Reduced-Complexity Semi-Distributed Multi-Channel Multi-Frame MVDR Filter

Raziyeh Ranjbaryan and Hamid Reza Abutalebi
Electrical Engineering Department
Yazd University, Yazd, Iran
Emails: ranjbaryan@stu.yazd.ac.ir, habutalebi@yazd.ac.ir

Simon Doclo
Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4All
University of Oldenburg, Germany
Email: simon.doclo@uni-oldenburg.de

***Abstract*—In this paper, we propose a semi-distributed multi-channel noise reduction method which considers the inter-frame correlation in the short-time Fourier transform (STFT) domain. Although exploiting the correlation of speech STFT coefficients enables to achieve impressive results, it also increases the computational complexity, especially in the case of a large number of frames and/or microphones. To address this issue in each time-frequency unit we propose to utilize the information of the current frame and a compressed signal from the previous frames in a distributed way. Simulation results show that the computational complexity can be substantially reduced by the proposed method without impairing speech quality.**

I. Introduction

Undesired background noise makes speech communication unpleasant or in some cases, even impossible. Extracting the clean speech signal from the noisy observed microphone signals has been an important research topic due to its applications in, e.g., hands-free communication, teleconferencing, and hearing aids.

Noise reduction algorithms are usually implemented in the short-time Fourier transform (STFT) domain. A common assumption which simplifies the processing is that successive speech frames are uncorrelated. However, it is well known that this assumption is not accurate and that the inherent correlation of the speech signal in addition to the overlap procedure applied in the STFT processing introduces a large correlation between consecutive frames [1].

In [2] the authors exploited the inter-frame correlation (IFC) in the STFT domain for single-microphone noise reduction. In each time-frequency unit (TFU) they employed the noisy observation of the current and previous frames. Consequently, they developed a single-channel multi-frame minimum variance distortion-less response (MVDR) filter, showing that impressive results in terms of signal-to-noise ratio (SNR) improvement and low signal distortion can be achieved if the IFC can be accurately estimated. In [3] blind maximum-likelihood and maximum *a-posteriori* estimators for the speech IFC vector have been proposed. In addition, to set the trade-off between speech distortion and noise reduction, speech-distortion weighted inter-frame Wiener filters have been proposed in [4].

This work was supported in part by Iran National Science Foundation (INSF) and German Academic Exchange Service (DAAD).

The concept of exploiting the IFC has also been considered in multi-microphone noise reduction. In this way, in each TFU the received signals from different microphones at the current and previous frames are taken into account to improve the overall performance. It has been shown that the noise reduction performance provided by the multi-channel multi-frame algorithms is better than the improvement achieved by the multi-channel algorithms, which ignore the IFC [1]. Although exploiting the IFC results in SNR improvement, it also increases the computational complexity, especially in the case of a large number of frames and/or microphones. In order to overcome this challenge, in this paper we propose a reduced-complexity multi-channel multi-frame MVDR filter inspired by existing distributed algorithms in wireless acoustic sensor networks (WASNs).

WASNs consist of several spatially distributed nodes, where each node contains a small-sized array of sensors which connects to the other nodes via a wireless link. In a centralized algorithm, all nodes transmit their data to a fusion center that combines all information. The fusion center hence has direct access to all data of the network, enabling it to achieve a better performance than each individual node. However, this comes at the cost of a larger required bandwidth and computational complexity, especially in the case of a large number of nodes. Hence, distributed algorithms have been proposed in order to decrease the complexity and the required bandwidth. In these algorithms, nodes share data between themselves, such that there is no need for a fusion center.

In [5] the distributed multi-channel Wiener filter (DB-MWF) was proposed for two nodes. The DB-MWF was generalized in [6] to multiple nodes, leading to the so-called distributed adaptive node specific estimation (DANSE) algorithm. In addition, a distributed version of the generalized sidelobe canceller (GSC) was proposed in [7]. A comprehensive review of distributed algorithms for noise reduction in WASNs has been presented in [8], indicating that the main idea of all mentioned techniques is similar: each node performs calculations using its local information and a compressed signal from the other nodes.

