Katholieke Universiteit Leuven

Departement Elektrotechniek ESAT-SISTA/TR 2003-47

Stochastic Gradient based Implementation of Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener Filtering for Noise Reduction in Hearing Aids [1]

Ann Spriet[2], Marc Moonen[3] ,Jan Wouters[4]

IEEE Transactions on Signal Processing, vol. 53, no. 3, March 2005

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/spriet/reports/03-47.pdf*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium, Tel. 32/16/32 18 99, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista*. E-mail: *ann.spriet@esat.kuleuven.ac.be*. K.U.Leuven, Lab. Exp. ORL/ ENT-Dept., Kapucijnenvoer 33, 3000 Leuven, Belgium, Tel. 32/16/33 24 15, Fax 32/16/33 23 35, WWW: *http://www.kuleuven.ac.be/exporl/Lab/Default.htm*. Ann Spriet is a Research Assistant supported by the Fonds voor Wetenschappelijk Onderzoek (FWO) - Vlaanderen. This research work was carried out at the ESAT laboratory and Lab. Exp. ORL of the Katholieke Universiteit Leuven, in the the frame of the Belgian State, Prime Minister's Office - Federal Office for Scientific, Technical and Cultural Affairs - Interuniversity Poles of Attraction Programme (2002-2007) - IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling'), the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) of the Flemish Government, Research Project FWO nr.G.0233.01 ('Signal processing and automatic patient fitting for advanced auditory prostheses'), IWT project 020540: 'Innovative Speech Processing Algorithms for Improved Performance of Cochlear Implants' and was partially sponsored by Cochlear. The scientific responsibility is assumed by its authors.

[3]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium, Tel. 32/16/32 17 09, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista*. E-mail: *marc.moonen@esat.kuleuven.ac.be*. Marc Moonen is a professor at the Katholieke Universiteit Leuven.

[4]K.U.Leuven, Lab. Exp. ORL, Dept. Neurowetenschappen, Kapucijnenvoer 33, 3000 Leuven, Belgium, Tel. 32/16/33 23 42, Fax 32/16/33 23 35, WWW: *http://www.kuleuven.ac.be/exporl/Lab/Default.htm* E-mail: *jan.wouters@uz.kuleuven.ac.be*. Jan Wouters is a professor at the Katholieke Universiteit Leuven.

Stochastic Gradient based Implementation of Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener Filtering for Noise Reduction in Hearing Aids

Ann Spriet*[1,2], *Student Member, IEEE,* Marc Moonen[1], *Member, IEEE,* Jan Wouters[2]

[1]ESAT/SCD-SISTA, K.U. Leuven

Kasteelpark Arenberg 10

B-3001 Leuven-Heverlee, Belgium

Tel.: +32 16 321795, Fax: +32 16 321970

Email: ann.spriet@esat.kuleuven.ac.be; marc.moonen@esat.kuleuven.ac.be

[2]Lab. Exp. ORL / ENT-Dept., K.U. Leuven

Kapucijnenvoer 33, B-3000 Leuven, Belgium

Fax: +32 16 332335; Email: jan.wouters@uz.kuleuven.ac.be

Manuscript received August 26, 2003; revised March 23, 2004. From 1999 until 2003, Ann Spriet was supported by FWO-Vlaanderen. This research work was carried out at the ESAT laboratory and Lab. Exp. ORL of the K. U. Leuven, in the frame of the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling'), the Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) of the Flemish Government, Research Project FWO nr. G.0233.01 ('Signal processing and automatic patient fitting for advanced auditory prostheses') and IWT project 020540 ('Innovative speech processing algorithms for improved performance of cochlear implants'). The scientific responsibility is assumed by its authors. The associate editor coordinating the review of this paper and approving it for publication was Prof. Yuri Abramovich.

A. Spriet and M. Moonen are with the Katholieke Universiteit Leuven, Department of Electrical Engineering (ESAT-SCD), B-3001 Leuven, Belgium (e-mail: ann.spriet@esat.kuleuven.ac.be, marc.moonen@esat.kuleuven.ac.be).

A. Spriet and J. Wouters are with the Katholieke Universiteit Leuven, ENT-Department (Lab. Exp. ORL), B-3000 Leuven, Belgium (e-mail: jan.wouters@uz.kuleuven.ac.be).

Abstract

Recently, a generalized noise reduction scheme has been proposed, called the Spatially Pre-processed, Speech Distortion Weighted, Multi-channel Wiener Filter (SP-SDW-MWF). It encompasses the Generalized Sidelobe Canceller (GSC) and a multi-channel Wiener filtering technique as extreme cases. Compared to the widely studied GSC with Quadratic Inequality Constraint (QIC-GSC), the SP-SDW-MWF achieves a better noise reduction performance, for a given maximum speech distortion level. In this paper, we develop a low-cost, stochastic gradient implementation of the SP-SDW-MWF. To speed up convergence and reduce computational complexity, the algorithm is implemented in the frequency-domain. Experimental results with a behind-the-ear hearing aid show that the proposed frequency-domain stochastic gradient algorithm preserves the benefit of the exact SP-SDW-MWF over the QIC-GSC, while its computational cost is comparable to the least mean square based scaled projection algorithm for QIC-GSC.

