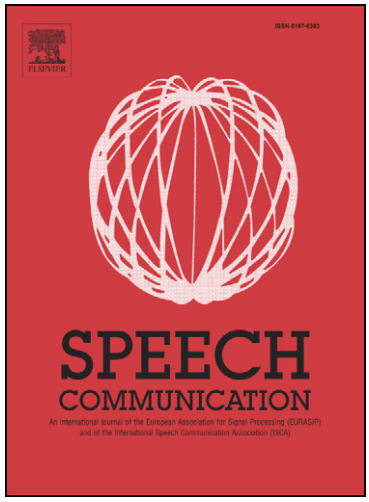
Frequency-Domain Criterion for the Speech Distortion Weighted Multichannel Wiener Filter for Robust Noise Reduction

Simon Doclo, Ann Spriet, Jan Wouters, Marc Moonen



Frequency-Domain Criterion for the Speech Distortion Weighted Multichannel Wiener Filter for Robust Noise Reduction

Simon Doclo, Ann Spriet, Jan Wouters, Marc Moonen

PII: S0167-6393(07)00031-3
DOI: 10.1016/j.specom.2007.02.001
Reference: SPECOM 1620

To appear in: *Speech Communication*

Received Date: 1 February 2006
Revised Date: 22 October 2006
Accepted Date: 4 February 2007

Please cite this article as: Doclo, S., Spriet, A., Wouters, J., Moonen, M., Frequency-Domain Criterion for the Speech Distortion Weighted Multichannel Wiener Filter for Robust Noise Reduction , *Speech Communication* (2007), doi: 10.1016/j.specom.2007.02.001

Frequency-Domain Criterion for the Speech Distortion Weighted Multichannel Wiener Filter for Robust Noise Reduction

Simon Doclo *, Ann Spriet, Jan Wouters, Marc Moonen

Katholieke Universiteit Leuven, Dept. of Electrical Engineering (ESAT - SCD), Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium

Abstract

Recently, a generalized multi-microphone noise reduction scheme, referred to as the spatially pre-processed speech distortion weighted multichannel Wiener filter (SP-SDW-MWF), has been presented. This scheme consists of a fixed spatial pre-processor and a multichannel adaptive noise canceler (ANC) optimizing the SDW-MWF cost function. By taking speech distortion explicitly into account in the design criterion of the multichannel ANC, the SP-SDW-MWF adds robustness to the standard generalized sidelobe canceler (GSC). In this paper, we present a multichannel frequency-domain criterion for the SDW-MWF, from which several – existing and novel – adaptive frequency-domain algorithms can be derived. The main difference between these adaptive algorithms consists in the calculation of the step size matrix (constrained vs. unconstrained, block-structured vs. diagonal) used in the update formula for the multichannel adaptive filter. We investigate the noise reduction performance, the robustness and the tracking performance of these adaptive algorithms, using a perfect voice activity detection (VAD) mechanism and using an energy-based VAD. Using experimental results with a small-sized microphone array in a hearing aid, it is shown that the SP-SDW-MWF is more robust against signal model errors than the GSC, and that the block-structured step size matrix gives rise to a faster convergence and a better tracking performance than the diagonal step size matrix, only at a slightly higher computational cost.

Key words: multi-microphone noise reduction, adaptive frequency-domain algorithms, multichannel Wiener filter, generalized sidelobe canceler, hearing aids
PACS: 43.60.Fg, 43.60.Mn, 43.72.-p, 43.60.Dh

* Corresponding author. Tel: +32/16/321899, Fax: +32/16/321970
Email address: simon.doclo@esat.kuleuven.be (Simon Doclo).

1 Introduction

In many speech communication applications, such as hands-free mobile telephony, hearing aids and voice-controlled systems, the recorded speech signals are corrupted by acoustic background noise. Generally speaking, background noise is broadband and non-stationary, and the signal-to-noise ratio (SNR) may be quite low. Background noise causes a signal degradation that can lead to total unintelligibility of the speech signal and that substantially decreases the performance of speech coding and speech recognition systems. Therefore efficient speech enhancement techniques are called for.

Since the desired speech signal and the undesired noise signal usually occupy overlapping frequency bands, single-microphone speech enhancement techniques, such as spectral subtraction, Kalman filtering, and signal subspace-based techniques, often fail to reduce the background noise without introducing artifacts (e.g. musical noise) or speech distortion. However, when the speech and noise sources are physically located at different positions, it is possible to exploit this spatial diversity by using a microphone array, such that both the spectral and the spatial characteristics of the sources can be used.

Well-known multi-microphone speech enhancement techniques are fixed and adaptive beamforming [1]. In a minimum variance distortionless response (MVDR) beamformer [2], the energy of the output signal is minimized under the constraint that signals arriving from the look direction, i.e. the assumed direction of the speech source, are processed without distortion. A widely studied adaptive implementation of this beamformer is the generalized sidelobe canceler (GSC) [3], which consists of a fixed spatial pre-processor, i.e. a fixed beamformer and a blocking matrix, combined with a multichannel adaptive noise canceler (ANC). The fixed beamformer creates a so-called speech reference, the blocking matrix creates so-called noise references, and the multichannel ANC eliminates the noise components in the speech reference that are correlated with the noise references.

Due to room reverberation, microphone mismatch, look direction error and spatially distributed sources, speech components may however leak into the noise references of the standard GSC, giving rise to speech distortion and possibly signal cancelation. Several techniques have been proposed to limit the speech distortion resulting from this speech leakage, e.g.

- *reducing the speech leakage components in the noise references*, e.g. using a more robust fixed blocking matrix design [4–7]; using an adaptive blocking matrix [8–10]; or by constructing a blocking matrix based on estimating the ratios of the acoustic transfer functions from the speech source to the microphone array [11];
- *limiting the distorting effect of the remaining speech leakage components* by

- updating the multichannel ANC only during periods (and for frequencies) where the noise component is dominant, i.e. where the SNR is low [4,8–15]; and
- constraining the update formula for the multichannel adaptive filter, e.g. by imposing a quadratic inequality constraint (QIC) [9,16–18]; by using the leaky least mean square (LMS) algorithm [5,6]; or by taking speech distortion due to speech leakage into account using the so-called speech distortion weighted multichannel Wiener filter (SDW-MWF) [19–21].

In this paper, we will focus on implementation aspects of the SDW-MWF. In [22–24], recursive matrix-decomposition-based implementations for the SDW-MWF have been presented, which are computationally quite expensive. In [20] cheaper (time-domain and frequency-domain) stochastic gradient algorithms have been proposed. These algorithms however require large circular data buffers, resulting in a large memory requirement. In [21], adaptive frequency-domain algorithms for the SDW-MWF have been presented using frequency-domain correlation matrices, reducing the memory requirement and the computational complexity.

Recently, a generalized multichannel frequency-domain filtering framework has been proposed, which takes into account both the autocorrelation of the individual channels as well as the cross-correlation between the different channels [25,26]. Using this framework, several adaptive algorithms can be derived, which have been applied to e.g. multichannel acoustic echo cancelation and the GSC. In this paper, we will use this framework to formulate a *frequency-domain criterion* for the SDW-MWF, trading off noise reduction and speech distortion. From the proposed criterion several *adaptive frequency-domain algorithms* for the SDW-MWF can be derived. The main difference between these algorithms consists in the calculation of the step size matrix in the update formula for the multichannel adaptive filter and in the calculation of a particular regularization term (cf. Sections 3 and 4).

The paper is organized as follows. In Section 2, the GSC and the spatially pre-processed SDW-MWF are briefly reviewed. In Section 3, the frequency-domain criterion for the SDW-MWF is presented. A recursive (RLS-type) algorithm is derived from this criterion and it is shown how this algorithm can be implemented in practice. In Section 4, several approximations are proposed for reducing the computational complexity, leading to adaptive (LMS-type) frequency-domain algorithms, some of which have already been presented in the literature [21]. Section 5 discusses the computational complexity of the different adaptive algorithms. In Section 6, the noise reduction performance, the robustness against signal model errors, and the tracking performance of the proposed algorithms are illustrated using experimental results for a small-sized microphone array in a hearing aid. In addition, the impact of using a non-perfect VAD on the performance is analyzed.

2 GSC and Spatially Pre-Processed SDW-MWF

2.1 Notation and General Structure

Consider a microphone array with M microphones, where each microphone signal $u_i[k]$, $i = 1 \dots M$, at time k, consists of a filtered version of the clean speech signal $s[k]$ and additive noise, i.e.

$$u_i[k] = h_i[k] * s[k] + u_i^v[k],\ i = 1 \dots M, \tag{1}$$

where $h_i[k]$ represents the acoustic impulse response between the speech source and the ith microphone and $*$ denotes convolution. The additive noise $u_i^v[k]$ can be colored and is assumed to be uncorrelated with the clean speech signal.

The spatially pre-processed speech distortion weighted multichannel Wiener Filter (SP-SDW-MWF) [19] is depicted in Figure 1. It consists of a fixed spatial pre-processor, i.e. a fixed beamformer and a blocking matrix, and a multichannel ANC. Note that the structure of the SP-SDW-MWF strongly resembles the standard GSC, but the difference lies in the fact that the SDW-MWF cost function is used in the multichannel ANC and that it is possible to include an extra filter $\mathbf{w}_0$ on the speech reference.

The *fixed beamformer* creates a so-called speech reference

$$y_0[k] = x_0[k] + v_0[k], \tag{2}$$

with $x_0[k]$ and $v_0[k]$ respectively the speech and the noise component of the speech reference, by steering a beam towards the assumed direction of the speaker. The fixed beamformer should be designed such that the distortion of the speech component $x_0[k]$, due to possible errors in the assumed signal model (e.g. look direction error, microphone mismatch) is small. A delay-and-sum beamformer, which time-aligns the microphone signals, offers sufficient robustness against signal model errors since it minimizes the noise sensitivity. However, in order to achieve a better spatial selectivity while still preserving robustness, the fixed beamformer can be optimized, e.g. by using statistical knowledge about the signal model errors that occur in practice [7].

The *blocking matrix* creates $M - 1$ so-called noise references

$$y_n[k] = x_n[k] + v_n[k], \quad n = 1 \dots M - 1, \tag{3}$$

by steering zeroes towards the assumed direction of the speaker. A simple technique to create the noise references consists of pair-wisely subtracting the time-aligned microphone signals. Under ideal conditions (i.e. no reverberation, point speech source, no look direction error, no microphone mismatch), the noise references only contain noise components $v_n[k]$. Since these conditions are never fulfilled in practice, undesired speech components $x_n[k]$, i.e. so-called

speech leakage components, are present in the noise references. Although several techniques have been proposed for reducing the speech leakage components in the noise references [4–11], speech leakage can never be completely avoided in practice.

During *speech periods*, the speech and the noise references consist of speech and noise components, i.e. $y_n[k] = x_n[k] + v_n[k]$, whereas during *noise-only periods* (speech pauses), only the noise components $v_n[k]$ are observed. We assume that the second-order statistics of the noise are sufficiently stationary such that they can be estimated during noise-only periods and used during subsequent speech periods. This requires the use of a voice activity detection (VAD) mechanism [27,28] or an on-line SNR estimation procedure [29].

The goal of the *multichannel ANC* is to estimate the noise component $v_0[k]$ in the speech reference and to subtract this noise estimate from the speech reference in order to obtain an enhanced output signal $z[k]$. Let N be the number of input channels to the multichannel filter ($N = M$ if the filter $\mathbf{w}_0$ on the speech reference is present, $N = M - 1$ otherwise). Let the FIR filters $\mathbf{w}_n[k]$, $n = M-N, \ldots, M-1$, have filter length L, and consider the L-dimensional data vectors $\mathbf{y}_n[k]$, the NL-dimensional stacked data vector $\mathbf{y}[k]$, and the NL-dimensional stacked filter $\mathbf{w}[k]$, defined as

$$\mathbf{y}_n[k] = \Big[\, y_n[k] \; y_n[k-1] \; \ldots y_n[k-L+1] \,\Big]^T, \; n = M-N, \ldots, M-1, \quad (4)$$

$$\mathbf{y}[k] = \Big[\, \mathbf{y}_{M-N}^T[k] \; \mathbf{y}_{M-N+1}^T[k] \; \ldots \; \mathbf{y}_{M-1}^T[k] \,\Big]^T, \quad (5)$$

$$\mathbf{w}[k] = \Big[\, \mathbf{w}_{M-N}^T[k] \; \mathbf{w}_{M-N+1}^T[k] \; \ldots \; \mathbf{w}_{M-1}^T[k] \,\Big]^T, \quad (6)$$

where ^{T} denotes transpose of a vector or a matrix. The stacked data vector can be decomposed into a speech and a noise component, i.e. $\mathbf{y}[k] = \mathbf{x}[k] + \mathbf{v}[k]$, where $\mathbf{x}[k]$ and $\mathbf{v}[k]$ are defined similarly as in (4) and (5). The goal of the filter $\mathbf{w}[k]$ is to estimate the delayed noise component $v_0[k-\Delta]$ in the speech reference[1]. This noise estimate is then subtracted from the speech reference in order to obtain the enhanced output signal $z[k]$, i.e.

$$z[k] = y_0[k-\Delta] - \mathbf{w}^T[k]\mathbf{y}[k] \quad (7)$$

$$= x_0[k-\Delta] + \underbrace{(v_0[k-\Delta] - \mathbf{w}^T[k]\mathbf{v}[k])}_{e_v[k]} - \underbrace{\mathbf{w}^T[k]\mathbf{x}[k]}_{e_x[k]}. \quad (8)$$

Hence, the output signal $z[k]$ consists of 3 terms: the delayed speech component $x_0[k-\Delta]$ in the speech reference, residual noise $e_v[k]$, and (linear) speech

[1] The delay Δ is applied to the speech reference in order to allow for non-causal filter taps. This delay is usually set equal to $\lceil L/2 \rceil$, where $\lceil x \rceil$ denotes the smallest integer larger than or equal to x.

distortion $e_x[k]$. The goal of any speech enhancement algorithm is to reduce the residual noise as much as possible, while simultaneously limiting the speech distortion. The speech distortion can e.g. be limited by reducing the speech leakage components $\mathbf{x}[k]$ and/or by constraining the filter $\mathbf{w}[k]$.

In this paper we will assume a fixed blocking matrix, such that speech leakage components are always present, especially when microphone mismatch occurs. We will not consider techniques here that aim to minimize the speech leakage components by using an adaptive blocking matrix (ABM) [8–11]. One should however realize that these ABM-techniques may be used as an alternative or even in combination with the SDW-MWF.

2.2 Generalized Sidelobe Canceler (GSC)

The standard GSC aims to minimize the residual noise energy $\varepsilon_v^2[k]$ without taking into account speech distortion, i.e.

$$J_{GSC}(\mathbf{w}[k]) = \varepsilon_v^2[k] = E\Big\{|v_0[k-\Delta] - \mathbf{w}^T[k]\mathbf{v}[k]|^2\Big\}, \tag{9}$$

where E denotes the expected value operator. The filter $\mathbf{w}[k]$ minimizing this cost function is equal to

$$\mathbf{w}[k] = E\{\mathbf{v}[k]\mathbf{v}^T[k]\}^{-1}\, E\{\mathbf{v}[k]v_0[k-\Delta]\}, \tag{10}$$

where the noise correlation matrix $E\{\mathbf{v}[k]\mathbf{v}^T[k]\}$ and the noise cross-correlation vector $E\{\mathbf{v}[k]v_0[k-\Delta]\}$ are estimated during noise-only periods. Hence, in a typical adaptive implementation, the filter $\mathbf{w}[k]$ is allowed to be updated only during noise-only periods [4,8–15], since adaptation during speech periods would lead to an incorrect solution and possibly signal cancelation. Note however that signal distortion due to speech leakage still occurs even when the adaptive filter is updated only during noise-only periods, since the speech distortion term $e_x[k]$ is still present in the output signal $z[k]$.

A commonly used approach to increase the robustness against signal model errors is to apply a *quadratic inequality constraint (QIC)* [16–18], i.e.

$$\mathbf{w}^T[k]\mathbf{w}[k] \leq \beta^2. \tag{11}$$

The QIC avoids excessive growth of the filter coefficients $\mathbf{w}[k]$, and hence limits speech distortion $\mathbf{w}^T[k]\mathbf{x}[k]$ due to speech leakage.

In the GSC the number of input channels to the adaptive filter is typically equal to $N = M-1$. It is however not possible to include the filter $\mathbf{w}_0$ on the speech reference, since in this case the filter $\mathbf{w}[k]$ in (10) would be equal to

$$\mathbf{w}_0[k] = \mathbf{u}_{\Delta+1}, \quad \mathbf{w}_n[k] = \mathbf{0}, \;\; n = 1 \ldots M-1\,, \tag{12}$$

with $\mathbf{u}_l$ the lth canonical L-dimensional vector, i.e. a vector of which the lth element is equal to 1 and all other elements are equal to 0, such that the output signal $z[k] = 0$.