Motivated by the idea behind the above-mentioned distributed algorithms, in this paper we propose a distributed version of the multi-channel multi-frame MVDR filter which

ISBN 978-90-827970-1-5 © EURASIP 2018

takes the IFC into account. In order to realize causal processing, we assume that in each TFU the filter has access to the information of the current frame and a compressed signal from the previous frames (hence the term "semi-distributed").

The paper is organized as follows. After formulating the multi-frame signal model in section II, we review the multi-channel multi-frame MVDR filter in section III. The proposed algorithm is introduced in section IV. Simulation results are presented in Section V.

II. Multi-Channel Multi-Frame Signal Model

Consider an N-element microphone array which captures a speech signal in a noisy environment. In the STFT domain, the received signal at the n-th microphone can be expressed as

$$Y_n(m,k) = X_n(m,k) + V_n(m,k), \tag{1}$$

where $Y_n(m,k)$, $X_n(m,k)$ and $V_n(m,k)$ denote the noisy microphone signal, the clean speech signal and the additive noise respectively, with m the frame index and k the discrete frequency index. We assume that the speech and noise signals are zero mean random processes. Without loss of generality the clean speech signal at the first microphone ($X_1(m,k)$) is considered the desired signal.
In vector notation (1) can be written as

$$\overline{\mathbf{y}}(m,k) = \overline{\mathbf{x}}(m,k) + \overline{\mathbf{v}}(m,k) \in \mathbb{C}^{N\times 1}, \tag{2}$$

where $\overline{\mathbf{y}}(m,k) = [Y_1(m,k),\ Y_2(m,k),\ ...,\ Y_N(m,k)]^T$, with ^{T} denoting the matrix transpose operation, the vectors $\overline{\mathbf{x}}(m,k)$ and $\overline{\mathbf{v}}(m,k)$ can be defined similarly. By considering the current frame and $L-1$ previous frames, we define the vector

$$\mathbf{y}(m,k) = [\overline{\mathbf{y}}^T(m,k),\ ...,\ \overline{\mathbf{y}}^T(m-L+1,k)]^T \in \mathbb{C}^{NL\times 1}, \tag{3}$$

where the vectors $\mathbf{x}(m,k)$ and $\mathbf{v}(m,k)$ can be defined similarly, such that

$$\mathbf{y}(m,k) = \mathbf{x}(m,k) + \mathbf{v}(m,k). \tag{4}$$

The clean speech signal is estimated as

$$\hat{X}(m,k) = \mathbf{h}^H(m,k)\mathbf{y}(m,k), \tag{5}$$

where $\mathbf{h}(m,k) = [\mathbf{h}_0^T(m,k),\ \mathbf{h}_1^T(m,k),\ ...,\ \mathbf{h}_{L-1}^T(m,k)]^T$ are filter coefficients and ^{H} denotes the Hermitian operator.

Assuming that the speech and noise signals are uncorrelated, the noisy correlation matrix is given by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{yy}}(m,k) &= \mathbb{E}\left\{\mathbf{y}(m,k)\mathbf{y}^H(m,k)\right\}\\ &= \mathbf{\Phi}_{\mathbf{xx}}(m,k) + \mathbf{\Phi}_{\mathbf{vv}}(m,k),\end{aligned} \tag{6}$$

where $\mathbb{E}\{.\}$ denotes the expectation operator, $\mathbf{\Phi}_{\mathbf{xx}}(m,k)$ denotes the clean speech correlation matrix and $\mathbf{\Phi}_{\mathbf{vv}}(m,k)$ denotes the noise correlation matrix.

The noisy and noise correlation matrices are estimated recursively as

$$\hat{\mathbf{\Phi}}_{\mathbf{yy}}(m,k) = \lambda_y\hat{\mathbf{\Phi}}_{\mathbf{yy}}(m-1,k) + (1-\lambda_y)\mathbf{y}(m,k)\mathbf{y}^H(m,k), \tag{7}$$

$$\hat{\mathbf{\Phi}}_{\mathbf{vv}}(m,k) = \lambda_v\hat{\mathbf{\Phi}}_{\mathbf{vv}}(m-1,k) + (1-\lambda_v)\mathbf{v}(m,k)\mathbf{v}^H(m,k), \tag{8}$$

where λ_y and λ_v denote forgetting factors. Using (6), the clean speech correlation matrix can be estimated as $\hat{\mathbf{\Phi}}_{\mathbf{xx}}(m,k) = \hat{\mathbf{\Phi}}_{\mathbf{yy}}(m,k) - \hat{\mathbf{\Phi}}_{\mathbf{vv}}(m,k)$. Considering estimation errors in the noisy and noise correlation matrices, negative eigenvalues of $\hat{\mathbf{\Phi}}_{\mathbf{xx}}(m,k)$ are set to zero to ensure that the resulting speech correlation matrix is positive definite.