Index Terms

Speech enhancement, robust adaptive beamforming, stochastic gradient based algorithms, generalized sidelobe canceller, multi-channel Wiener filtering, hearing aids

I. INTRODUCTION

Noise reduction algorithms are crucial for hearing impaired people to improve speech intelligibility in background noise and/or reverberation. Multi-microphone systems exploit spatial in addition to temporal and spectral information of the desired and noise signal and are thus preferred to single microphone procedures (such as spectral subtraction). For small-sized arrays such as hearing aids, multi-microphone noise reduction goes together with an increased sensitivity to errors in the assumed signal model such as microphone mismatch, reverberation, etc. [1], [2]

A widely studied, adaptive noise reduction algorithm for hearing aid applications is the *Generalized Sidelobe Canceller* (GSC) [2]-[10]. The GSC consists of a fixed spatial pre-processor, which includes a fixed beamformer and a blocking matrix, and an Adaptive Noise Canceller (ANC). To reduce speech distortion, the ANC is adapted during periods of noise only [5], [9], [11], [12]. The ANC then minimizes the output noise power while the blocking matrix is designed to avoid so-called speech leakage into the noise references. The GSC assumes the desired speaker location, the microphone characteristics and positions to be known, and reflections of the speech signal to be absent. If these assumptions are satisfied, it provides an undistorted enhanced speech signal in which the amount of residual noise is minimized.

However, in reality these assumptions are often violated, resulting in speech leakage and hence speech distortion. When used in combination with small-sized arrays, e.g., in hearing aids, a robustness constraint is required to guarantee performance in the presence of errors in the assumed signal model, e.g., due to microphone mismatch [13], [14], even when the ANC is adapted during noise only. A widely applied method consists of imposing a quadratic inequality constraint to the ANC (QIC-GSC) [8], [9], [12], [15], [16], [17]. For Least Mean Square (LMS) based updating, the Scaled Projection Algorithm (SPA) [15] is a simple and effective technique that imposes this constraint. However, the QIC goes at the expense of less noise reduction [13], [14].

Recently, a *Multi-channel Wiener Filtering* (MWF) technique has been proposed that provides a Minimum Mean Square Error (MMSE) estimate of the speech component in one of the received microphone signals [18]-[23]. In contrast to the ANC of the GSC, the MWF takes speech distortion into account in its optimization criterion. The MMSE optimization criterion can also be generalized to allow for a trade-off between speech distortion and noise reduction [19]. We will refer to this generalization as Speech Distortion Weighted MWF (SDW-MWF). The SDW-MWF technique is uniquely based on estimates of the second order statistics of the recorded speech signal and the noise signal. A robust speech detection is thus (again) needed. In contrast to the GSC, the MWF does not make any a priori assumptions about the signal model so that no or a less severe robustness constraint is needed to guarantee performance when used in combination with small-sized arrays [13], [14]. Especially in complex noise scenarios such as multiple noise sources or diffuse noise, the MWF filter outperforms the GSC, even when the GSC is supplied with a robustness constraint [13], [14].

In [24], the GSC and the (SDW-)MWF have been integrated into one generalized scheme, called the *Spatially Pre-processed Speech Distortion Weighted Multi-channel Wiener Filter* (SP-SDW-MWF), which consists of a fixed spatial pre-processor and an adaptive SDW-MWF stage. The new scheme encompasses the GSC and (SDW-)MWF as extreme cases and allows for in-between solutions that can be interpreted as a Speech Distortion Regularized GSC (SDR-GSC). The SDR-GSC or more general the SP-SDW-MWF adds robustness to the GSC by taking speech distortion explicitly into account in the design criterion of the adaptive stage. Compared to the widely studied QIC-GSC, the SP-SDW-MWF achieves a better noise reduction performance, for a given maximum speech distortion level.

In [20], [21] recursive implementations of the SDW-MWF have been proposed based on a Generalized Singular Value Decomposition (GSVD) or an QR decomposition. A subband implementation [22], [23] results in improved speech intelligibility at a significantly lower cost compared to a fullband approach. Similar techniques can be applied to implement the SP-SDW-MWF [24]. However, in contrast to the

GSC and the QIC-GSC [15], no cheap stochastic gradient based implementation of the SP-SDW-MWF is available yet. In [25], an LMS based algorithm for the MWF has been developed. The algorithm requires recordings of calibration signals. Since room acoustics, microphone characteristics and the location of the desired speaker change over time, frequent re-calibration is required, making this approach cumbersome and expensive. In [26], a stochastic gradient SDW-MWF has been proposed that avoids the need for calibration. The algorithm however relies on some independence assumptions that are not necessarily satisfied [27]. In this paper, we derive a Frequency-Domain Stochastic Gradient (FD-SG) algorithm for the SP-SDW-MWF. Experiments demonstrate that the proposed FD-SG algorithm preserves the benefit of the exact SP-SDW-MWF over the LMS based SPA, while its computational cost is comparable.