2.3 Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF)

The SDW-MWF takes speech distortion due to speech leakage explicitly into account in the design criterion of the filter $\mathbf{w}[k]$ and aims to minimize a weighted sum of the residual noise energy $\varepsilon_v^2[k]$ and the speech distortion energy $\varepsilon_x^2[k]$, i.e.

$$\begin{aligned} J_{SDW-MWF}(\mathbf{w}[k]) &= \varepsilon_v^2[k] + \frac{1}{\mu}\varepsilon_x^2[k] \\ &= E\Big\{|v_0[k-\Delta] - \mathbf{w}^T[k]\mathbf{v}[k]|^2\Big\} + \frac{1}{\mu}E\Big\{|\mathbf{w}^T[k]\mathbf{x}[k]|^2\Big\} \end{aligned} \tag{13}$$

where the parameter $\mu \in [0, \infty]$ provides a trade-off between noise reduction and speech distortion [19,23,30]. If $\mu = 1$, the minimum mean square error (MMSE) criterion is obtained. If $\mu < 1$, speech distortion is reduced at the expense of increased residual noise energy. On the other hand, if $\mu > 1$, residual noise is reduced at the expense of increased speech distortion.

The filter $\mathbf{w}[k]$ minimizing the cost function in (13) is equal to

$$\mathbf{w}[k] = \left[E\{\mathbf{v}[k]\mathbf{v}^T[k]\} + \frac{1}{\mu}E\{\mathbf{x}[k]\mathbf{x}^T[k]\}\right]^{-1} E\{\mathbf{v}[k]v_0[k-\Delta]\}, \tag{14}$$

where, using the independence assumption between speech and noise, the speech correlation matrix $E\{\mathbf{x}[k]\mathbf{x}^T[k]\}$ can be computed as

$$E\{\mathbf{x}[k]\mathbf{x}^T[k]\} = E\{\mathbf{y}[k]\mathbf{y}^T[k]\} - E\{\mathbf{v}[k]\mathbf{v}^T[k]\}. \tag{15}$$

The correlation matrix $E\{\mathbf{y}[k]\mathbf{y}^T[k]\}$ is estimated during speech periods and the noise correlation matrix $E\{\mathbf{v}[k]\mathbf{v}^T[k]\}$ is estimated during noise-only periods. As already mentioned, we assume that the spectral and/or spatial characteristics of the noise are sufficiently stationary.

Since the SDW-MWF takes speech distortion explicitly into account in its optimization criterion, it is now possible to include an extra filter $\mathbf{w}_0$ on the speech reference. Depending on the setting of the parameter μ and the presence/absence of the filter $\mathbf{w}_0$, different algorithms are obtained:

- Without a filter $\mathbf{w}_0$ ($N = M-1$), we obtain the *speech distortion regularized GSC* (SDR-GSC), where the standard optimization criterion of the GSC in (9) is supplemented with a regularization term $1/\mu\,\varepsilon_x^2$. For $\mu = \infty$, speech distortion is completely ignored, which corresponds to the standard GSC.

For $\mu = 0$, all emphasis is put on speech distortion, such that $\mathbf{w}[k] = \mathbf{0}$ and the output signal $z[k]$ is equal to the delayed speech reference $y_0[k - \Delta]$. Compared to the QIC-GSC, the SDR-GSC is less conservative, since the regularization term is proportional to the actual amount of speech leakage in the noise references. In [19] it has been shown that in comparison with the QIC-GSC, the SDR-GSC obtains a better noise reduction for small model errors, while guaranteeing robustness against large model errors.

- With a filter $\mathbf{w}_0$ ($N = M$), we obtain the *spatially pre-processed speech distortion weighted multichannel Wiener filter* (SP-SDW-MWF). For $\mu = 1$, the output signal $z[k]$ is the MMSE estimate of the delayed speech component $x_0[k - \Delta]$ in the speech reference. In [19] it has been shown that, for infinite filter lengths, the performance of the SP-SDW-MWF is not affected by microphone mismatch. Hence, the extra filter on the speech reference further improves the performance.

In [22–24], recursive matrix-decomposition-based implementations have been presented, which are computationally quite expensive. Starting from the cost function in (13), a cheaper time-domain stochastic gradient algorithm has been derived. To speed up convergence and reduce the computational complexity, this algorithm has been implemented in the frequency-domain [20]. It has been shown that for highly non-stationary noise, this stochastic gradient algorithm suffers from a large excess error, which can be reduced by low-pass filtering a particular regularization term, i.e. the part of the gradient estimate that limits speech distortion. The computation of this regularization term however requires the storage of circular data buffers, giving rise to a large memory requirement. In [21], the regularization term has been approximated in the frequency-domain, using (diagonal) speech and noise correlation matrices in the frequency-domain. This approximation leads to a drastic decrease in memory requirement and also further reduces the computational complexity.

In the following section, a novel *frequency-domain criterion* for the SDW-MWF is presented, which is similar to the cost function in (13). This frequency-domain criterion is an extension of the criterion used in [25,26] for multichannel echo cancelation. Furthermore, it provides a way for linking existing adaptive frequency-domain algorithms for the SDW-MWF [21] and for deriving novel adaptive algorithms, as will be shown in Section 4.

3 Frequency-Domain Criterion for the SDW-MWF

We first define block signals for the residual noise and the speech distortion, which can be computed using frequency-domain operations. Using these block signals, we define a frequency-domain cost function for the SDW-MWF. By setting the derivative of this cost function to zero, we obtain the normal equations, from which a recursive (RLS-type) algorithm can be derived. Next, we discuss some practical implementation issues, i.e. adaptation during noise-only

periods and computation of the regularization term. The general block diagram of the frequency-domain implementation of the SDW-MWF is depicted in Figure 2.

3.1 Frequency-Domain Notation

We define the L-dimensional block signals $\mathbf{e}_v[m]$ and $\mathbf{e}_x[m]$ as

$$\mathbf{e}_v[m] = \Big[\, e_v[mL] \; e_v[mL+1] \; \ldots \; e_v[mL+L-1] \,\Big]^T, \tag{16}$$

$$\mathbf{e}_x[m] = \Big[\, e_x[mL] \; e_x[mL+1] \; \ldots \; e_x[mL+L-1] \,\Big]^T, \tag{17}$$

with m the block time index. Using (8), the block signal $\mathbf{e}_v[m]$, representing the residual noise, can be computed using frequency-domain operations as [25,26,31]

$$\mathbf{e}_v[m] = \mathbf{d}[m] - \Big[\, \mathbf{0}_L \; \mathbf{I}_L \,\Big]\, \mathbf{F}_{2L}^{-1} \sum_{n=M-N}^{M-1} \mathbf{D}_{v,n}[m]\, \mathbf{F}_{2L} \begin{bmatrix} \mathbf{I}_L \\ \mathbf{0}_L \end{bmatrix} \mathbf{w}_n \;, \tag{18}$$

with

$$\mathbf{d}[m] = \Big[\, v_0[mL-\Delta] \; v_0[mL-\Delta+1] \; \ldots \; v_0[mL-\Delta+L-1] \,\Big]^T \;. \tag{19}$$

$\mathbf{0}_L$ represents the $L \times L$-dimensional zero matrix, $\mathbf{I}_L$ represents the $L \times L$-dimensional identity matrix, $\mathbf{F}_{2L}$ is the $2L \times 2L$-dimensional discrete Fourier transform matrix and $\mathbf{D}_{v,n}[m]$ is a $2L \times 2L$-dimensional diagonal matrix whose elements are the discrete Fourier transform of the $2L$-dimensional vector

$$\Big[\, v_n[mL-L] \; \ldots \; v_n[mL-1] \; v_n[mL] \; \ldots \; v_n[mL+L-1] \,\Big]^T. \tag{20}$$

The block signal $\mathbf{e}_v[m]$ can also be written as

$$\mathbf{e}_v[m] = \mathbf{d}[m] - \Big[\, \mathbf{0}_L \; \mathbf{I}_L \,\Big]\, \mathbf{F}_{2L}^{-1}\, \mathbf{U}_v[m]\, \mathbf{w}, \tag{21}$$

with the $2L \times NL$-dimensional matrix $\mathbf{U}_v[m]$ defined as

$$\mathbf{U}_v[m] = \left[\mathbf{D}_{v,M-N}[m]\, \mathbf{F}_{2L} \begin{bmatrix} \mathbf{I}_L \\ \mathbf{0}_L \end{bmatrix} \; \ldots \; \mathbf{D}_{v,M-1}[m]\, \mathbf{F}_{2L} \begin{bmatrix} \mathbf{I}_L \\ \mathbf{0}_L \end{bmatrix} \right] \tag{22}$$

$$= \mathbf{D}_v[m]\, \mathbf{F}_{2NL\times NL}^{10}, \tag{23}$$

and the $2L \times 2NL$-dimensional matrix $\mathbf{D}_v[m]$ and the $2NL \times NL$-dimensional block diagonal matrix $\mathbf{F}_{2NL\times NL}^{10}$ equal to

$$\mathbf{D}_v[m] = \Big[\, \mathbf{D}_{v,M-N}[m] \;\ldots\; \mathbf{D}_{v,M-1}[m] \,\Big] \tag{24}$$

$$\mathbf{F}^{10}_{2NL\times NL} = \mathrm{diag}\left[\mathbf{F}_{2L} \begin{bmatrix} \mathbf{I}_L \\ \mathbf{0}_L \end{bmatrix} \;\ldots\; \mathbf{F}_{2L} \begin{bmatrix} \mathbf{I}_L \\ \mathbf{0}_L \end{bmatrix} \right]. \tag{25}$$

Similarly, the block signal $\mathbf{e}_x[m]$, representing the speech distortion, can be computed as

$$\mathbf{e}_x[m] = \Big[\, \mathbf{0}_L \;\; \mathbf{I}_L \,\Big]\, \mathbf{F}^{-1}_{2L}\, \mathbf{U}_x[m]\, \mathbf{w} = \Big[\, \mathbf{0}_L \;\; \mathbf{I}_L \,\Big]\, \mathbf{F}^{-1}_{2L}\, \mathbf{D}_x[m]\, \mathbf{F}^{10}_{2NL\times NL}\, \mathbf{w}, \tag{26}$$

where $\mathbf{U}_x[m]$ and $\mathbf{D}_x[m]$ are defined similarly as $\mathbf{U}_v[m]$ and $\mathbf{D}_v[m]$ for the speech component instead of the noise component.

If we multiply the block signals in (21) and (26) with the $L\times L$-dimensional discrete Fourier transform matrix $\mathbf{F}_L$, we obtain the error signals in the frequency-domain (denoted by underbars), i.e.

$$\underline{\mathbf{e}}_v[m] = \mathbf{F}_L\, \mathbf{e}_v[m] = \underline{\mathbf{d}}[m] - \mathbf{G}^{01}_{L\times 2L} \mathbf{U}_v[m]\, \mathbf{w}, \tag{27}$$

$$\underline{\mathbf{e}}_x[m] = \mathbf{F}_L\, \mathbf{e}_x[m] = \mathbf{G}^{01}_{L\times 2L} \mathbf{U}_x[m]\, \mathbf{w}, \tag{28}$$

with $\underline{\mathbf{d}}[m] = \mathbf{F}_L\, \mathbf{d}[m]$ and $\mathbf{G}^{01}_{L\times 2L} = \mathbf{F}_L \Big[\, \mathbf{0}_L \;\; \mathbf{I}_L \,\Big]\, \mathbf{F}^{-1}_{2L}$.

Using these frequency-domain signals, we now define a *frequency-domain criterion* for the SDW-MWF, minimizing the weighted sum of the residual noise energy and the speech distortion energy, i.e.

$$\boxed{J_f[m] = (1-\lambda_v) \sum_{i=0}^{m} \lambda_v^{m-i}\, \underline{\mathbf{e}}_v^H[i]\, \underline{\mathbf{e}}_v[i] + \frac{1}{\mu}(1-\lambda_x) \sum_{i=0}^{m} \lambda_x^{m-i}\, \underline{\mathbf{e}}_x^H[i]\, \underline{\mathbf{e}}_x[i]} \tag{29}$$

where ^{H} denotes complex conjugate of a vector or a matrix, λ_v and λ_x are exponential forgetting factors respectively for noise and speech ($0 < \lambda_v < 1$, $0 < \lambda_x < 1$), and $1/\mu$ is the trade-off parameter between noise reduction and speech distortion. Note that typically quite large values are used for the exponential forgetting factors (cf. Section 6.2), implying that mainly the long-term spatial and spectral characteristics of the speech and the noise sources are used.

3.2 Normal Equations

The cost function $J_f[m]$ can be minimized by setting its derivative with respect to the (time-domain) filter coefficients $\mathbf{w}[m]$ equal to zero. Using (27) and (28), the derivative is equal to

$$\frac{\partial J_f[m]}{\partial \mathbf{w}[m]} = (1-\lambda_v)\sum_{i=0}^{m}\lambda_v^{m-i}\Big(\mathbf{U}_v^H[i]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_v[i]\mathbf{w}[m] - \mathbf{U}_v^H[i]\,\underline{\mathbf{d}}_{2L}[i]\Big) + \frac{1}{\mu}(1-\lambda_x)\sum_{i=0}^{m}\lambda_x^{m-i}\,\mathbf{U}_x^H[i]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_x[i]\mathbf{w}[m], \tag{30}$$

with

$$\underline{\mathbf{d}}_{2L}[m] = 2(\mathbf{G}_{L\times 2L}^{01})^H\underline{\mathbf{d}}[m] = \mathbf{F}_{2L}\begin{bmatrix}\mathbf{0}_L\\ \mathbf{I}_L\end{bmatrix}\mathbf{d}[m] \tag{31}$$

$$\mathbf{G}_{2L\times 2L}^{01} = 2(\mathbf{G}_{L\times 2L}^{01})^H\mathbf{G}_{L\times 2L}^{01}, = \mathbf{F}_{2L}\begin{bmatrix}\mathbf{0}_L & \mathbf{0}_L\\ \mathbf{0}_L & \mathbf{I}_L\end{bmatrix}\mathbf{F}_{2L}^{-1}. \tag{32}$$

Hence, the *normal equations* can be written as

$$\boxed{\left[\mathbf{S}_v[m] + \frac{1}{\mu}\mathbf{S}_x[m]\right]\mathbf{w}[m] = \mathbf{s}[m]} \tag{33}$$

with the $NL\times NL$-dimensional correlation matrices $\mathbf{S}_v[m]$ and $\mathbf{S}_x[m]$, and the NL-dimensional cross-correlation vector $\mathbf{s}[m]$ defined as

$$\mathbf{S}_v[m] = (1-\lambda_v)\sum_{i=0}^{m}\lambda_v^{m-i}\,\mathbf{U}_v^H[i]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_v[i] \tag{34}$$

$$= \lambda_v\mathbf{S}_v[m-1] + (1-\lambda_v)\,\mathbf{U}_v^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_v[m], \tag{35}$$

$$\mathbf{S}_x[m] = (1-\lambda_x)\sum_{i=0}^{m}\lambda_x^{m-i}\,\mathbf{U}_x^H[i]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_x[i] \tag{36}$$

$$= \lambda_x\mathbf{S}_x[m-1] + (1-\lambda_x)\,\mathbf{U}_x^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_x[m], \tag{37}$$

$$\mathbf{s}[m] = (1-\lambda_v)\sum_{i=0}^{m}\lambda_v^{m-i}\,\mathbf{U}_v^H[i]\,\underline{\mathbf{d}}_{2L}[i] \tag{38}$$

$$= \lambda_v\mathbf{s}[m-1] + (1-\lambda_v)\,\mathbf{U}_v^H[m]\,\underline{\mathbf{d}}_{2L}[m]. \tag{39}$$

3.3 Recursive Algorithm

A recursive (RLS-type) algorithm for updating $\mathbf{w}[m]$ can be found by enforcing the normal equations (33) at block time m and $m-1$, i.e.