Considering that the clean speech signal $X_1(m,k)$ is the desired signal at the current TFU, in [2] it was proposed to decompose the vector $\mathbf{x}(m,k)$ into a correlated and an uncorrelated component with regard to $X_1(m,k)$, i.e.,

$$\mathbf{x}(m,k) = \boldsymbol{\rho}_{\mathbf{x}}(m,k)X_1(m,k) + \mathbf{x}'(m,k), \tag{9}$$

where the speech IFC vector is defined as

$$\boldsymbol{\rho}_{\mathbf{x}}(m,k) = \frac{\mathbb{E}\left\{\mathbf{x}(m,k)X_1^*(m,k)\right\}}{\mathbb{E}\left\{|X_1(m,k)|^2\right\}} = \frac{\mathbf{\Phi}_{\mathbf{xx}}(m,k)\mathbf{e}}{\mathbf{e}^T\mathbf{\Phi}_{\mathbf{xx}}(m,k)\mathbf{e}}, \tag{10}$$

and $\mathbb{E}\left\{|X_1(m,k)|^2\right\} = \phi_{X_1}(m,k)$ is the variance of the clean speech signal at the first microphone and $\mathbf{e}$ denotes a selection vector with the first element equal to 1 and all other elements equal to 0. Using this decomposition, (4) can be re-written as

$$\mathbf{y}(m,k) = \boldsymbol{\rho}_{\mathbf{x}}(m,k)X_1(m,k) + \mathbf{x}'(m,k) + \mathbf{v}(m,k). \tag{11}$$

Since $\mathbf{x}'(m,k)$ is uncorrelated with $X_1(m,k)$, it can be interpreted as an interference. By defining the undesired signal as the sum of interference and additive noise, i.e., $\mathbf{n}(m,k) = \mathbf{x}'(m,k) + \mathbf{v}(m,k)$, (6) can be re-written as

$$\mathbf{\Phi}_{\mathbf{yy}}(m,k) = \phi_{X_1}(m,k)\boldsymbol{\rho}_{\mathbf{x}}(m,k)\boldsymbol{\rho}_{\mathbf{x}}^H(m,k) + \mathbf{\Phi}_{\mathbf{nn}}(m,k), \tag{12}$$

where $\mathbf{\Phi}_{\mathbf{nn}}(m,k) = \mathbf{\Phi}_{\mathbf{x'x'}}(m,k) + \mathbf{\Phi}_{\mathbf{vv}}(m,k)$ denotes the correlation matrix of the undesired signal.

III. Multi-Channel Multi-Frame MVDR Filter

The multi-channel multi-frame MVDR (MCMF-MVDR) filter exploiting the IFC was proposed in [1]. This filter aims at minimizing the power spectral density of undesired signal while not distorting the desired signal. This constrained optimization problem can be expressed as

$$\begin{aligned}&\min_{\mathbf{h}(m,k)} \mathbf{h}^H(m,k)\mathbf{\Phi}_{\mathbf{nn}}(m,k)\mathbf{h}(m,k)\\ &\text{s.t.}\quad \mathbf{h}^H(m,k)\boldsymbol{\rho}_{\mathbf{x}}(m,k) = 1.\end{aligned} \tag{13}$$

The solution of this constrained optimization problem is [1]

$$\mathbf{h}_{MCMF-MVDR}(m,k) = \frac{\mathbf{\Phi}_{\mathbf{nn}}^{-1}(m,k)\boldsymbol{\rho}_{\mathbf{x}}(m,k)}{\boldsymbol{\rho}_{\mathbf{x}}^H(m,k)\mathbf{\Phi}_{\mathbf{nn}}^{-1}(m,k)\boldsymbol{\rho}_{\mathbf{x}}(m,k)}. \tag{14}$$

