upper parts of the bars represent the performance in speech-and-noise and noise-only, respectively.

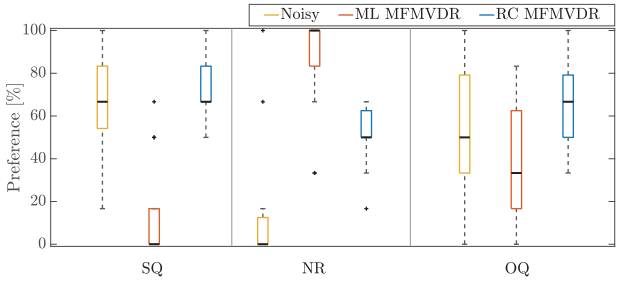

Figure 4: Boxplots of the average preference ratings of the unprocessed noisy speech signal, the ML MFMVDR filter and the proposed RC MFMVDR filter for the criteria speech quality, noise reduction and overall quality.

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. Furthermore, a pairwise preference test indicates that although the proposed RC MFMVDR filter leads to a more conservative noise reduction than the ML MFMVDR filter, it produces less speech and noise distortions, such that the speech sounds more natural and less musical noise is present.

ISBN 978-3-8007-4767-2

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), pp. 273–276, May 2011.

[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, 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, 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, pp. 4–24, Apr. 1988.

[7] 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 Processing Magazine*, vol. 32, 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, pp. 1355–1365, Sept. 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), pp. 603–607, Aug. 2017.

[10] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, pp. 1702–1715, July 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, pp. 2974–2987, Feb. 2012.

[12] S. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *IEEE Trans. Signal Process.*, vol. 93, 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), pp. 86–90, 2017.

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

[15] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, 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), pp. 292–296, Oct. 2016.

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

[18] M. Krawczyk-Becker and T. Gerkmann, "An evaluation of the perceptual quality of phase-aware single-channel speech enhancement," *J. Acoust. Soc. Amer.*, vol. 140, pp. EL364–EL369, Oct. 2016.

ISBN 978-3-8007-4767-2