$$\left[\mathbf{S}_v[m] + \frac{1}{\mu}\mathbf{S}_x[m]\right]\mathbf{w}[m] = \lambda_v\mathbf{s}[m-1] + (1-\lambda_v)\,\mathbf{U}_v^H[m]\,\underline{\mathbf{d}}_{2L}[m]$$
$$= \lambda_v\left[\mathbf{S}_v[m-1] + \frac{1}{\mu}\mathbf{S}_x[m-1]\right]\mathbf{w}[m-1] + (1-\lambda_v)\,\mathbf{U}_v^H[m]\,\underline{\mathbf{d}}_{2L}[m]$$

$$= \Big[\mathbf{S}_v[m] - (1-\lambda_v)\,\mathbf{U}_v^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_v[m] + \frac{1}{\mu}\frac{\lambda_v}{\lambda_x}\Big(\mathbf{S}_x[m] - (1-\lambda_x)\,\mathbf{U}_x^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_x[m]\Big)\Big]\cdot \mathbf{w}[m-1] + (1-\lambda_v)\,\mathbf{U}_v^H[m]\,\underline{\mathbf{d}}_{2L}[m],$$

such that the recursive update formula for $\mathbf{w}[m]$ can be written as

$$\mathbf{w}[m] = \left[\mathbf{S}_v[m] + \frac{1}{\mu}\mathbf{S}_x[m]\right]^{-1}\left\{\left[\mathbf{S}_v[m] + \frac{1}{\mu}\frac{\lambda_v}{\lambda_x}\mathbf{S}_x[m]\right]\mathbf{w}[m-1] + (1-\lambda_v)\mathbf{U}_v^H[m]\underline{\mathbf{e}}_{v,2L}[m] - \frac{1}{\mu}\frac{\lambda_v}{\lambda_x}(1-\lambda_x)\mathbf{U}_x^H[m]\underline{\mathbf{e}}_{x,2L}[m]\right\}, \quad (40)$$

$$\underline{\mathbf{e}}_{v,2L}[m] = \mathbf{F}_{2L}\begin{bmatrix}\mathbf{0}_L\\ \mathbf{I}_L\end{bmatrix}\mathbf{e}_v[m] = \underline{\mathbf{d}}_{2L}[m] - \mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_v[m]\,\mathbf{w}[m-1], \quad (41)$$

$$\underline{\mathbf{e}}_{x,2L}[m] = \mathbf{F}_{2L}\begin{bmatrix}\mathbf{0}_L\\ \mathbf{I}_L\end{bmatrix}\mathbf{e}_x[m] = \mathbf{G}_{2L\times 2L}^{01}\mathbf{U}_x[m]\,\mathbf{w}[m-1]. \quad (42)$$

For convenience, we now define the $2NL\times 2NL$-dimensional correlation matrices $\mathbf{Q}_v[m]$ and $\mathbf{Q}_x[m]$ as

$$\mathbf{S}_v[m] = (\mathbf{F}_{2NL\times NL}^{10})^H\,\mathbf{Q}_v[m]\,\mathbf{F}_{2NL\times NL}^{10}, \quad (43)$$

$$\mathbf{S}_x[m] = (\mathbf{F}_{2NL\times NL}^{10})^H\,\mathbf{Q}_x[m]\,\mathbf{F}_{2NL\times NL}^{10}, \quad (44)$$

such that

$$\mathbf{Q}_v[m] = \lambda_v\mathbf{Q}_v[m-1] + (1-\lambda_v)\,\mathbf{D}_v^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{D}_v[m], \quad (45)$$

$$\mathbf{Q}_x[m] = \lambda_x\mathbf{Q}_x[m-1] + (1-\lambda_x)\,\mathbf{D}_x^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{D}_x[m]. \quad (46)$$

In addition, we define the $2NL$-dimensional frequency-domain filter $\underline{\mathbf{w}}_{2NL}[m]$ as

$$\underline{\mathbf{w}}_{2NL}[m] = \mathbf{F}_{2NL\times NL}^{10}\mathbf{w}[m] = \Big[\underline{\mathbf{w}}_{M-N,2L}^T[m]\ \ldots\ \underline{\mathbf{w}}_{M-1,2L}^T[m]\Big]^T, \quad (47)$$

with

$$\underline{\mathbf{w}}_{n,2L}[m] = \mathbf{F}_{2L}\begin{bmatrix}\mathbf{I}_L\\ \mathbf{0}_L\end{bmatrix}\mathbf{w}_n[m]. \quad (48)$$

By pre-multiplying both sides of (40) with $\mathbf{F}_{2NL\times NL}^{10}$, and by using (43) and (44), we obtain

$$\underline{\mathbf{w}}_{2NL}[m] = \mathbf{F}^{10}_{2NL\times NL}\left[\mathbf{S}_v[m] + \frac{1}{\mu}\mathbf{S}_x[m]\right]^{-1}(\mathbf{F}^{10}_{2NL\times NL})^H \left\{\left[\mathbf{Q}_v[m] + \frac{1}{\mu}\frac{\lambda_v}{\lambda_x}\mathbf{Q}_x[m]\right]\underline{\mathbf{w}}_{2NL}[m-1] + (1-\lambda_v)\mathbf{D}_v^H[m]\underline{\mathbf{e}}_{v,2L}[m] - \frac{1}{\mu}\frac{\lambda_v}{\lambda_x}(1-\lambda_x)\,\mathbf{D}_x^H[m]\,\underline{\mathbf{e}}_{x,2L}[m]\right\}, \tag{49}$$

$$\underline{\mathbf{e}}_{v,2L}[m] = \underline{\mathbf{d}}_{2L}[m] - \mathbf{G}^{01}_{2L\times 2L}\mathbf{D}_v[m]\,\underline{\mathbf{w}}_{2NL}[m-1], \tag{50}$$

$$\underline{\mathbf{e}}_{x,2L}[m] = \mathbf{G}^{01}_{2L\times 2L}\mathbf{D}_x[m]\,\underline{\mathbf{w}}_{2NL}[m-1]. \tag{51}$$

In [25], it has been shown that

$$\mathbf{F}^{10}_{2NL\times NL}\,\mathbf{S}_v^{-1}[m]\,(\mathbf{F}^{10}_{2NL\times NL})^H = \mathbf{G}^{10}_{2NL\times 2NL}\,\mathbf{Q}_v^{-1}[m], \tag{52}$$

with the $2NL\times 2NL$-dimensional block diagonal matrix $\mathbf{G}^{10}_{2NL\times 2NL}$ defined as

$$\mathbf{G}^{10}_{2NL\times 2NL} = \mathrm{diag}\left[\mathbf{G}^{10}_{2L\times 2L}\ \ldots\ \mathbf{G}^{10}_{2L\times 2L}\right], \tag{53}$$

with

$$\mathbf{G}^{10}_{2L\times 2L} = \mathbf{F}_{2L}\begin{bmatrix}\mathbf{I}_L & \mathbf{0}_L\\ \mathbf{0}_L & \mathbf{0}_L\end{bmatrix}\mathbf{F}_{2L}^{-1}, \tag{54}$$

such that (49) can be written as

$$\underline{\mathbf{w}}_{2NL}[m] = \mathbf{G}^{10}_{2NL\times 2NL}\left[\mathbf{Q}_v[m] + \frac{1}{\mu}\mathbf{Q}_x[m]\right]^{-1} \left\{\left[\mathbf{Q}_v[m] + \frac{1}{\mu}\frac{\lambda_v}{\lambda_x}\mathbf{Q}_x[m]\right]\underline{\mathbf{w}}_{2NL}[m-1] + (1-\lambda_v)\mathbf{D}_v^H[m]\underline{\mathbf{e}}_{v,2L}[m] - \frac{1}{\mu}\frac{\lambda_v}{\lambda_x}(1-\lambda_x)\,\mathbf{D}_x^H[m]\,\underline{\mathbf{e}}_{x,2L}[m]\right\}. \tag{55}$$

In the sequel, we will assume equal exponential forgetting factors for speech and noise, i.e. $\lambda_x = \lambda_v = \lambda$, such that using $\mathbf{G}^{10}_{2NL\times 2NL}\underline{\mathbf{w}}_{2NL}[m-1] = \underline{\mathbf{w}}_{2NL}[m-1]$, (55) reduces to

$$\boxed{\underline{\mathbf{w}}_{2NL}[m] = \underline{\mathbf{w}}_{2NL}[m-1] + (1-\lambda)\,\mathbf{G}^{10}_{2NL\times 2NL}\left[\mathbf{Q}_v[m] + \frac{1}{\mu}\mathbf{Q}_x[m]\right]^{-1}\cdot \left\{\mathbf{D}_v^H[m]\,\underline{\mathbf{e}}_{v,2L}[m] - \frac{1}{\mu}\mathbf{D}_x^H[m]\,\underline{\mathbf{e}}_{x,2L}[m]\right\}} \tag{56}$$

When the trade-off parameter $1/\mu = 0$, this algorithm is equal to the multi-channel frequency-domain adaptive filtering algorithm derived in [25,26], ap-

plied to the GSC. For $1/\mu > 0$, the $2NL$-dimensional additional *regularization term*

$$\underline{\mathbf{r}}_{2NL}[m] = \frac{1}{\mu}\mathbf{D}_x^H[m]\,\underline{\mathbf{e}}_{x,2L}[m] = \frac{1}{\mu}\mathbf{D}_x^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{D}_x[m]\,\underline{\mathbf{w}}_{2NL}[m-1] \quad (57)$$

limits speech distortion due to speech leakage components in the noise references.

3.4 Practical Implementation

If we take a closer look at (56), we notice that $\mathbf{D}_v[m]$ and $\underline{\mathbf{e}}_{v,2L}[m]$ can be computed only during noise-only periods, whereas $\mathbf{D}_x[m]$ and $\underline{\mathbf{e}}_{x,2L}[m]$ can be computed only during speech periods. We will now take a similar approach as in the standard GSC, i.e. we will *update the filter coefficients only during noise-only periods*. Since during noise-only periods the (instantaneous) correlation matrix $\mathbf{D}_x^H[m]\mathbf{G}_{2L\times 2L}^{01}\mathbf{D}_x[m]$ of the clean speech signal, required in the computation of the regularization term $\underline{\mathbf{r}}_{2NL}[m]$, is not available, we will approximate this term by the (average) correlation matrix $\mathbf{Q}_x[m]$ [2], i.e. the regularization term will be computed as

$$\underline{\mathbf{r}}_{2NL}[m] \approx \frac{1}{\mu}\mathbf{Q}_x[m]\,\underline{\mathbf{w}}_{2NL}[m-1]. \quad (59)$$

In fact, using the correlation matrix $\mathbf{Q}_x[m]$ instead of $\mathbf{D}_x^H[m]\mathbf{G}_{2L\times 2L}^{01}\mathbf{D}_x[m]$ is quite similar to low-pass filtering a similar time-domain regularization term, which has been proposed in [20] to improve the performance in highly non-stationary noise. Using the assumption that speech and noise components are uncorrelated, the speech correlation matrix will be computed in practice as

$$\mathbf{Q}_x[m] \approx \mathbf{Q}_y[m] - \mathbf{Q}_v[m], \quad (60)$$

where $\mathbf{Q}_y[m]$ is the $2NL\times 2NL$-dimensional correlation matrix updated during speech periods, i.e.

$$\mathbf{Q}_y[m] = \lambda\mathbf{Q}_y[m-1] + (1-\lambda)\,\mathbf{D}_y^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{D}_y[m], \quad (61)$$

where $\mathbf{D}_y[m]$ is defined similarly as $\mathbf{D}_x[m]$. The complete recursive frequency-domain algorithm for updating the filter $\underline{\mathbf{w}}_{2NL}[m]$ is summarized in Table 1.

[2] Note that a similar reasoning for computing the term $\mathbf{D}_v^H[m]\,\underline{\mathbf{e}}_{v,2L}[m]$ during speech periods is not possible, since

$$\mathbf{D}_v^H[m]\,\underline{\mathbf{e}}_{v,2L}[m] = \mathbf{D}_v^H[m]\,\underline{\mathbf{d}}_{2L}[m] - \mathbf{D}_v^H[m]\,\mathbf{G}_{2L\times 2L}^{01}\mathbf{D}_v[m]\,\underline{\mathbf{w}}_{2NL}[m-1] \quad (58)$$

cannot easily be approximated, because of the term $\mathbf{D}_v^H[m]\,\underline{\mathbf{d}}_{2L}[m]$.

4 Frequency-domain adaptive algorithms

The algorithm in Table 1 constitutes a general framework from which different adaptive algorithms can be derived by introducing different types of approximations. Some of these algorithms have already been presented in the literature [21], whereas other algorithms represent novel techniques for implementing the SDW-MWF cost function in the frequency-domain. Figure 3 depicts the block diagram of the algorithms for updating the filter coefficients that will be discussed in this section. The difference between these algorithms consists of whether block-structured or diagonal correlation matrices are used (cf. Sections 4.1 and 4.2) and whether the update formula is constrained or unconstrained (cf. Section 4.3).

4.1 Block-structured Correlation Matrices (Algo 1)

Since the correlation matrices $\mathbf{Q}_v[m]$ and $\mathbf{Q}_y[m]$ do not have a special structure, both updating these correlation matrices according to (45) and (61), and the matrix inversion in (56) are computationally expensive operations $[O((NL)^3)]$, such that in fact the algorithm in Table 1 is not very useful in practice. However, in [25,26] it has been shown that the matrix $\mathbf{G}^{01}_{2L\times 2L}$ may be well approximated by $\mathbf{I}_{2L}/2$, because – for large L – the off-diagonal elements of $\mathbf{G}^{01}_{2L\times 2L}$ are small compared to the diagonal elements.

Using this approximation, we obtain the following update formula for the block-structured correlation matrices $\tilde{\mathbf{Q}}_v[m]$ and $\tilde{\mathbf{Q}}_y[m]$,

$$\tilde{\mathbf{Q}}_v[m] = \lambda\tilde{\mathbf{Q}}_v[m-1] + (1-\lambda)\,\mathbf{D}_v^H[m]\,\mathbf{D}_v[m]/2, \tag{62}$$
$$\tilde{\mathbf{Q}}_y[m] = \lambda\tilde{\mathbf{Q}}_y[m-1] + (1-\lambda)\,\mathbf{D}_y^H[m]\,\mathbf{D}_y[m]/2, \tag{63}$$

which are $N\times N$ block matrices with $2L\times 2L$-dimensional diagonal blocks $\tilde{\mathbf{Q}}_{v,np}[m]$ and $\tilde{\mathbf{Q}}_{y,np}[m]$, $n = M-N,\ldots,M-1$, $p = M-N,\ldots,M-1$. Hence, we obtain the following update formula for the filter coefficients,

$$\boxed{\begin{aligned}\underline{\mathbf{w}}_{2NL}[m] = \underline{\mathbf{w}}_{2NL}[m-1] + \rho(1-\lambda)\,\mathbf{G}^{10}_{2NL\times 2NL}\left[\tilde{\mathbf{Q}}_v[m] + \frac{1}{\mu}\tilde{\mathbf{Q}}_x[m]\right]^{-1}\cdot \\ \left\{\mathbf{D}_v^H[m]\,\underline{\mathbf{e}}_{v,2L}[m] - \underline{\mathbf{r}}_{2NL}[m]\right\}\end{aligned}} \tag{64}$$

where ρ is a step size parameter and the regularization term now is defined as

$$\underline{\mathbf{r}}_{2NL}[m] = \frac{1}{\mu}\tilde{\mathbf{Q}}_x[m]\,\underline{\mathbf{w}}_{2NL}[m-1]\,, \tag{65}$$

with $\tilde{\mathbf{Q}}_x[m] = \tilde{\mathbf{Q}}_y[m] - \tilde{\mathbf{Q}}_v[m]$. This update formula will be referred to as *Algo 1*.

The update formula in (64) involves computing the inverse of the matrix $\tilde{\mathbf{Q}}_v[m] + 1/\mu\,\tilde{\mathbf{Q}}_x[m]$. It is well known that the inverse of an $N \times N$ block matrix $\mathbf{Q}$ with $2L \times 2L$-dimensional diagonal blocks $\mathbf{Q}_{np}$, i.e.

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{M-N,M-N} & \cdots & \mathbf{Q}_{M-N,M-1} \\ \vdots & & \vdots \\ \mathbf{Q}_{M-1,M-N} & \cdots & \mathbf{Q}_{M-1,M-1} \end{bmatrix}, \tag{66}$$

is again a block matrix with diagonal blocks. Computing the inverse corresponds to inverting $2L$ $N \times N$-dimensional matrices, which is attractive from a computational complexity point of view. More in particular, the block matrix $\mathbf{Q}$ can be permuted into the block diagonal matrix $\bar{\mathbf{Q}}$,

$$\bar{\mathbf{Q}} = \mathrm{diag}\left[\, \bar{\mathbf{Q}}_0 \,\ldots\, \bar{\mathbf{Q}}_{2L-1} \,\right], \tag{67}$$

with $N \times N$-dimensional sub-matrices $\bar{\mathbf{Q}}_l$, $l = 0 \ldots 2L-1$, on its diagonal, by means of row and column permutations, i.e.

$$\bar{\mathbf{Q}} = \mathbf{A}^T\,\mathbf{Q}\,\mathbf{A}. \tag{68}$$

The matrix $\mathbf{A}$ is a $2NL \times 2NL$-dimensional column permutation matrix (and hence $\mathbf{A}^T$ is a row permutation matrix), consisting of $2NL$ $2L \times N$-dimensional sub-matrices $\mathbf{A}_{nl}$, $n = M-N, \ldots, M-1$, $l = 0 \ldots 2L-1$, where the (l,n)-th element of $\mathbf{A}_{nl}$ is equal to 1. It readily follows that

$$\mathbf{Q}^{-1} = \mathbf{A}\,\bar{\mathbf{Q}}^{-1}\,\mathbf{A}^T, \tag{69}$$

where $\bar{\mathbf{Q}}^{-1}$ can be computed by inverting the $N \times N$-dimensional sub-matrices $\bar{\mathbf{Q}}_l$ on its diagonal, i.e.

$$\bar{\mathbf{Q}}^{-1} = \mathrm{diag}\left[\, \bar{\mathbf{Q}}_0^{-1} \,\ldots\, \bar{\mathbf{Q}}_{2L-1}^{-1} \,\right]. \tag{70}$$

In addition, one should make sure that the matrix $\tilde{\mathbf{Q}}_v[m] + 1/\mu\,\tilde{\mathbf{Q}}_x[m]$ in (64) is positive definite. When this matrix is not positive definite, this actually has the same effect as a negative step size ρ, i.e. leading to divergence of the filter coefficients. The noise correlation matrix $\tilde{\mathbf{Q}}_v[m]$ is always positive definite, but the speech correlation matrix $\tilde{\mathbf{Q}}_x[m]$ may not always be positive definite (especially for non-stationary signals), since it is computed as $\tilde{\mathbf{Q}}_x[m] = \tilde{\mathbf{Q}}_y[m] - \tilde{\mathbf{Q}}_v[m]$, where $\tilde{\mathbf{Q}}_y[m]$ and $\tilde{\mathbf{Q}}_v[m]$ are estimated during (different) speech periods and noise-only periods. Checking the positive definiteness of a matrix comes down to computing its eigenvalues. By using (68) and the fact that $\mathbf{A}\mathbf{A}^T = \mathbf{I}_{2NL}$ and $\det(\mathbf{A}) = \pm 1$, it readily follows that

$$\det(\mathbf{Q} - \gamma\mathbf{I}_{2NL}) = \det\left(\mathbf{A}(\bar{\mathbf{Q}} - \gamma\mathbf{I}_{2NL})\mathbf{A}^T\right) = \det(\bar{\mathbf{Q}} - \gamma\mathbf{I}_{2NL})\,, \tag{71}$$

such that the eigenvalues γ of the block matrix $\mathbf{Q}$ are equal to the set of eigenvalues of its $N \times N$-dimensional sub-matrices $\bar{\mathbf{Q}}_l$, $l = 0 \dots 2L-1$.