By applying the matrix inverse lemma to (12) this filter can be written as [1]

$$\mathbf{h}_{MCMF-MVDR}(m,k) = \frac{\mathbf{\Phi}_{\mathbf{yy}}^{-1}(m,k)\boldsymbol{\rho}_{\mathbf{x}}(m,k)}{\boldsymbol{\rho}_{\mathbf{x}}^H(m,k)\mathbf{\Phi}_{\mathbf{yy}}^{-1}(m,k)\boldsymbol{\rho}_{\mathbf{x}}(m,k)}. \tag{15}$$

[1] Although this actually corresponds to the minimum power distortionless response (MPDR) filter [9], we decided to keep the original terminology from [1].

ISBN 978-90-827970-1-5

The MCMF-MVDR filter uses the noisy observations at the current and $L-1$ previous frames to estimate the clean speech signal. Especially for a large number of frames and/or microphones, this may however lead to a large computational complexity, as $NL \times NL$-dimensional correlation matrices need to be estimated and inverted (which may in addition lead to numerical problems).

IV. Semi-Distributed Multi-Channel Multi-Frame MVDR Filter

Motivated by distributed processing approaches employed in WASNs, in this section we propose a distributed version of the MCMF-MVDR filter in (15) to reduce its computational complexity.

It should be realized that there are some important differences between WASNs and our considered problem. First, the IFC vectors are more time-varying in comparison with WASNs. In addition, in WASNs each node can share its information with all other nodes, whereas in our application, considering the causality constraint, only the information from the previous frames can be shared (no future frames).

In order to reduce computational complexity, instead of using the NL-dimensional vector in (3) the proposed semi-distributed MCMF-MVDR (SDMCMF-MVDR) filter employs the noisy observation at the current frame and a compressed signal from the $L-1$ previous frames, i.e.,

$$\begin{aligned}&\mathbf{y}_{sd}(m,k) = \\ &[\overline{\mathbf{y}}^T(m,k),\ ...,\ Z(m-L+1,k)]^T \in \mathbb{C}^{(N+L-1)\times 1},\end{aligned} \tag{16}$$

where $Z(m-l,k)$ denotes the compressed signal delayed with l frames. The computation of the compressed signal $Z(m,k)$ can be explained as follows. The vectors $\mathbf{x}_{sd}(m,k)$ and $\mathbf{v}_{sd}(m,k)$ are defined similar to $\mathbf{y}_{sd}(m,k)$. Similarly as in (5), the clean speech signal is estimated as

$$\hat{X}(m,k) = \mathbf{h}_{sd}^H(m,k)\mathbf{y}_{sd}(m,k), \tag{17}$$

with

$$\mathbf{h}_{sd}(m,k) = [\mathbf{h}_{sd_0}^T(m,k),\ H_{sd_1}(m,k),\ ...,\ H_{sd_{L-1}}(m,k)]^T, \tag{18}$$

the $N+L-1$ filter coefficients. Using (16) and (18), (17) can be written as

$$\hat{X}(m,k) = \underbrace{\mathbf{h}_{sd_0}^H(m,k)\overline{\mathbf{y}}(m,k)}_{Z(m,k)} + \sum_{i=1}^{L-1} H_{sd_i}^*(m,k)Z(m-i,k). \tag{19}$$

Motivated by [6], the compressed signal is defined as the noisy observation at the current frame $\overline{\mathbf{y}}(m,k)$ filtered with $\mathbf{h}_{sd_0}(m,k)$, i.e. the first N elements of $\mathbf{h}_{sd}(m,k)$. The compressed signals are computed as a part of the enhanced signal. (In [6], it was shown that each node is able to converge to the centralized solution of fully connected WASN by broadcasting a compressed signal, which is defined as the filtered version of the recorded signals of that node.) It is evident that using the compressed signal from previous frames decreases the vector dimension from NL to $N+L-1$, considerably reducing computational complexity.