Hence, instead of directly computing the inverse of the matrix $\tilde{\mathbf{Q}}_v[m]+1/\mu\,\tilde{\mathbf{Q}}_x[m]$ in (64), we first compute the eigenvalues of the matrix $\tilde{\mathbf{Q}}_x[m]$, and then use the inverse of the positive definite matrix

$$\tilde{\mathbf{Q}}_v[m] + \frac{1}{\mu}\left[\tilde{\mathbf{Q}}_x[m] - \min(\gamma_{min}, 0)\,\mathbf{I}_{2NL}\right] + \delta\,\mathbf{I}_{2NL} \tag{72}$$

in (64), with γ_{min} the smallest eigenvalue of $\tilde{\mathbf{Q}}_x[m]$ and δ a small positive regularization factor (a typical value is $\delta = 1e-6$). Whereas in general computing the smallest eigenvalue of an $N \times N$-dimensional Hermitian matrix is computationally quite complex, for $N = 2$ (e.g. in a two- or a three-microphone application) the smallest eigenvalue $\gamma_{l,min}$ of the sub-matrix

$$\bar{\mathbf{Q}}_l = \begin{bmatrix} \bar{q}_{l,11} & \bar{q}_{l,12} \\ \bar{q}^*_{l,12} & \bar{q}_{l,22} \end{bmatrix}, \tag{73}$$

with $\bar{q}_{l,11}$ and $\bar{q}_{l,22}$ real-valued, is equal to

$$\gamma_{l,min} = \frac{(\bar{q}_{l,11} + \bar{q}_{l,22}) - \sqrt{(\bar{q}_{l,11} - \bar{q}_{l,22})^2 + 4|\bar{q}_{l,12}|^2}}{2}. \tag{74}$$

4.2 Diagonal Correlation Matrices (Algo 2 and 3)

In a further approximation, we can decouple the updates for the N filters $\underline{\mathbf{w}}_{n,2L}[m]$ in (64) by neglecting the off-diagonal elements of the matrix $\tilde{\mathbf{Q}}_v[m]+1/\mu\,\tilde{\mathbf{Q}}_x[m]$, which represent the inter-channel correlation. Hence, the update formula for the filter coefficients $\underline{\mathbf{w}}_{n,2L}[m]$, $n = M-N, \dots, M-1$ becomes

$$\boxed{\begin{aligned}\underline{\mathbf{w}}_{n,2L}[m] = \underline{\mathbf{w}}_{n,2L}[m-1] + \rho(1-\lambda)\,\mathbf{G}^{10}_{2L\times 2L}\left[\tilde{\mathbf{Q}}_{v,nn}[m] + \frac{1}{\mu}\tilde{\mathbf{Q}}_{x,nn}[m]\right]^{-1} \cdot \\ \left\{\mathbf{D}^H_{v,n}[m]\,\underline{\mathbf{e}}_{v,2L}[m] - \underline{\mathbf{r}}_{n,2L}[m]\right\}\end{aligned}} \tag{75}$$

with $\tilde{\mathbf{Q}}_{v,nn}[m]$ and $\tilde{\mathbf{Q}}_{x,nn}[m]$ the $2L \times 2L$-dimensional diagonal sub-matrices on the diagonal of $\tilde{\mathbf{Q}}_v[m]$ and $\tilde{\mathbf{Q}}_x[m]$, and $\underline{\mathbf{r}}_{n,2L}[m]$ a $2L$-dimensional sub-vector of $\underline{\mathbf{r}}_{2NL}[m]$ [3]. This update formula will be referred to as *Algo 2*.

Ensuring the positive definiteness of $\tilde{\mathbf{Q}}_{x,nn}[m]$ now is straightforward, since the eigenvalues of $\tilde{\mathbf{Q}}_{x,nn}[m]$ are equal to the diagonal elements. As will be

[3] Note that we still use the off-diagonal elements of $\tilde{\mathbf{Q}}_x[m]$ for computing the regularization term $\underline{\mathbf{r}}_{2NL}[m]$, i.e. (65).

shown in the experimental results in Section 6, updating the filter coefficients using block-structured correlation matrices gives rise to a faster convergence than using diagonal correlation matrices, since the inter-channel correlation is taken into account. This has also been observed in [26] when applying this algorithm to the GSC, i.e. for $N = 2$ and $1/\mu = 0$.

Where in (75) a different step size matrix $\tilde{\mathbf{Q}}_{v,nn}[m] + 1/\mu\tilde{\mathbf{Q}}_{x,nn}[m]$ is used for each channel n, it is also possible to use a common step size matrix $\tilde{\mathbf{Q}}_c$, e.g. the sum or the average over all channels, i.e.

$$\begin{aligned} \underline{\mathbf{w}}_{n,2L}[m] &= \underline{\mathbf{w}}_{n,2L}[m-1] + \rho(1-\lambda)\,\mathbf{G}^{10}_{2L\times 2L}\tilde{\mathbf{Q}}_c^{-1}[m]\cdot \\ &\qquad \left\{\mathbf{D}^H_{v,n}[m]\,\underline{\mathbf{e}}_{v,2L}[m] - \underline{\mathbf{r}}_{n,2L}[m]\right\} \\ \tilde{\mathbf{Q}}_c[m] &= \left(\frac{1}{N}\right)\sum_{n=M-N}^{M-1}\tilde{\mathbf{Q}}_{v,nn}[m] + \frac{1}{\mu}\tilde{\mathbf{Q}}_{x,nn}[m] \end{aligned} \tag{76}$$

This update formula will be referred to as *Algo 3*. In fact, this algorithm is very similar to the algorithm already presented in [21]. Note however that the algorithm in [21] has been derived as a frequency-domain implementation of a time-domain stochastic gradient algorithm for minimizing the (time-domain) cost function in (13).

4.3 Unconstrained Algorithms

In Section 4.1 the term $\mathbf{G}^{01}_{2L\times 2L}$ in the calculation of the correlation matrices has been approximated by $\mathbf{I}_{2L}/2$. It is also possible to use the same approximation for the term $\mathbf{G}^{10}_{2L\times 2L}$ and hence approximate $\mathbf{G}^{10}_{2NL\times 2NL}$ in the update formula for the filter coefficients in (56) by

$$\mathbf{G}^{10}_{2NL\times 2NL} \approx \mathrm{diag}\left[\mathbf{I}_{2L}/2 \;\ldots\; \mathbf{I}_{2L}/2\right] = \mathbf{I}_{2NL}/2, \tag{77}$$

resulting in the following so-called unconstrained update formula, i.e.

$$\begin{aligned} \underline{\mathbf{w}}_{2NL}[m] = \underline{\mathbf{w}}_{2NL}[m-1] + \frac{(1-\lambda)}{2}\left[\mathbf{Q}_v[m] + \frac{1}{\mu}\mathbf{Q}_x[m]\right]^{-1}\cdot \\ \left\{\mathbf{D}^H_v[m]\,\underline{\mathbf{e}}_{v,2L}[m] - \underline{\mathbf{r}}_{2NL}[m]\right\} \end{aligned} \tag{78}$$

This update formula gives rise to a lower computational complexity, since it requires $2N$ less FFT operations, cf. Section 5. However, when using this update formula one cannot guarantee that the second half of $\mathbf{F}^{-1}_{2L}\underline{\mathbf{w}}_{n,2L}[m]$, $n = M-N,\ldots,M-1$, is equal to zero, cf. (48). In addition, for the unconstrained algorithms one can also approximate the correlation matrices $\mathbf{Q}_v[m]$ and $\mathbf{Q}_x[m]$ by block-structured or diagonal matrices.

4.4 Summary

Summarizing all presented algorithms in Section 4, the update formula for the filter coefficients $\underline{\mathbf{w}}_{2NL}[m]$ can be written as

$$\boxed{\begin{aligned}\underline{\mathbf{w}}_{2NL}[m] &= \underline{\mathbf{w}}_{2NL}[m-1] + \rho(1-\lambda)\,\mathbf{\Lambda}[m]\Big\{\mathbf{D}_v^H[m]\,\underline{\mathbf{e}}_{v,2L}[m] - \underline{\mathbf{r}}_{2NL}[m]\Big\} \\ \underline{\mathbf{r}}_{2NL}[m] &= \frac{1}{\mu}\tilde{\mathbf{Q}}_x[m]\,\underline{\mathbf{w}}_{2NL}[m-1]\end{aligned}} \tag{79}$$

where the $2NL \times 2NL$-dimensional step size matrix $\mathbf{\Lambda}[m]$ is summarized in Table 2. For all algorithms, the matrix $\tilde{\mathbf{Q}}_x[m]$ needs to be regularized in order to make sure that it is positive definite. The algorithm already presented in [21] corresponds to the constrained version of Algo 3. Figure 3 depicts the block diagram of these algorithms for updating the filter coefficients.

5 Computational complexity

Table 3 summarizes the computational complexity of several frequency-domain adaptive algorithms for robust multi-microphone noise reduction: the QIC-GSC using the Scaled Projection Algorithm (SPA) [17], the stochastic gradient buffer-based implementation of the SDW-MWF [20], and the different adaptive algorithms implementing the frequency-domain criterion for the SDW-MWF, which have been discussed in this paper. The computational complexity is expressed as the number of operations, i.e. real multiplications and additions (MAC), per second. We assume that one complex multiplication is equivalent to 4 real multiplications and 2 real additions and that a $2L$-point FFT of a real input vector requires $2L \log_2 2L$ real MACs (using the radix-2 FFT algorithm). For Algo 1 the cost of ensuring the positive definiteness of the block-structured step size matrix, and hence calculating the smallest eigenvalue of $\tilde{\mathbf{Q}}_x[m]$, has been included in the computational complexity. Therefore the computational complexity for Algo 1 in Table 3 is only valid for $N = 2$, i.e. when a closed-form expression is available for calculating the smallest eigenvalue, cf. (74). The computational complexity has been explicitly calculated for the parameter values used in the simulations in Section 6, i.e. $M = 3$, $L = 128$, sampling frequency $f_s = 16\,\text{kHz}$, and either $N = M - 1$ or $N = M$ input channels to the multichannel adaptive filter.

From this table we can draw the following conclusions:

- The complexity of all SDW-MWF algorithms (constrained version) is higher than the complexity of the QIC-GSC. However, as has been shown in [19], the SDW-MWF obtains a better noise reduction than the QIC-GSC for small model errors, while guaranteeing robustness against large model errors.

- The complexity of the adaptive algorithms implementing the frequency-domain criterion for the SDW-MWF is lower than the stochastic gradient buffer-based implementation of the SDW-MWF [20]. However, this only remains true for a small number of input channels, since the complexity of these frequency-domain algorithms contains a quadratic term $O(N^2)$.
- The complexity of the algorithms using a diagonal step size matrix (Algo 2 and Algo 3) is smaller than the complexity of Algo 1 using a block-structured step size matrix. As will be shown, these algorithms however give rise to a slower convergence behavior.
- The unconstrained algorithms require $2N$ less FFT operations than the constrained algorithms.

6 Experimental Results

In this section, experimental results are presented for a hearing aid application. For small-sized microphone arrays as typically used in hearing aids, robustness is very important, since these microphone arrays exhibit a large sensitivity to signal model errors [32]. Section 6.1 describes the setup and defines the performance measures used here. In Section 6.2 the performance, i.e. SNR improvement and speech distortion, and the convergence behavior of different adaptive algorithms is analyzed, and the effect of different parameter settings (i.e. filter $\mathbf{w}_0$ and $1/\mu$) on the performance and the robustness against signal model errors is evaluated. In Section 6.3 the performance difference between using a perfect voice activity detection (VAD) mechanism and using a non-perfect VAD is investigated for different input SNRs. In Section 6.4 the tracking performance is analyzed for a time-varying scenario.

6.1 Setup and Performance Measures

A hearing aid with $M = 3$ omni-directional microphones (Knowles FG-3452) in an end-fire configuration has been mounted on the right ear of a dummy head in an office room. The distance between the first and the second microphone is about 1 cm and the distance between the second and the third microphone is about 1.5 cm. The reverberation time T_{60} of the room is approximately 700 ms. The speech and the noise sources are positioned at a distance of 1 m from the head: the speech source in front of the head (0°), and the noise sources at an angle θ with respect to the speech source. The recording environment is depicted in Figure 4. Both the speech and the noise signal have a level of 70 dB at the center of the head. For evaluation purposes, the speech and the noise signal are recorded separately. The sampling frequency is equal to 16 kHz.

The microphone signals are pre-whitened prior to processing in order to improve the intelligibility, and the output signal $z[k]$ is de-whitened accordingly [33]. The microphones are calibrated using anechoic recordings of a speech-

weighted noise signal at 0° with the microphone array mounted on the head. A delay-and-sum beamformer is used for the fixed beamformer, since – in the case of small microphone distances – this beamformer is quite robust against signal model errors. The blocking matrix pair-wisely subtracts the time-aligned calibrated microphone signals to generate the noise references.

To assess the performance of the different algorithms, the broadband intelligibility weighted signal-to-noise ratio improvement $\Delta\mathrm{SNR}_{\mathrm{intellig}}$ is used, which is defined as [34]

$$\Delta\mathrm{SNR}_{\mathrm{intellig}} = \sum_i I_i \left(\mathrm{SNR}_{i,\mathrm{out}} - \mathrm{SNR}_{i,\mathrm{in}}\right), \tag{80}$$

where the band importance function I_i expresses the importance of the ith one-third octave band with center frequency f_i^c for intelligibility, and where $\mathrm{SNR}_{i,\mathrm{out}}$ and $\mathrm{SNR}_{i,\mathrm{in}}$ represent respectively the output SNR and the input SNR (in dB) in this band. The center frequencies f_i^c and the values I_i are defined in [35]. The intelligibility weighted SNR improvement reflects how much the speech intelligibility is improved by the noise reduction algorithms, but does not take into account speech distortion.

In order to measure the amount of (linear) speech distortion, we similarly define an intelligibility weighted spectral distortion measure $\mathrm{SD}_{\mathrm{intellig}}$,

$$\mathrm{SD}_{\mathrm{intellig}} = \sum_i I_i \,\mathrm{SD}_i \ , \tag{81}$$

with SD_i the average spectral distortion (dB) in the ith one-third octave band,

$$\mathrm{SD}_i = \frac{1}{\left(2^{1/6} - 2^{-1/6}\right) f_i^c} \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} \left|10 \log_{10} G_x(f)\right| df, \tag{82}$$

with $G_x(f)$ the power transfer function for the speech component from the input to the output of the noise reduction algorithm.

In order to exclude the effect of the spatial pre-processor, the performance measures (80) and (81) are calculated with respect to the output of the fixed beamformer, i.e. the speech reference $y_0[k]$. In some experiments, a microphone gain mismatch of 4 dB is applied to the second microphone in order to illustrate the sensitivity to signal model errors. Among the different possible signal model errors, microphone mismatch has been found to be quite harmful to the performance of the GSC in a hearing aid application [32]. In hearing aids, microphones are rarely matched in gain and phase, with typical gain and phase differences of up to 6 dB and 10° [36].

All algorithms are evaluated with a filter length $L = 128$. In Sections 6.2 and 6.4, the input SNR of the microphone signals is equal to 0 dB, whereas in

Section 6.3 different input SNRs, ranging from $-10\,\text{dB}$ to $5\,\text{dB}$, are used. In Section 6.2 a (non-perfect) energy-based VAD [27] is used, whereas in Section 6.4 a perfect VAD is used, i.e. the speech periods and the noise-only periods have been marked manually. In Section 6.3 the performance difference between using a perfect and a non-perfect VAD is investigated.

6.2 SNR Improvement and Robustness Against Microphone Mismatch

For the experiments in this section, the desired speech source at 0° consists of sentences from the HINT-database [37] spoken by a male speaker, and a complex noise scenario consisting of 5 spectrally non-stationary multi-talker babble noise sources at 75°, 120°, 180°, 240° and 285°, is used. The input SNR of the microphone signals is equal to $0\,\text{dB}$ and an energy-based VAD [27] is used. As will be seen in Section 6.3, the effect of using an energy-based VAD instead of a perfect VAD is quite small for SNR=$0\,\text{dB}$.

Figure 5 plots the convergence of the SNR improvement for different adaptive algorithms (constrained vs. unconstrained, block-structured vs. diagonal step size matrix) for different values of the step size parameter ρ and the exponential forgetting factor λ. Instead of λ we use the corresponding time T_λ, i.e. the factor λ corresponds to an averaging of the correlation matrices over approximately $1/(1-\lambda)$ blocks of L samples, such that

$$T_\lambda = \frac{1}{1-\lambda}\frac{L}{f_s}\,. \tag{83}$$

Hence, for $L = 128$, $T_\lambda = 0.8\,\text{s}$ corresponds to $\lambda = 0.99$, $T_\lambda = 1.6\,\text{s}$ corresponds to $\lambda = 0.995$, and $T_\lambda = 3.2\,\text{s}$ corresponds to $\lambda = 0.9975$. Typically, quite large values are used for the exponential forgetting factor, implying that mainly the long-term spatial and spectral characteristics of the speech and the noise sources are used. In this experiment, we have used the SDR-GSC ($N = 2$) with trade-off parameter $1/\mu = 0.5$ and with no microphone mismatch present. Obviously, similar plots can be obtained for the SP-SDW-MWF ($N = 3$), for different values of the trade-off parameter and when microphone mismatch is present. From Figure 5 it can be seen that a block-structured step size matrix gives rise to a substantially faster convergence than a diagonal step size matrix, which can be explained by the fact that a block-structured step size matrix takes into account the inter-channel correlation. Hence, the observations in [26] for the GSC are also valid for the SDR-GSC and the SP-SDW-MWF. In addition, the main factor affecting the convergence speed is $\rho(1-\lambda)$, i.e. the larger ρ, the faster the convergence and the larger λ, the slower the convergence. However, the SNR improvement at convergence will be worse for larger $\rho(1-\lambda)$ because of the larger misadjustment of the adaptive filter coefficients (taking $\rho(1-\lambda)$ too large obviously even leads to divergence). The SNR improvement at convergence is slightly better for larger λ, because a

better estimate of the regularization term is obtained (for spectrally and/or spatially stationary sources). Taking λ too small results in a highly time-varying regularization term, which is undesirable. Moreover, for this scenario, the performance difference between the constrained and the unconstrained update formula is quite small. For the experimental results in this section and in Section 6.3 we will use $\rho = 2$ and $T_\lambda = 1.6\,\text{s}$.

Figure 6 plots the SNR improvement and the speech distortion at convergence for the SDR-GSC ($N = 2$) and for the SP-SDW-MWF ($N = 3$) as a function of the trade-off parameter $1/\mu$, using the unconstrained update formula with block-structured step size matrix. This figure also depicts the effect of a gain mismatch of 4 dB at the second microphone. Similar conclusions as in [19,20] can be drawn:

- *SDR-GSC ($N = 2$)*: In the absence of microphone mismatch, the amount of speech leakage into the noise references is limited, such that the speech distortion is small for all $1/\mu$. However, since there is some speech leakage present due to reverberation, the SNR improvement decreases for increasing $1/\mu$. In the presence of microphone mismatch, the amount of speech leakage into the noise references grows. For the standard GSC, i.e. $1/\mu = 0$, significant speech distortion now occurs and the SNR improvement is seriously degraded. Setting $1/\mu > 0$ improves the performance of the GSC in the presence of signal model errors, i.e. the speech distortion decreases and the SNR degradation becomes smaller.
- *SP-SDW-MWF ($N = 3$)*: The SNR improvement and the speech distortion also decrease for increasing $1/\mu$. Compared to the SDR-GSC, the speech distortion however is larger [4], but both the SNR improvement and the speech distortion are hardly affected by microphone mismatch.

Figure 8 shows the spectrograms of the microphone signal $u_1[k]$, the speech reference signal $y_0[k]$, and the output signal $z[k]$ for the GSC ($1/\mu = 0$), the SDR-GSC ($1/\mu = 0.5$) and the SP-SDW-MWF ($1/\mu = 0.1, 0.5$), with and without mismatch. As can be observed from this figure, in the presence of mismatch significant speech distortion occurs for the GSC, whereas less distortion occurs for the SDR-GSC ($1/\mu = 0.5$). Although the SP-SDW-MWF seems to reduce substantially more noise than the SDR-GSC, also more spectral distortion occurs. However, the performance difference for the SP-SDW-MWF between mismatch and no mismatch is hardly noticeable.

Figure 7 depicts the SNR improvement and the speech distortion of the QIC-GSC as a function of the constraint value β^2, with and without microphone mismatch. Like the SDR-GSC, the QIC-GSC increases the robustness of the GSC: in the presence of mismatch, the speech distortion decreases for de-

[4] In [19], it has been shown that the SP-SDW-MWF can be interpreted as an SDR-GSC with a single-channel post-filter in the absence of speech leakage.

creasing β^2 (but also the SNR improvement decreases). The constraint value β^2 should be chosen such that the maximum allowable speech distortion level is not exceeded for the largest possible model errors. E.g. a maximum allowable speech distortion level of 4 dB for a gain mismatch of 4 dB, corresponding to $\beta^2 = 0.3$, results in an SNR improvement of 4.8 dB with mismatch and 5.0 dB without mismatch. On the other hand, for the SDR-GSC the emphasis on speech distortion is only increased when the amount of speech leakage grows. As a result, a better SNR improvement is obtained without mismatch (6.8 dB for $1/\mu = 0.9$), while guaranteeing sufficient robustness when mismatch occurs (4.8 dB). The SP-SDW-MWF even further improves the performance in the presence of mismatch (6.3 dB).

6.3 Impact of energy-based VAD

In this section, we compare the performance, i.e. the SNR improvement and the speech distortion, between using a perfect VAD and using an energy-based VAD [27]. This comparison is performed for different input SNRs, ranging from -10 dB to 5 dB, which is an important range for hearing aid applications. We have used the same speech and noise scenario as in Section 6.2.

Figure 9 depicts the speech component $x_0[k]$ in the speech reference, together with the perfect VAD and the output of the energy-based VAD for different input SNRs. For each input SNR, the percentage of speech frames classified as noise and noise frames classified as speech is indicated. As can be seen, the percentage of speech frames classified as noise decreases as the input SNR grows, whereas the percentage of noise frames classified as speech increases as the input SNR grows. However, wrongly classified speech frames have a larger impact on the performance than wrongly classified noise frames, as already shown in [38]. Hence, we expect the performance difference between using a perfect and an energy-based VAD to be larger for low input SNRs.

Figure 10 plots the SNR improvement and the speech distortion at convergence for the GSC ($1/\mu = 0$) and the SDR-GSC ($1/\mu = 0.5$) as a function of the input SNR, when using a perfect VAD and when using the energy-based VAD, with and without microphone mismatch. We have used the unconstrained update formula with block-structured step size matrix. For all input SNRs, the conclusions from Section 6.2 still hold, i.e. in comparison with the GSC the SDR-GSC gives rise to an improved robustness (lower speech distortion and smaller SNR degradation) when microphone mismatch occurs. These effects are more pronounced for high SNRs, presumably due to the fact that relatively more speech leakage components are present in the noise references. Compared to the perfect VAD, the energy-based VAD gives rise to a degraded performance, i.e. lower SNR improvement and slightly higher speech distortion. This effect is more pronounced for low SNRs, since at low SNRs the energy-based VAD generates more speech detection errors.

Figure 11 plots the SNR improvement and the speech distortion at convergence for the SP-SDW-MWF ($1/\mu = 0.1, 0.5$) as a function of the input SNR, when using a perfect VAD and when using the energy-based VAD, with and without microphone mismatch. It can be observed that the trade-off parameter $1/\mu$ mainly has an influence on the speech distortion and to a smaller extent on the SNR improvement. Moreover, for all conditions the performance measures are hardly affected by microphone mismatch. However, it can be observed that compared to the perfect VAD, the energy-based VAD gives rise to a degraded performance, especially for low SNRs. In general, the performance of the SP-SDW-MWF is better than the SDR-GSC when microphone mismatch occurs, also when using the energy-based VAD.

6.4 Tracking Performance

To investigate the tracking performance of the frequency-domain adaptive algorithms, we consider a noise scenario consisting of 5 multi-talker babble noise sources at 75°, 120°, 180°, 240° and 285°, and a switching speech scenario with a speech source at 0° (scenario 1) and 45° (scenario 2). Every 20 seconds, the speech scenario suddenly changes between scenario 1 and 2. We have used a stationary speech-weighted noise signal both for the speech source and for the noise sources. The speech component consists of alternating segments of signal and silence, each with a length of 1600 samples. The input SNR of the microphone signals is equal to 0 dB and we have used a perfect VAD.

In addition to the SNR improvement and the speech distortion, we will also compare the filter convergence, defined as

$$\Delta\mathbf{w}[m] = \frac{||\mathbf{w}[m] - \mathbf{w}_{opt}||}{||\mathbf{w}_{opt}||}, \tag{84}$$

where for each of the two noise scenarios the "optimal" filter $\mathbf{w}_{opt}$ is calculated using (14) and where the correlation matrices in (14) are constructed using all available speech and noise samples.

Figure 12 plots the filter convergence $\Delta\mathbf{w}[m]$ for the SDR-GSC ($N = 2$), using the unconstrained update formula (block-structured vs. diagonal step size matrix), for different values of ρ and T_λ. The trade-off parameter $1/\mu = 0.5$ and a microphone mismatch of 4 dB is present. For the switching scenario, similar results as in Figure 5 are obtained: the block-structured step size matrix gives rise to a substantially faster convergence than the diagonal step size matrix and the main factor affecting the convergence speed is $\rho(1-\lambda)$, i.e. the larger ρ, the faster the convergence and the larger λ, the slower the convergence. For equal $\rho(1-\lambda)$, the convergence behavior is smoother for larger λ.

Figure 13 plots the SNR improvement, the speech distortion and the filter convergence for the GSC ($1/\mu = 0$) and the SDR-GSC ($1/\mu = 0.5$), both us-

ing the unconstrained update formula with block-structured step size matrix, with and without mismatch. The step size parameter $\rho = 2$ and $T_\lambda = 0.8\,\mathrm{s}$. Again, this figure shows that when microphone mismatch is present, the noise reduction performance of the GSC decreases (quite substantially for scenario 2) and the speech distortion substantially increases (more for scenario 2 than for scenario 1). Compared to the GSC, the SDR-GSC ($1/\mu = 0.5$) gives rise to considerably less speech distortion when microphone mismatch is present, whereas the SNR improvement for both scenarios only slightly decreases.

7 Conclusion

In this paper, we have presented a novel frequency-domain criterion for the SDW-MWF cost function, trading off noise reduction and speech distortion. From this frequency-domain criterion several adaptive algorithms have been derived for implementing the SDW-MWF. The main difference between these algorithms consists in the calculation of the step size matrix (constrained vs. unconstrained, block-structured vs. diagonal) used in the update formula for the multichannel adaptive filter. The computational complexity for all adaptive algorithms is quite similar, where the complexity for the unconstrained algorithms is smaller than the constrained algorithms and the complexity for the diagonal step size matrix is smaller than the block-structured step size matrix. Experimental results with a small-sized microphone array in a hearing aid show that the SDR-GSC and the SP-SDW-MWF are more robust against signal model errors than the GSC, both in stationary and in time-varying scenarios. The main factor affecting the convergence speed is $\rho(1-\lambda)$, and the block-structured step size matrix gives rise to a substantially faster convergence than the diagonal step size matrix, only at a slightly higher computational cost. Compared to a perfect VAD, an energy-based VAD generally gives rise to a degraded performance, especially at low input SNRs ($< 0\,\mathrm{dB}$), since at these SNRs an energy-based VAD generates more detection errors.

References

[1] B. D. Van Veen, K. M. Buckley, Beamforming: A Versatile Approach to Spatial Filtering, IEEE ASSP Magazine 5 (2) (1988) 4–24.

[2] O. L. Frost III, An Algorithm for Linearly Constrained Adaptive Array Processing, Proc. IEEE 60 (1972) 926–935.

[3] L. J. Griffiths, C. W. Jim, An alternative approach to linearly constrained adaptive beamforming, IEEE Trans. Antennas Propagat. 30 (1) (1982) 27–34.

[4] S. Nordholm, I. Claesson, B. Bengtsson, Adaptive Array Noise Suppression of Handsfree Speaker Input in Cars, IEEE Trans. Veh. Technol. 42 (4) (1993) 514–518.

[5] I. Claesson, S. Nordholm, A Spatial Filtering Approach to Robust Adaptive Beaming, IEEE Trans. Antennas Propagat. 40 (9) (1992) 1093–1096.

[6] S. Nordebo, I. Claesson, S. Nordholm, Adaptive beamforming: Spatial filter designed blocking matrix, IEEE Journal of Oceanic Engineering 19 (4) (1994) 583–590.

[7] S. Doclo, M. Moonen, Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics, IEEE Trans. Signal Processing 51 (10) (2003) 2511–2526.

[8] D. Van Compernolle, Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque NM, USA, 1990, pp. 833–836.

[9] O. Hoshuyama, A. Sugiyama, A. Hirano, A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters, IEEE Trans. Signal Processing 47 (10) (1999) 2677–2684.

[10] W. Herbordt, W. Kellermann, Adaptive Beamforming for Audio Signal Acquisition, Springer-Verlag, 2003, Ch. 6 in "Adaptive Signal Processing: Applications to Real-World Problems" (Benesty, J. and Huang, Y. , Eds.), pp. 155–194.

[11] S. Gannot, D. Burshtein, E. Weinstein, Signal Enhancement Using Beamforming and Non-Stationarity with Applications to Speech, IEEE Trans. Signal Processing 49 (8) (2001) 1614–1626.

[12] J. E. Greenberg, P. M. Zurek, Evaluation of an adaptive beamforming method for hearing aids, Journal of the Acoustical Society of America 91 (3) (1992) 1662–1676.

[13] J. Vanden Berghe, J. Wouters, An adaptive noise canceller for hearing aids using two nearby microphones, Journal of the Acoustical Society of America 103 (6) (1998) 3621–3626.

[14] W. Herbordt, H. Buchner, W. Kellermann, An acoustic human-machine front-end for multimedia applications, EURASIP Journal on Applied Signal Processing 2003 (1) (2003) 21–31.

[15] O. Hoshuyama, B. Begasse, A. Sugiyama, A new adaptation-mode control based on cross correlation for a robust adaptive microphone array, IEICE Trans. Fundamentals E84-A (2) (2001) 406–413.

[16] N. K. Jablon, Adaptive beamforming with the Generalized Sidelobe Canceller in the presence of array imperfections, IEEE Trans. Antennas Propagat. 34 (8) (1986) 996–1012.

[17] H. Cox, R. M. Zeskind, M. M. Owen, Robust Adaptive Beamforming, IEEE Trans. Acoust., Speech, Signal Processing 35 (10) (1987) 1365–1376.

[18] M. W. Hoffman, K. M. Buckley, Robust time-domain processing of broadband microphone array data, IEEE Trans. Speech and Audio Processing 3 (3) (1995) 193–203.

[19] A. Spriet, M. Moonen, J. Wouters, Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction, Signal Processing 84 (12) (2004) 2367–2387.

[20] A. Spriet, M. Moonen, J. Wouters, Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids, IEEE Trans. Signal Processing 53 (3) (2005) 911–925.

[21] S. Doclo, A. Spriet, M. Moonen, Efficient frequency-domain implementation of speech distortion weighted multi-channel Wiener filtering for noise reduction, in: Proc. European Signal Processing Conf. (EUSIPCO), Vienna, Austria, 2004, pp. 2007–2010.

[22] S. Doclo, M. Moonen, GSVD-Based Optimal Filtering for Multi-Microphone Speech Enhancement, Springer-Verlag, 2001, Ch. 6 in "Microphone Arrays: Signal Processing Techniques and Applications" (Brandstein, M. S. and Ward, D. B., Eds.), pp. 111–132.

[23] S. Doclo, M. Moonen, GSVD-based optimal filtering for single and multimicrophone speech enhancement, IEEE Trans. Signal Processing 50 (9) (2002) 2230–2244.

[24] G. Rombouts, M. Moonen, QRD-based unconstrained optimal filtering for acoustic noise reduction, Signal Processing 83 (9) (2003) 1889–1904.