Similarly to (15), the filter coefficients of the SDMCMF-MVDR filter can be computed as

$$\mathbf{h}_{sd}(m,k) = \frac{\boldsymbol{\Phi}_{\mathbf{y}_{sd}\mathbf{y}_{sd}}^{-1}(m,k)\boldsymbol{\rho}_{\mathbf{x_{sd}}}(m,k)}{\boldsymbol{\rho}_{\mathbf{x_{sd}}}^H(m,k)\boldsymbol{\Phi}_{\mathbf{y}_{sd}\mathbf{y}_{sd}}^{-1}(m,k)\boldsymbol{\rho}_{\mathbf{x_{sd}}}(m,k)}, \tag{20}$$

where $\boldsymbol{\rho}_{\mathbf{x_{sd}}}(m,k)$ denotes the semi-distributed speech IFC vector, and the semi-distributed noisy correlation matrix is given by

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{y}_{sd}\mathbf{y}_{sd}}(m,k) &= \mathbb{E}\left\{\mathbf{y}_{sd}(m,k)\mathbf{y}_{sd}^H(m,k)\right\} \\ &= \boldsymbol{\Phi}_{\mathbf{x}_{sd}\mathbf{x}_{sd}}(m,k) + \boldsymbol{\Phi}_{\mathbf{v}_{sd}\mathbf{v}_{sd}}(m,k),\end{aligned} \tag{21}$$

where $\boldsymbol{\Phi}_{\mathbf{x}_{sd}\mathbf{x}_{sd}}(m,k)$ and $\boldsymbol{\Phi}_{\mathbf{v}_{sd}\mathbf{v}_{sd}}(m,k)$ denote the semi-distributed clean speech and noise correlation matrices, respectively. Similarly to (7) and (8) the correlation matrix $\boldsymbol{\Phi}_{\mathbf{y}_{sd}\mathbf{y}_{sd}}(m,k)$ and $\boldsymbol{\Phi}_{\mathbf{v}_{sd}\mathbf{v}_{sd}}(m,k)$ can be estimated as

$$\begin{aligned}&\hat{\boldsymbol{\Phi}}_{\mathbf{y}_{sd}\mathbf{y}_{sd}}(m,k) = \\ &\lambda_y\hat{\boldsymbol{\Phi}}_{\mathbf{y}_{sd}\mathbf{y}_{sd}}(m-1,k) + (1-\lambda_y)\mathbf{y}_{sd}(m,k)\mathbf{y}_{sd}^H(m,k),\end{aligned} \tag{22}$$

$$\begin{aligned}&\hat{\boldsymbol{\Phi}}_{\mathbf{v}_{sd}\mathbf{v}_{sd}}(m,k) = \\ &\lambda_v\hat{\boldsymbol{\Phi}}_{\mathbf{v}_{sd}\mathbf{v}_{sd}}(m-1,k) + (1-\lambda_v)\mathbf{v}_{sd}(m,k)\mathbf{v}_{sd}^H(m,k).\end{aligned} \tag{23}$$

The semi-distributed clean speech correlation matrix is given by

$$\hat{\boldsymbol{\Phi}}_{\mathbf{x}_{sd}\mathbf{x}_{sd}}(m,k) = \hat{\boldsymbol{\Phi}}_{\mathbf{y}_{sd}\mathbf{y}_{sd}}(m,k) - \hat{\boldsymbol{\Phi}}_{\mathbf{v}_{sd}\mathbf{v}_{sd}}(m,k), \tag{24}$$

to ensure that the resulting correlation matrix is positive definite, all negative eigenvalues are set to zero.

The semi-distributed speech IFC vector is defined as

$$\boldsymbol{\rho}_{\mathbf{x_{sd}}}(m,k) = \frac{\mathbb{E}\left\{\mathbf{x}_{sd}X_1^*(m,k)\right\}}{\mathbb{E}\left\{|X_1(m,k)|^2\right\}} = \frac{\boldsymbol{\Phi}_{\mathbf{x}_{sd}\mathbf{x}_{sd}}(m,k)\mathbf{e}_{sd}}{\mathbf{e}_{sd}^T\boldsymbol{\Phi}_{\mathbf{x}_{sd}\mathbf{x}_{sd}}(m,k)\mathbf{e}_{sd}}, \tag{25}$$

and $\mathbf{e}_{sd}$ denotes a selection vector with the first element equal to 1 and all other elements equal to 0.