[25] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, General Derivation of Frequency-Domain Adaptive Filtering, Springer-Verlag, 2001, Ch. 8 in "Advances in Network and Acoustic Echo Cancellation", pp. 157–176.

[26] H. Buchner, J. Benesty, W. Kellermann, Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication, Signal Processing 85 (3) (2005) 549–570.

[27] S. Van Gerven, F. Xie, A Comparative Study of Speech Detection Methods, in: Proc. EUROSPEECH, Vol. 3, Rhodos, Greece, 1997, pp. 1095–1098.

[28] J. Sohn, N. S. Kim, W. Sung, A Statistical Model-Based Voice Activity Detection, IEEE Signal Processing Lett. 6 (1) (1999) 1–3.

[29] W. Herbordt, T. Trini, W. Kellermann, Robust spatial estimation of the signal-to-interference ratio for non-stationary mixtures, in: Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC), Kyoto, Japan, 2003, pp. 247–250.

[30] Y. Ephraim, H. L. Van Trees, A Signal Subspace Approach for Speech Enhancement, IEEE Trans. Speech and Audio Processing 3 (4) (1995) 251–266.

[31] J. J. Shynk, Frequency-Domain and Multirate Adaptive Filtering, IEEE Signal Processing Magazine 9 (1) (1992) 14–37.

[32] A. Spriet, M. Moonen, J. Wouters, Robustness Analysis of Multi-channel Wiener Filtering and Generalized Sidelobe Cancellation for Multi-microphone Noise Reduction in Hearing Aid Applications, IEEE Trans. Speech and Audio Processing 13 (4) (2005) 487–503.

[33] M. J. Link, K. M. Buckley, Prewhitening for intelligibility gain in hearing aids arrays, Journal of the Acoustical Society of America 93 (4) (1993) 2139–2140.

[34] J. E. Greenberg, P. M. Peterson, P. M. Zurek, Intelligibility-weighted measures of speech-to-interference ratio and speech system performance, Journal of the Acoustical Society of America 94 (5) (1993) 3009–3010.

[35] Acoustical Society of America, ANSI S3.5-1997 American National Standard Methods for Calculation of the Speech Intelligibility Index (Jun. 1997).

[36] L. B. Jensen, Hearing aid with adaptive matching of input transducers, US Patent No. 6,741,714, 2004.

[37] M. Nilsson, S. D. Soli, A. Sullivan, Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise, Journal of the Acoustical Society of America 95 (2) (1994) 1085–1099.

[38] A. Spriet, M. Moonen, J. Wouters, The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and Generalized Sidelobe Cancellation, Signal Processing 85 (6) (2005) 1073–1088.

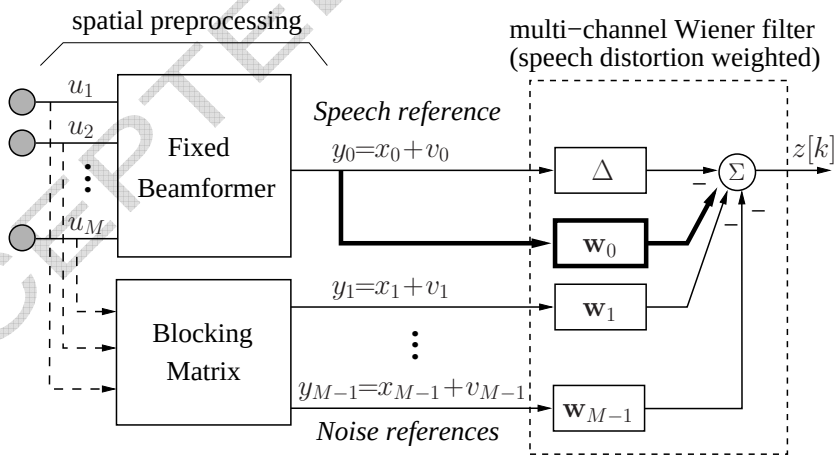


Fig. 1. Structure of the spatially pre-processed speech distortion weighted multi-channel Wiener filter (SP-SDW-MWF).

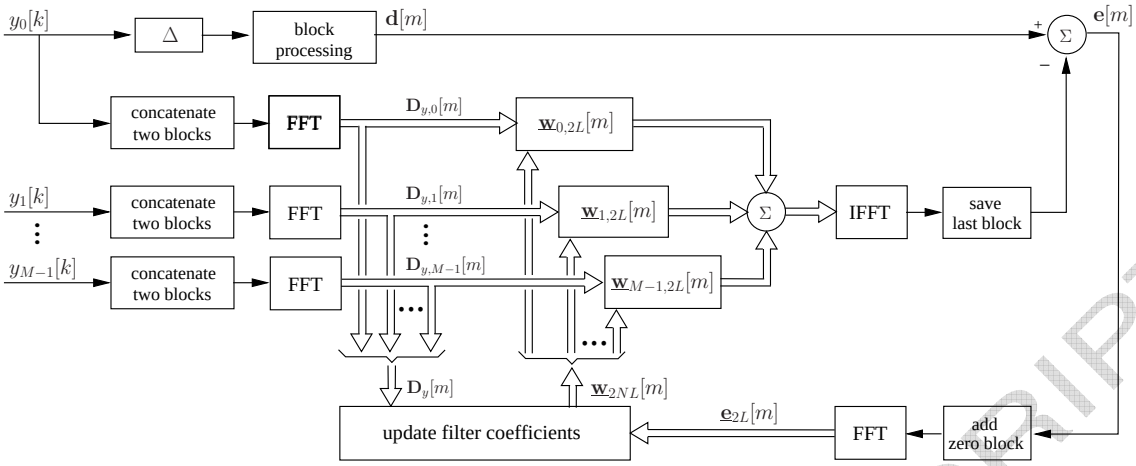


Fig. 2. General block diagram of the frequency-domain implementation of the SDW-MWF (all algorithms)

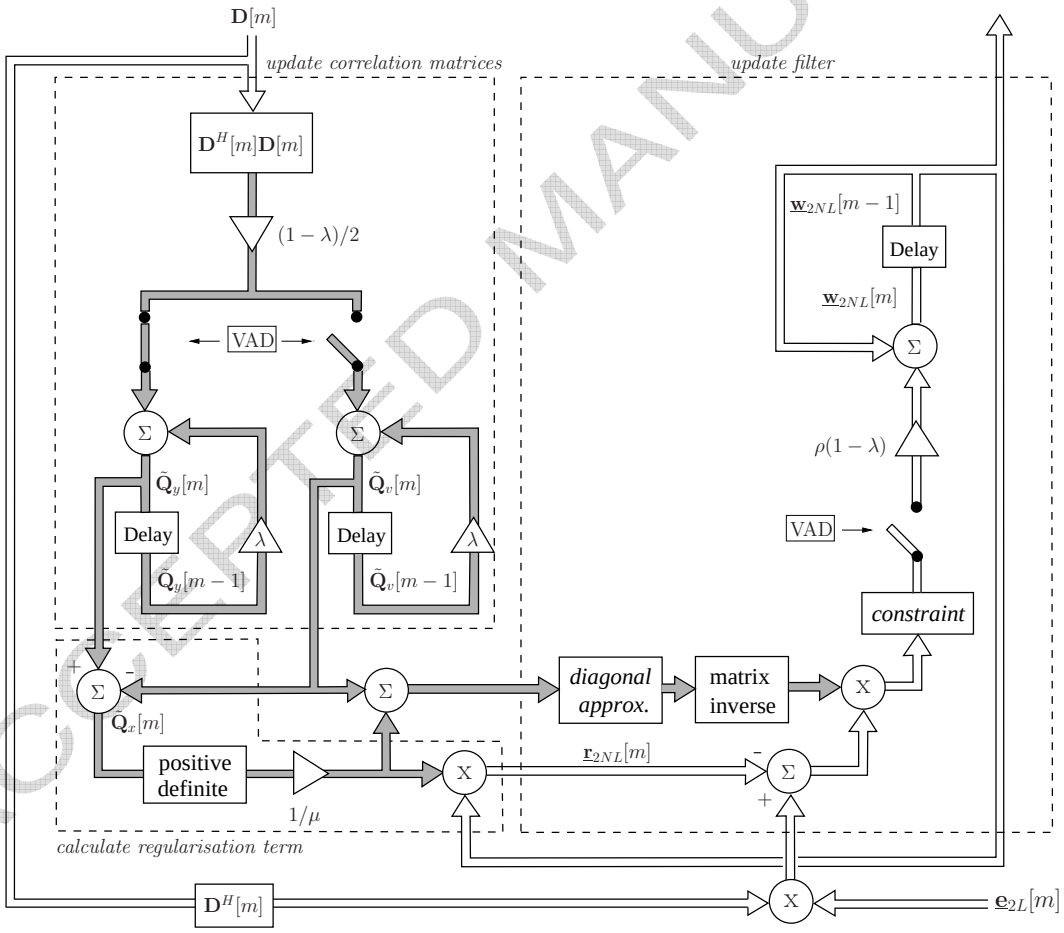


Fig. 3. Block diagram of the filter update procedure using block-structured and diagonal correlation matrices

Matrix definitions:

$\mathbf{F}_{2L} = 2L \times 2L$-dimensional DFT matrix

$\mathbf{0}_L = L \times L$-dimensional zero matrix, $\quad \mathbf{I}_L = L \times L$-dimensional identity matrix

$$\mathbf{G}^{01}_{2L\times 2L} = \mathbf{F}_{2L} \begin{bmatrix} \mathbf{0}_L & \mathbf{0}_L \\ \mathbf{0}_L & \mathbf{I}_L \end{bmatrix} \mathbf{F}^{-1}_{2L}, \quad \mathbf{G}^{10}_{2L\times 2L} = \mathbf{F}_{2L} \begin{bmatrix} \mathbf{I}_L & \mathbf{0}_L \\ \mathbf{0}_L & \mathbf{0}_L \end{bmatrix} \mathbf{F}^{-1}_{2L}$$

$$\mathbf{G}^{10}_{2NL\times 2NL} = \mathrm{diag}\left[\mathbf{G}^{10}_{2L\times 2L} \ \dots \ \mathbf{G}^{10}_{2L\times 2L} \right]$$

For each new block of L samples:

$$\mathbf{d}[m] = \left[y_0[mL-\Delta] \ y_0[mL-\Delta+1] \ \dots \ y_0[mL-\Delta+L-1] \right]^T$$

$$\mathbf{D}_{y,n}[m] = \mathrm{diag}\left\{ \mathbf{F}_{2L} \left[y_n[mL-L] \ \dots \ y_n[mL+L-1] \right]^T \right\}, \quad n = M-N, \dots, M-1$$

$$\mathbf{D}_y[m] = \left[\mathbf{D}_{y,M-N}[m] \ \dots \ \mathbf{D}_{y,M-1}[m] \right]$$

Output signal:

$$\mathbf{e}[m] = \mathbf{d}[m] - \left[\mathbf{0}_L \ \mathbf{I}_L \right] \mathbf{F}^{-1}_{2L} \, \mathbf{D}_y[m] \, \underline{\mathbf{w}}_{2NL}[m-1]$$

If speech detected:

$$\mathbf{Q}_y[m] = \lambda \mathbf{Q}_y[m-1] + (1-\lambda) \, \mathbf{D}^H_y[m] \, \mathbf{G}^{01}_{2L\times 2L} \mathbf{D}_y[m], \quad \mathbf{Q}_v[m] = \mathbf{Q}_v[m-1]$$

$$\underline{\mathbf{w}}_{2NL}[m] = \underline{\mathbf{w}}_{2NL}[m-1]$$

If noise detected: $\mathbf{D}_v[m] = \mathbf{D}_y[m]$

$$\mathbf{Q}_v[m] = \lambda \mathbf{Q}_v[m-1] + (1-\lambda) \, \mathbf{D}^H_v[m] \, \mathbf{G}^{01}_{2L\times 2L} \mathbf{D}_v[m], \quad \mathbf{Q}_y[m] = \mathbf{Q}_y[m-1]$$

$$\mathbf{Q}_x[m] = \mathbf{Q}_y[m] - \mathbf{Q}_v[m]$$

$$\underline{\mathbf{r}}_{2NL}[m] = \tfrac{1}{\mu} \mathbf{Q}_x[m] \, \underline{\mathbf{w}}_{2NL}[m-1]$$

$$\underline{\mathbf{e}}_{v,2L}[m] = \mathbf{F}_{2L} \begin{bmatrix} \mathbf{0}_L \\ \mathbf{I}_L \end{bmatrix} \mathbf{e}[m]$$

$$\underline{\mathbf{w}}_{2NL}[m] = \underline{\mathbf{w}}_{2NL}[m-1] + (1-\lambda) \, \mathbf{G}^{10}_{2NL\times 2NL} \left[\mathbf{Q}_v[m] + \tfrac{1}{\mu} \mathbf{Q}_x[m] \right]^{-1} \cdot \left\{ \mathbf{D}^H_v[m] \, \underline{\mathbf{e}}_{v,2L}[m] - \underline{\mathbf{r}}_{2NL}[m] \right\}$$

Table 1
Algorithmic description of recursive frequency-domain implementation of SDW-MWF

Algorithm	**Step size matrix**
Algo 1 - constr (64)	$\mathbf{G}_{2NL\times 2NL}^{10}\left[\tilde{\mathbf{Q}}_v[m]+\frac{1}{\mu}\tilde{\mathbf{Q}}_x[m]\right]^{-1}$
Algo 1 - unconstr	$\frac{1}{2}\left[\tilde{\mathbf{Q}}_v[m]+\frac{1}{\mu}\tilde{\mathbf{Q}}_x[m]\right]^{-1}$
Algo 2 - constr (75)	$\mathbf{G}_{2NL\times 2NL}^{10}\text{diag}\{\left[\tilde{\mathbf{Q}}_{v,nn}[m]+\frac{1}{\mu}\tilde{\mathbf{Q}}_{x,nn}[m]\right]^{-1}\}$
Algo 2 - unconstr	$\frac{1}{2}\,\text{diag}\{\left[\tilde{\mathbf{Q}}_{v,nn}[m]+\frac{1}{\mu}\tilde{\mathbf{Q}}_{x,nn}[m]\right]^{-1}\}$
Algo 3 - constr (76)	$\mathbf{G}_{2NL\times 2NL}^{10}\text{diag}\{\left[(1/N)\sum_{n=M-N}^{M-1}\tilde{\mathbf{Q}}_{v,nn}[m]+\frac{1}{\mu}\tilde{\mathbf{Q}}_{x,nn}[m]\right]^{-1}\}$
Algo 3 - unconstr	$\frac{1}{2}\,\text{diag}\{\left[(1/N)\sum_{n=M-N}^{M-1}\tilde{\mathbf{Q}}_{v,nn}[m]+\frac{1}{\mu}\tilde{\mathbf{Q}}_{x,nn}[m]\right]^{-1}\}$

Table 2
Step size matrix $\mathbf{\Lambda}[m]$ for different adaptive frequency-domain algorithms

Algorithm	**Computational complexity**	10^6 MAC
QIC-GSC-SPA (constr) [17]	$(3M-1)\text{FFT}+16M-9$	2.67
SDW-MWF (buffer - constr) [20]	$(3N+5)\text{FFT}+30N+10$	$3.94^{(a)}$, $5.18^{(b)}$
SDW-MWF (Algo 1 - constr, N=2)	$(3N+2)\text{FFT}+14N^2+10N+12$	$3.46^{(a)}$
SDW-MWF (Algo 1 - unconstr, N=2)	$(N+2)\text{FFT}+14N^2+12N+12$	$2.50^{(a)}$
SDW-MWF (Algo 2 - constr)	$(3N+2)\text{FFT}+8N^2+13N$	$2.98^{(a)}$, $4.59^{(b)}$
SDW-MWF (Algo 2 - unconstr)	$(N+2)\text{FFT}+8N^2+15N$	$2.02^{(a)}$, $3.15^{(b)}$
SDW-MWF (Algo 3 - constr)	$(3N+2)\text{FFT}+8N^2+12N$	$2.94^{(a)}$, $4.54^{(b)}$
SDW-MWF (Algo 3 - unconstr)	$(N+2)\text{FFT}+8N^2+14N$	$1.98^{(a)}$, $3.10^{(b)}$

Table 3
Computational complexity for frequency-domain adaptive algorithms ($M=3$, $L=128$, $f_s=16\,\text{kHz}$, (a) $N=M-1$, (b) $N=M$)