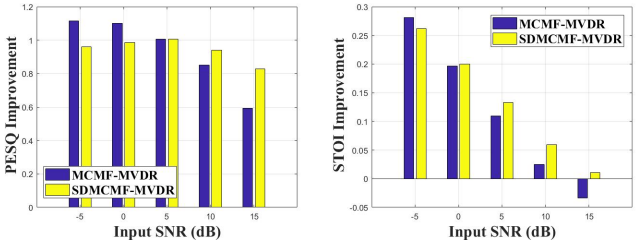
Due to the fact that for the processing of the current frame only compressed signals from the previous frames are used (which is different to distributed processing schemes in WASNs) we don't expect that the proposed SDMCMF-MVDR filter will obtain the same performance as the MCMF-MVDR filter.

V. Simulation Results

In this section, we compare the performance of the proposed SDMCMF-MVDR filter using the compressed signal from previous frames with the traditional MCMF-MVDR filter using the noisy vectors from previous frames.

For the simulations we consider a rectangular room with dimensions 4.5 m $\times$ 4.5 m $\times$ 3 m (width$\times$length$\times$height) with reverberation time $T60 = 200$ ms. We consider a uniform linear array with $N = 4$ microphones at positions $(x_n = x_{init} + (n-1)d,\ y = 1$ m, $z = 2$ m $)$, $n = 1, \cdots, 4$ with $x_{init} = 1$ m and inter-microphone distance $d = 0.02$ m. The speech source signal is located at $(x = 2$ m $; y = 1.5$ m; $z = 1.8$ m). The image method is used to generate the

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) (b)

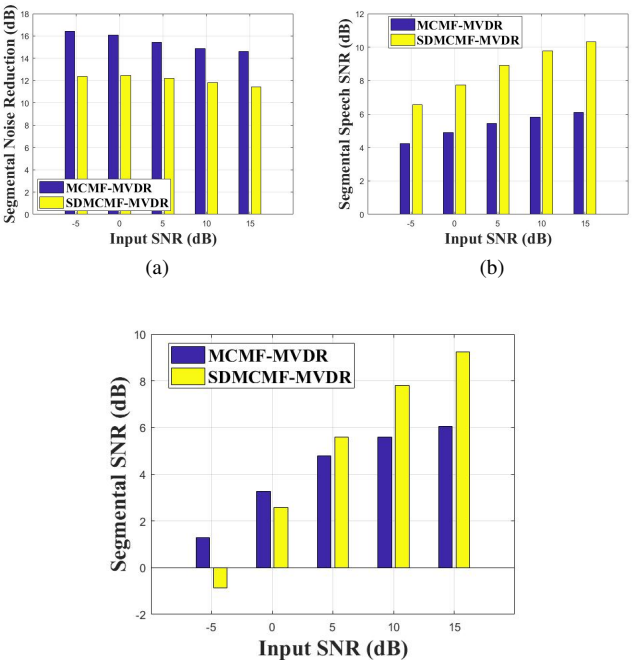
Fig. 1. Performance of the proposed SDMCMF-MVDR and the MCMF-MVDR filters in terms of PESQ and STOI improvement compared to the noisy signal for several SNRs (white Gaussian noise)

room impulse responses between the speech source and the microphones [10]. As clean speech signals we use samples from four male and four female speakers from the TIMIT database [11] and report average results. The microphone signals are corrupted by additive noise (either white Gaussian noise or factory noise) at SNRs ranging from -5 dB to 15 dB. In the case of factory noise, the noise source signal is located at $(x = 0.5$ m $; y = 3.5$ m; $z = 1.3$ m$)$. The sampling frequency is 16 kHz and the STFT is implemented using NFFT$= 256$ with 75% overlap and a Hamming window.

For the multi-frame MVDR filters we chose $L = 5$ to achieve a good compromise between performance and computational complexity. For this choice of parameters the length of the vector $\mathbf{y}(m,k)$ for the MCMF-MVDR filter is equal to $NL = 20$, while the length of the vector $\mathbf{y}_{sd}(m,k)$ for the SDMCMF-MVDR filter is equal to $N + L - 1 = 8$. The forgetting factor to update the noisy correlation matrices based on (7) and (22) is $\lambda_y = 0.92$. Also, similarly as in [2], to avoid the effect of errors in the estimation of noise correlation matrix, we compute the noise correlation matrices $\mathbf{\Phi}_{\mathbf{vv}}(m,k)$ and $\mathbf{\Phi}_{\mathbf{v}_{sd}\mathbf{v}_{sd}}(m,k)$ with the knowledge of noise signals based on the (8) and (23) using $\lambda_v = \lambda_y = 0.92$.

The sensitivity of the multi-frame MVDR filter to estimation errors in the speech and noise correlation matrix was investigated in [9] for single-microphone noise reduction. The main goal of this paper is to investigate how much the performance is affected when using the compressed signal in the proposed SDMCMF-MVDR filter instead of the noisy vectors. Analyzing the sensitivity of the proposed SDMCMF-MVDR filter to estimation errors of noise correlation matrix can be considered as a topic of future research.

For white Gaussian noise Fig. 1 depicts the performance of the proposed SDMCMF-MVDR and the MCMF-MVDR filters in terms of PESQ [12] and short-time objective intelligibility (STOI) [13] improvement compared to the noisy signal at the first microphone (clean speech signal at the first microphone is considered as reference). It can be observed from Fig. 1 (a) that MCMF-MVDR outperforms SDMCMF-MVDR in terms of PESQ for low SNRs. On the other hand, MCMF-MVDR needs to compute the inverse of a 20×20-dimensional correlation matrix, while SDMCMF-MVDR needs to compute the inverse of an 8×8-dimensional correlation matrix, hence considerably reducing the computational complexity.

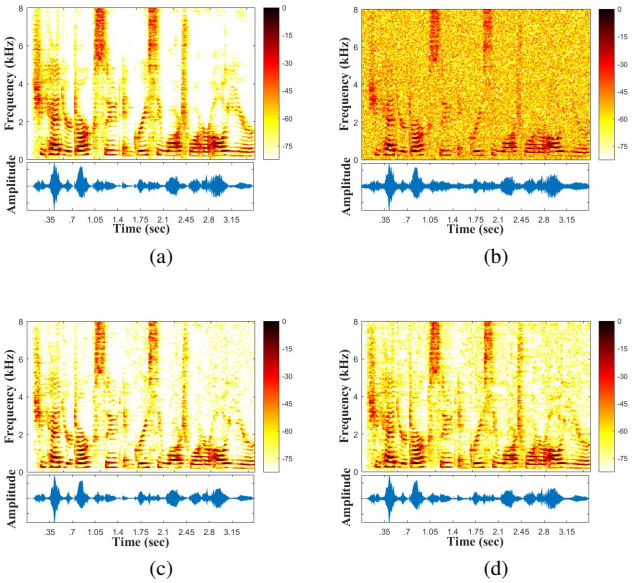


(a) (b) (c)

Fig. 2. Performance of the proposed SDMCMF-MVDR and the MCMF-MVDR filters in terms of SegNR, SSPSNR and SSNR for several SNRs (white Gaussian noise)

In terms of STOI, we observe a similar trend, where MCMF-MVDR achieves a larger STOI improvement than SDMCMF-MVDR for low SNRs, whereas SDMCMF-MVDR achieves a slightly larger STOI improvement than MCMF-MVDR at high SNRs.

Fig. 2 depicts the performance of the SDMCMF-MVDR and MCMF-MVDR filters in terms of the speech segmental SNR



(a) (b) (c) (d)

Fig. 3. Spectrograms of (a) clean signal, (b) noisy signal, (c) enhanced signal using MCMF-MVDR filter, and (d) the enhanced signal using proposed SDMCMF-MVDR filter (SNR = 10 dB, white Gaussian noise).