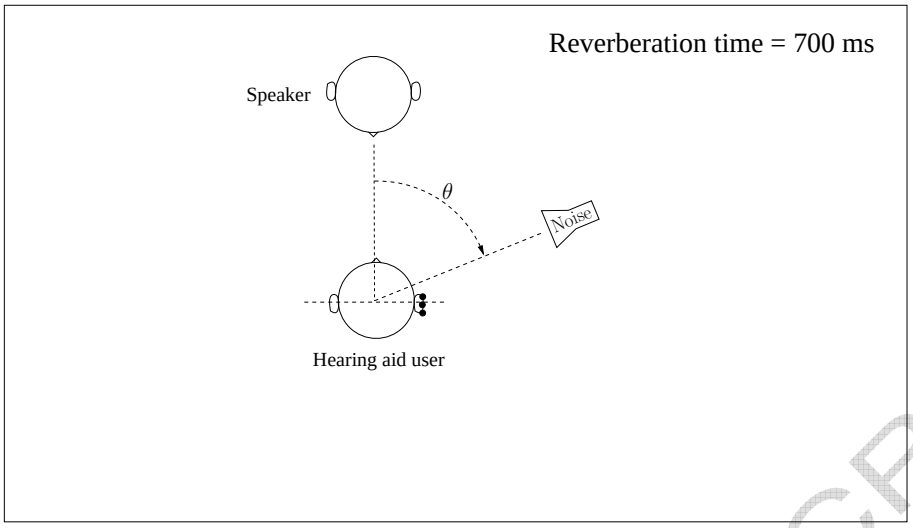


Fig. 4. Recording environment consisting of a speech source and one or more noise sources

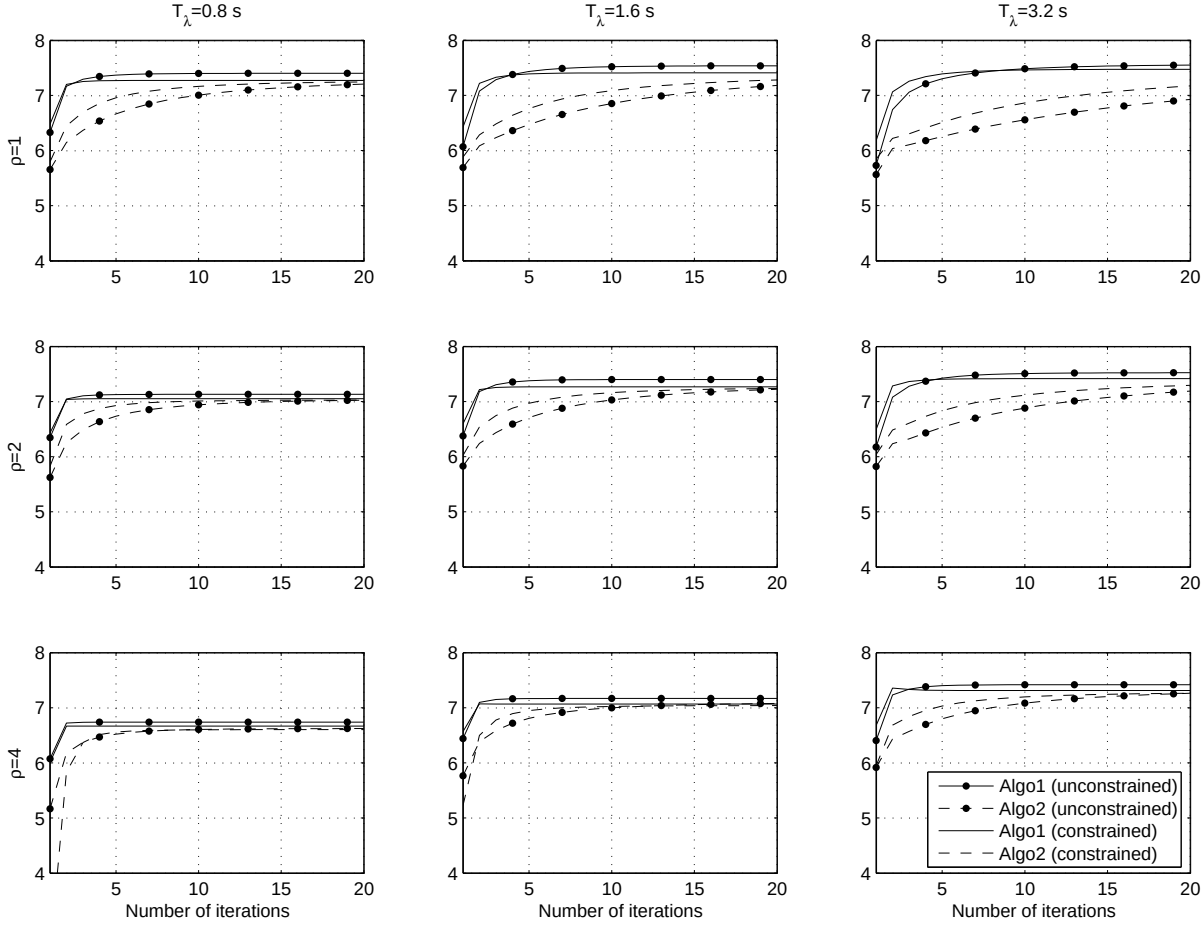


Fig. 5. Effect of the step size parameter ρ and the exponential forgetting factor λ on the convergence of the SNR improvement for different adaptive algorithms (SDR-GSC, $1/\mu = 0.5$, no microphone mismatch, energy-based VAD).

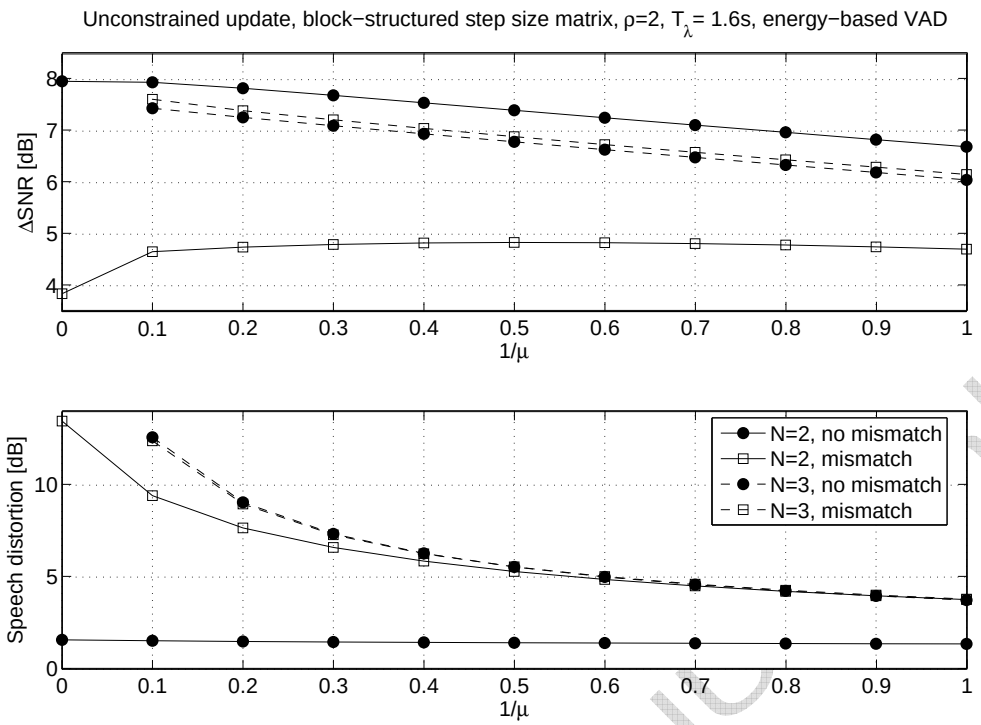


Fig. 6. SNR improvement and speech distortion of SDR-GSC ($N = 2$) and SP-SDW-MWF ($N = 3$) as a function of $1/\mu$, with and without microphone mismatch (unconstrained update, block-structured step size matrix, $\rho = 2$, $T_\lambda = 1.6\,\mathrm{s}$, energy-based VAD).

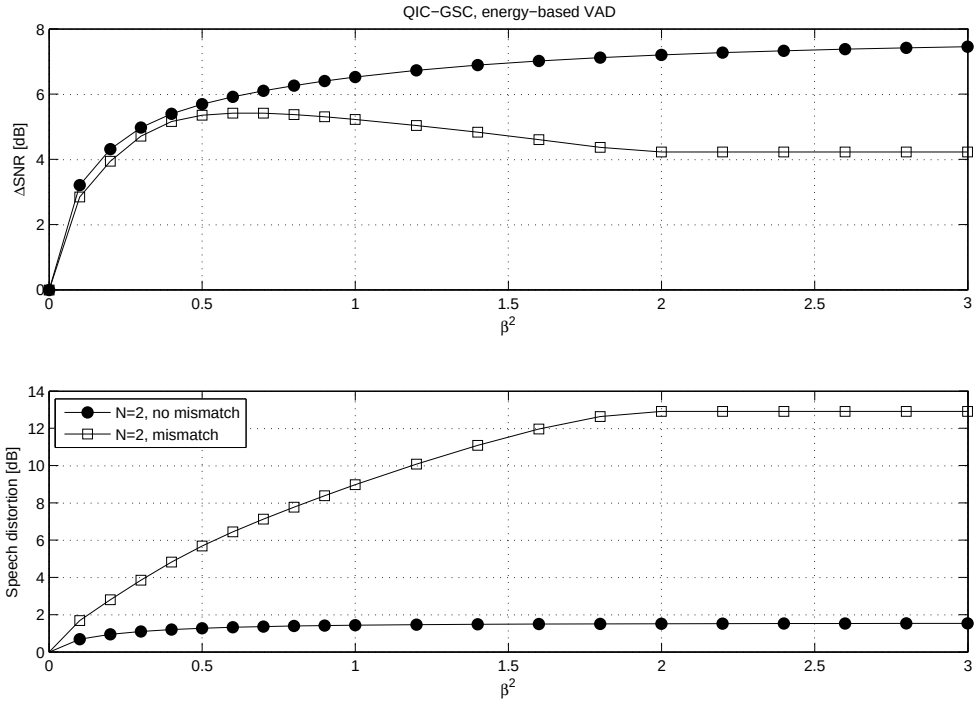


Fig. 7. SNR improvement and speech distortion of QIC-GSC as a function of β^2, with and without microphone mismatch.

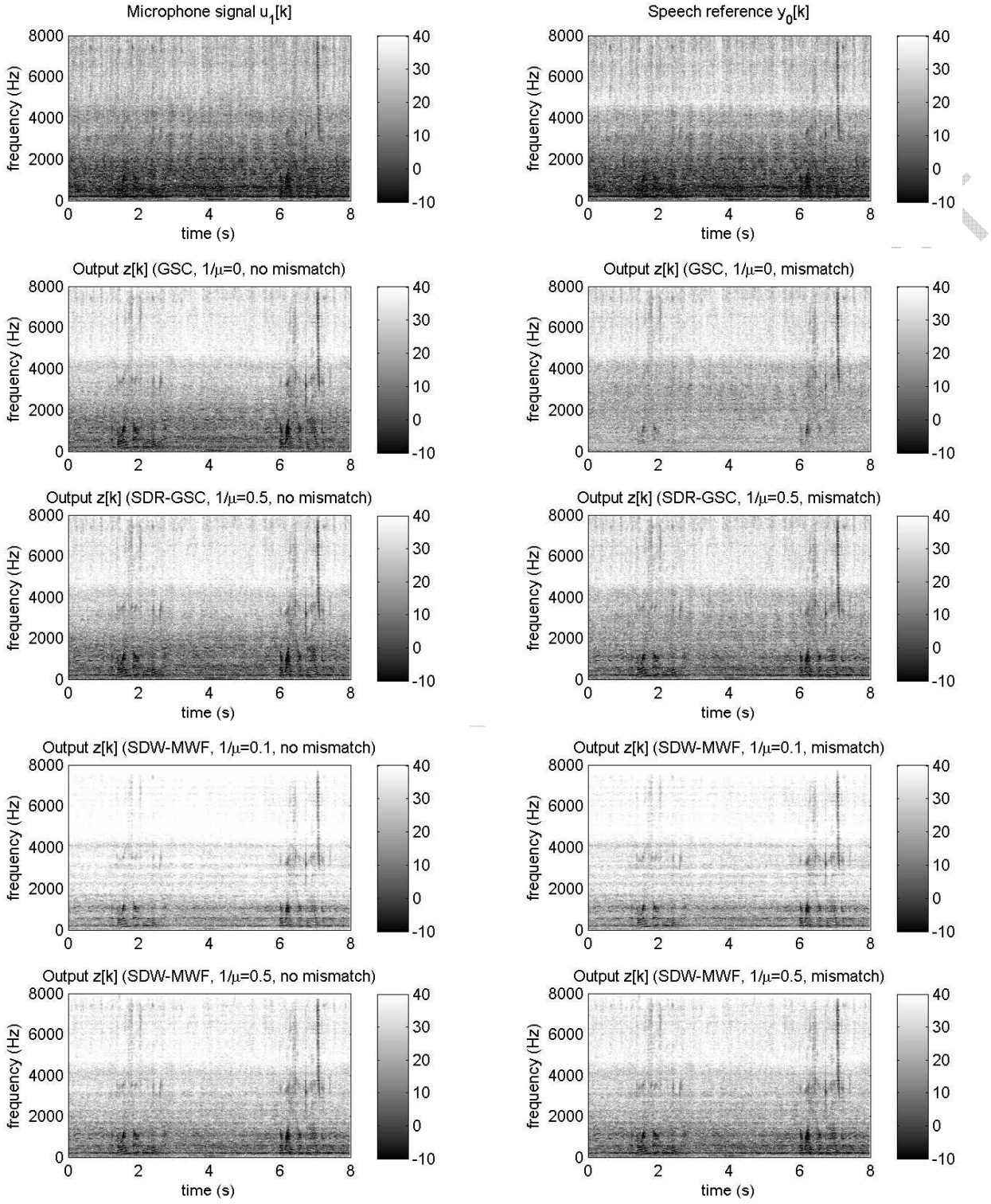


Fig. 8. Spectrogram of the microphone signal $u_1[k]$, the speech reference signal $y_0[k]$, and the output signal $z[k]$ for GSC ($1/\mu = 0$), SDR-GSC ($1/\mu = 0.5$) and SP-SDW-MWF ($1/\mu = 0.1, 0.5$), with and without mismatch (unconstrained update, block-structured step size matrix, $\rho = 2$, $T_\lambda = 1.6\,\mathrm{s}$, energy-based VAD).

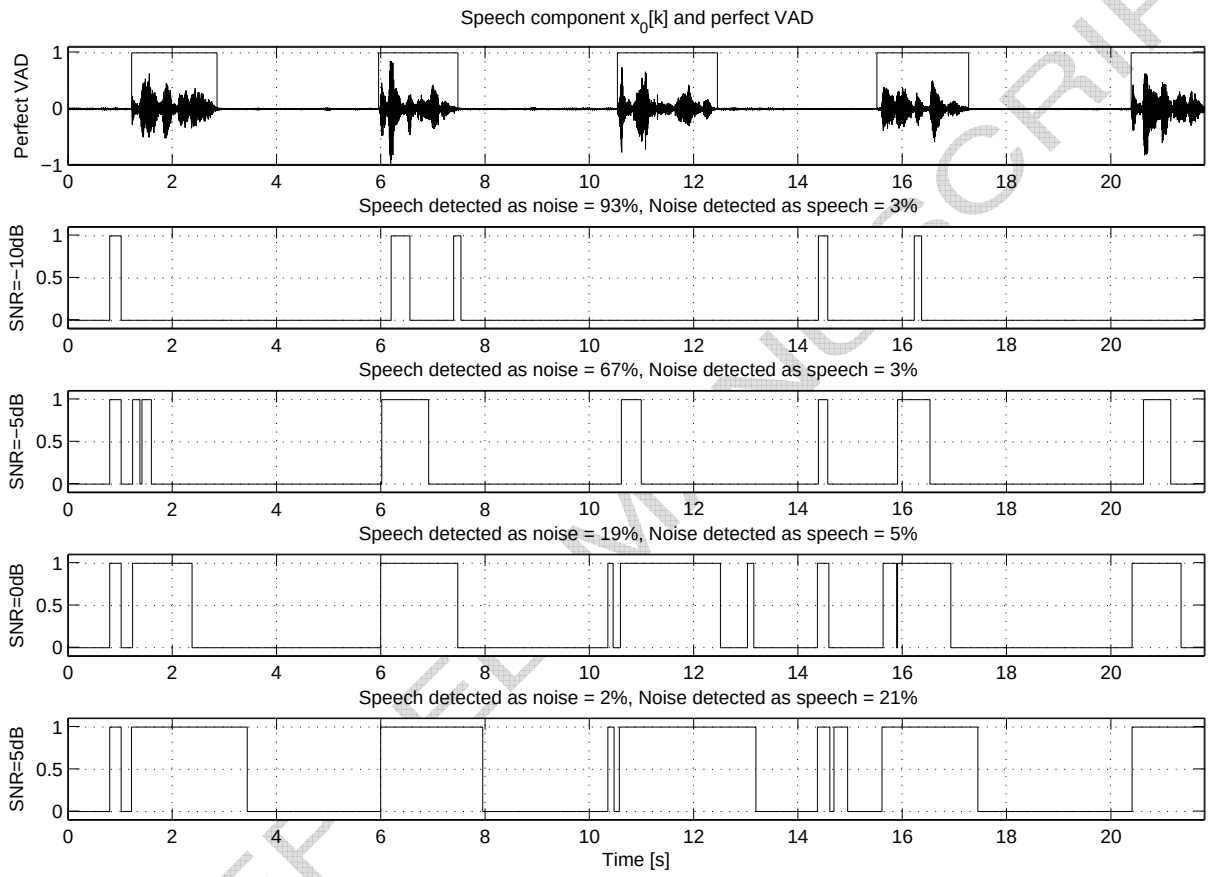


Fig. 9. VAD performance for different input SNRs, ranging from −10 dB to 5 dB. For each input SNR the percentage of speech frames classified as noise and noise frames classified as speech is indicated.