ISBN 978-90-827970-1-5

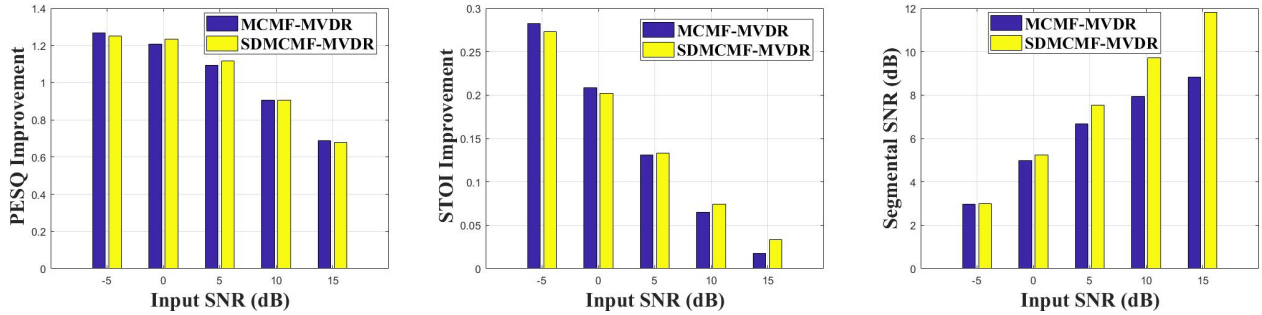


Fig. 4. Performance of the proposed SDMCMF-MVDR and MCMF-MVDR in terms of PESQ, STOI and SSNR in the case of factory noise

(SPSSNR), segmental noise reduction (SegNR) and segmental SNR (SSNR) as proposed in [14]. The SPSSNR has been defined as a measure for speech distortion; the larger SPSSNR the lower the speech distortion. From Fig. 2, the usual trade-off between noise reduction and speech distortion is evident. For all considered SNRs it can be observed that the MCMF-MVDR achieves better results in terms of noise reduction while the SDMCMF-MVDR achieves better results in terms of speech distortion. In terms of SSNR which considers both noise reduction as well as speech distortion, it can be observed that MCMF-MVDR achieves better results for low SNRs while SDMCMF-MVDR achieves better results for high SNRs.

Fig. 3 illustrates the spectrograms of the clean speech signal at the first microphone, the noisy signal (SNR $=$ 10 dB), the enhanced signal using the MCMF-MVDR filter, and the enhanced signal using the proposed SDMCMF-MVDR filter in case of white Gaussian noise. These spectrograms are highly consistent with the evaluation results shown in Fig. 2. The MCMF-MVDR filter considerably reduces the noise however, it removes some speech components which results in speech distortion. In the case of SDMCMF-MVDR, the noise has been significantly reduced, while the speech spectrum is not substantially degraded.

Fig. 4 depicts the performance in terms of PESQ improvement, STOI improvement and segmental SNR when using factory noise instead of white Gaussian noise. It can be observed that the proposed reduced-complexity SDMCMF-MVDR filter yields almost the same (sometimes even better) performance as the MCMF-MVDR filter.

VI. Conclusion

In this paper, the inter-frame correlation between speech components in the STFT domain was exploited to develop a semi-distributed MCMF-MVDR filter which substantially decreases the computational complexity. In each time-frequency unit a vector consisting of the observations at the current frame and compressed signals from previous frames was considered to derive the SDMCMF-MVDR filter. The compressed signals were computed as a part of the enhanced signal at previous frames. We compared the performance of the proposed filter with the traditional MCMF-MVDR filter for white Gaussian and factory noises, showing that the proposed filter yields a similar performance while substantially reducing computational complexity.

References

[1] J. Benesty, J. Chen, and E. A. Habets, *Speech Enhancement in the STFT Domain*, 1st ed. Springer Publishing Company, Incorporated, 2011.

[2] Y. A. 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.

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

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

[5] S. Doclo, M. Moonen, T. van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 1, pp. 38–51, Jan 2009.

[6] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks; part I: Sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, Oct 2010.

[7] S. Markovich-Golan, S. Gannot, and I. Cohen, "Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 2, pp. 343–356, Feb 2013.

[8] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Process.*, vol. 107, pp. 4 – 20, 2015.

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

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Acoustical Society of America Journal*, vol. 65, pp. 943–950, Apr. 1979.

[11] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburgh, MD, Tech. Rep., Dec. 1988, (prototype as of).

[12] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC press, 2007.

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, March 2010, pp. 4214–4217.

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

ISBN 978-90-827970-1-5 © EURASIP 2018