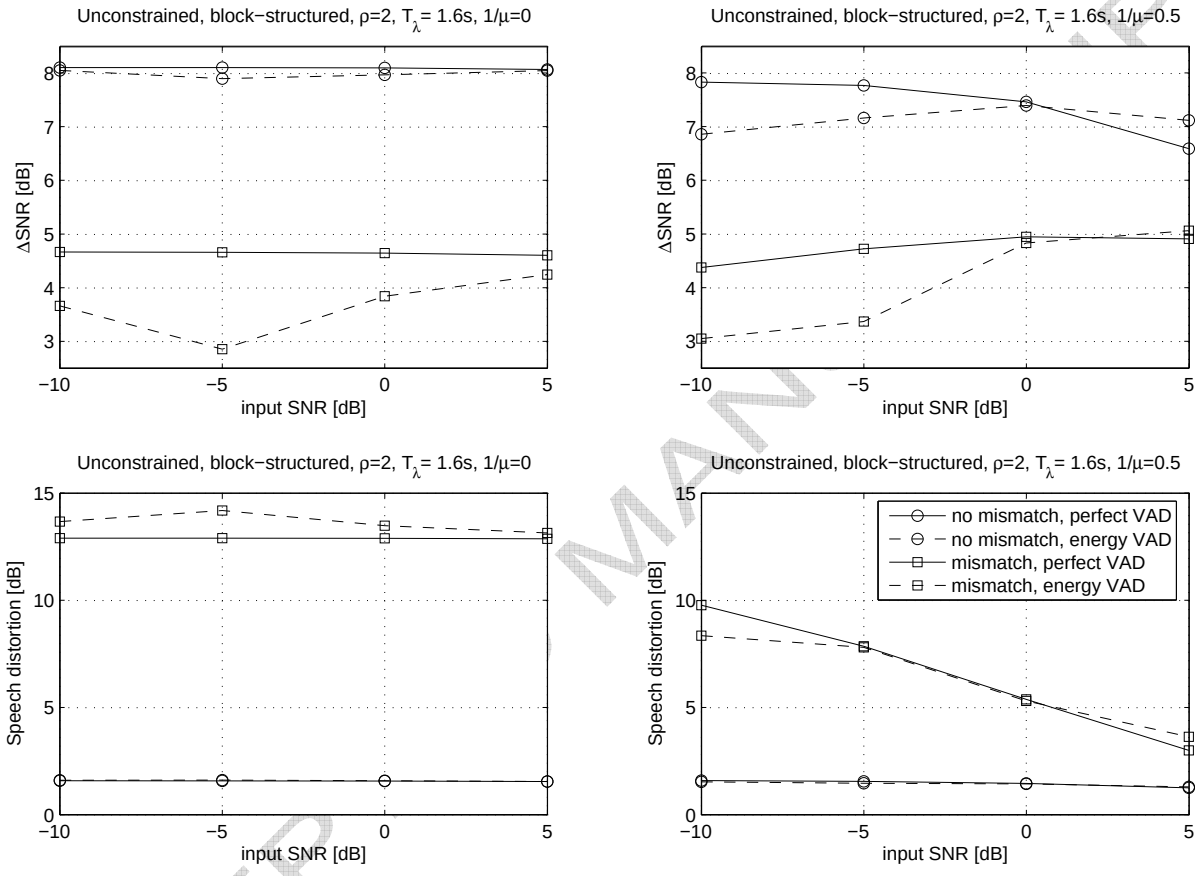


Fig. 10. Effect of energy-based VAD on SNR improvement and speech distortion for GSC ($1/\mu = 0$) and SDR-GSC ($1/\mu = 0.5$) for different input SNRs, with and without microphone mismatch (unconstrained update, block-structured step size matrix, $\rho = 2$, $T_\lambda = 1.6\,\mathrm{s}$).

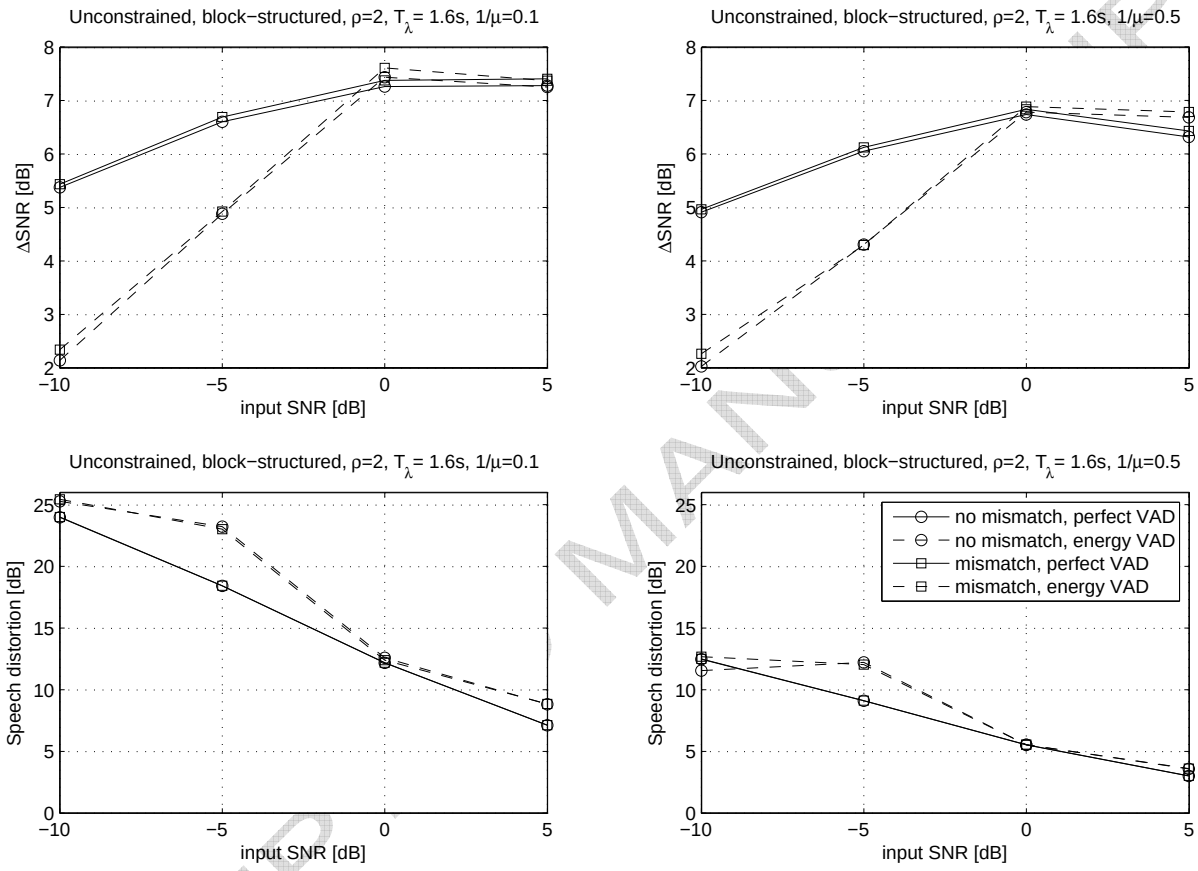


Fig. 11. Effect of energy-based VAD on SNR improvement and speech distortion for SP-SDW-MWF ($1/\mu = 0.1, 0.5$) for different input SNRs, with and without microphone mismatch (unconstrained update, block-structured step size matrix, $\rho = 2$, $T_\lambda = 1.6\,\mathrm{s}$).

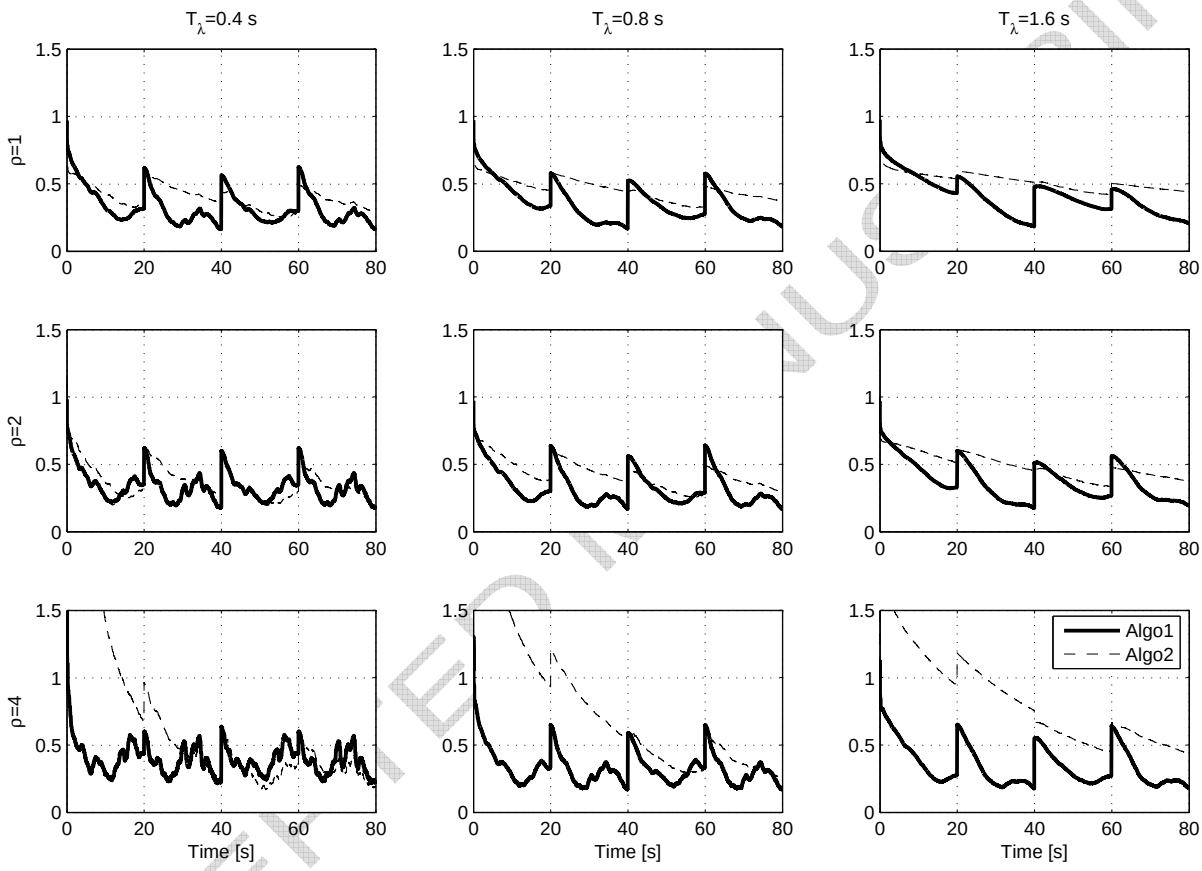


Fig. 12. Filter convergence $\Delta\mathbf{w}[m]$ of SDR-GSC for a switching speech scenario (unconstrained update, $1/\mu = 0.5$, mismatch, perfect VAD).

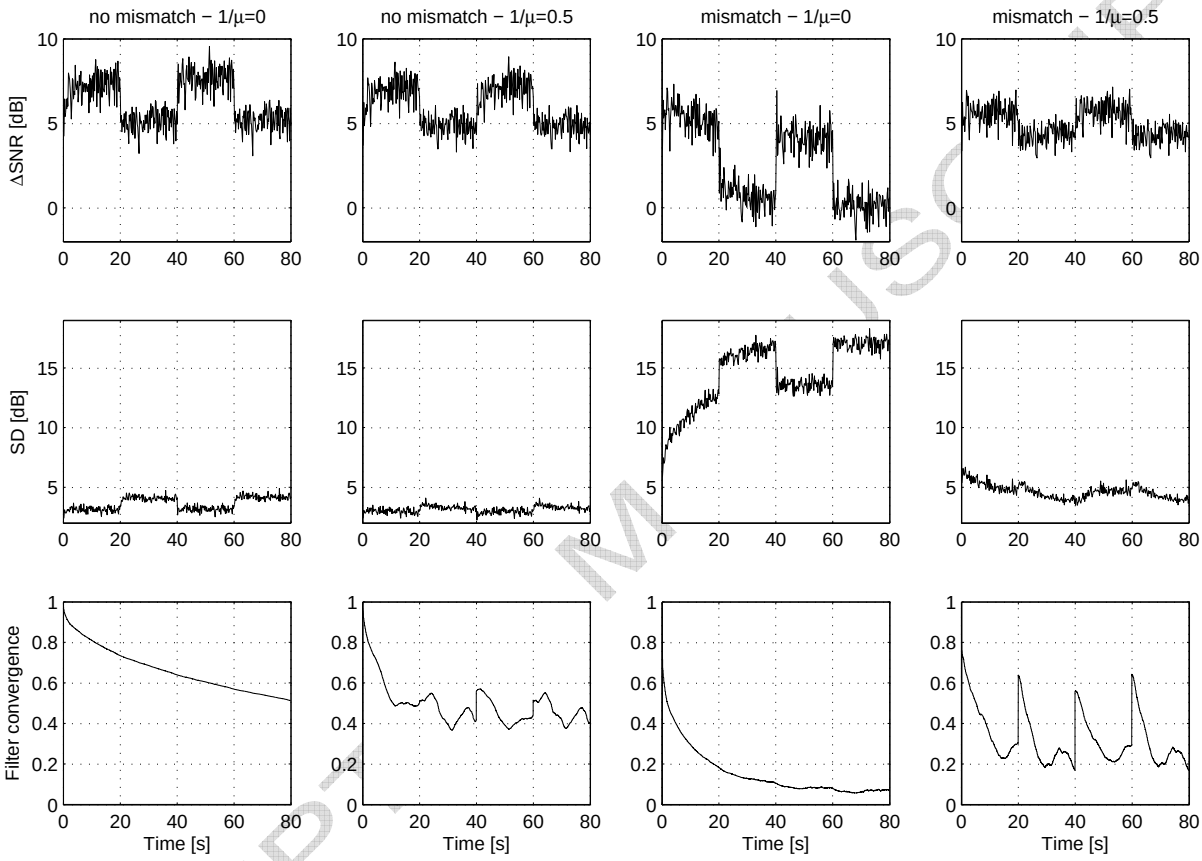


Fig. 13. SNR improvement, speech distortion and filter convergence of GSC ($1/\mu = 0$) and SDR-GSC ($1/\mu = 0.5$) for a switching speech scenario, with and without microphone mismatch (unconstrained update, block-structured step size matrix, $\rho = 2$, $T_\lambda = 0.8\,\mathrm{s}$, perfect VAD).

Frequency-Domain Criterion for the Speech Distortion Weighted Multichannel Wiener Filter for Robust Noise Reduction

Simon Doclo *, Ann Spriet, Jan Wouters, Marc Moonen

Katholieke Universiteit Leuven, Dept. of Electrical Engineering (ESAT - SCD), Kasteelpark Arenberg 10, 3001 Heverlee (Leuven), Belgium

Abstract

Recently, a generalized multi-microphone noise reduction scheme, referred to as the spatially pre-processed speech distortion weighted multichannel Wiener filter (SP-SDW-MWF), has been presented. This scheme consists of a fixed spatial pre-processor and a multichannel adaptive noise canceler (ANC) optimizing the SDW-MWF cost function. By taking speech distortion explicitly into account in the design criterion of the multichannel ANC, the SP-SDW-MWF adds robustness to the standard generalized sidelobe canceler (GSC). In this paper, we present a multichannel frequency-domain criterion for the SDW-MWF, from which several – existing and novel – adaptive frequency-domain algorithms can be derived. The main difference between these adaptive algorithms consists in the calculation of the step size matrix (constrained vs. unconstrained, block-structured vs. diagonal) used in the update formula for the multichannel adaptive filter. We investigate the noise reduction performance, the robustness and the tracking performance of these adaptive algorithms, using a perfect voice activity detection (VAD) mechanism and using an energy-based VAD. Using experimental results with a small-sized microphone array in a hearing aid, it is shown that the SP-SDW-MWF is more robust against signal model errors than the GSC, and that the block-structured step size matrix gives rise to a faster convergence and a better tracking performance than the diagonal step size matrix, only at a slightly higher computational cost.

Key words: multi-microphone noise reduction, adaptive frequency-domain algorithms, multichannel Wiener filter, generalized sidelobe canceler, hearing aids
PACS: 43.60.Fg, 43.60.Mn, 43.72.-p, 43.60.Dh

* Corresponding author. Tel: +32/16/321899, Fax: +32/16/321970
Email address: simon.doclo@esat.kuleuven.be (Simon Doclo).