The paper is organized as follows. In Section II, we review the principle of the SP-SDW-MWF. Starting from the cost function of the SP-SDW-MWF, a stochastic gradient algorithm is derived in Section III. To speed-up convergence and reduce computational complexity, the algorithm is implemented in the frequency-domain. In Section III-C, we suggest that the performance of the proposed FD-SG algorithm in spectrally highly non-stationary noise is improved by applying a low pass filter to the part of the gradient estimate that limits speech distortion. The low pass filtering reduces fluctuations in the filter coefficients caused by bad estimates of the short-term speech correlation matrix and/or by the highly non-stationary short-term speech spectrum, while providing good (i.e., similar to GSC) tracking of a change in the noise scenario. Section IV and Section V compare the computational complexity and the performance of the proposed frequency-domain stochastic gradient implementations with the LMS based SPA: the FD-SG with Low Pass Filtering (LPF) offers the best performance, especially in spectrally highly non-stationary noise. Experiments with a Behind-The-Ear (BTE) hearing aid show that the proposed FD-SG with LPF algorithm indeed preserves the performance benefit of the exact SP-SDW-MWF over the QIC-GSC.

Notation

Throughout the paper, the following notation is used.

Time-domain variables are denoted by lower-case letters whereas frequency-domain variables are denoted by upper-case letters. Vectors and matrices are written in bold font and scalars are written in italic. The discrete-time index is denoted by k. The superscripts T, H and $*$ denote matrix/vector transpose, complex conjugate transpose and complex conjugate, respectively. The expectation operator is denoted by $\mathcal{E}\{.\}$. The matrix $\mathbf{F}$ equals the $2L \times 2L$ DFT matrix. The matrix $\mathbf{I}_L$ is the $L \times L$-dimensional identity matrix. The matrix $\mathbf{0}_L$ is an $L \times L$-dimensional matrix of zeros.

II. SPATIALLY PRE-PROCESSED SDW-MWF [24]

A. Concept

Figure 1 describes the SP-SDW-MWF [24]. The SP-SDW-MWF consists of a fixed, spatial pre-processor, i.e., a fixed beamformer $\mathbf{A}(z)$ and a blocking matrix $\mathbf{B}(z)$, and an adaptive SDW-MWF.

Given M microphone signals

$$u_i[k] = u_i^s[k] + u_i^n[k],\ i = 1,\ ...,\ M \tag{1}$$

with $u_i^s[k]$ the desired speech contribution and $u_i^n[k]$ the noise contribution, the *fixed beamformer* $\mathbf{A}(z)$ creates a so-called speech reference

$$y_0[k] = y_0^s[k] + y_0^n[k], \tag{2}$$

by steering a beam towards the direction of the desired signal, with a speech contribution $y_0^s[k]$ and a noise contribution $y_0^n[k]$. In the sequel, the superscripts s and n are used to refer to the speech and noise contribution of a signal. An endfire array is assumed and the desired speaker is assumed to be in front at 0°. The fixed beamformer $\mathbf{A}(z)$ should be designed so that the distortion in the speech reference $y_0^s[k]$ is minimal for all possible errors in the assumed signal model, such as microphone mismatch. In the sequel, a delay-and-sum beamformer is used. For small-sized arrays, this beamformer offers sufficient robustness against signal model errors as it minimizes the noise sensitivity[1]. Given statistical knowledge about the signal model errors that occur in practice such as microphone mismatch, a further optimized beamformer $\mathbf{A}(z)$ can be designed, e.g., using the techniques in [28].

The *blocking matrix* $\mathbf{B}(z)$ creates $M-1$ so-called noise references

$$y_i[k] = y_i^s[k] + y_i^n[k],\ i = 1,\ ...,\ M-1 \tag{3}$$

by steering zeroes towards the front so that the noise contributions $y_i^n[k]$ are dominant compared to the speech leakage contributions $y_i^s[k]$. A simple technique to create the noise references consists of pairwisely subtracting the microphone signals that have been time-aligned for 0°. Further optimized noise references can be created with a spatial filter designed blocking matrix [29] and/or with the design methods for fixed robust broadband beamformers proposed in [28]. A spatial filter designed blocking matrix allows to minimize the speech leakage for a specified angular region around 0° instead of for 0°

[1]The noise sensitivity is defined as the ratio of the spatially white noise gain to the gain of the desired signal [2]. Its reciprocal is the white noise gain [15]. The noise sensitivity and the white noise gain are often used to quantify the sensitivity of an algorithm to errors in the assumed signal model [2], [15].

only [29]. In addition, given statistical knowledge about the microphone characteristics, speech leakage can be minimized for all possible microphone mismatches by using [28]. Instead of using a fixed blocking matrix, also an adaptive blocking matrix can be used to reduce speech leakage [9], [12].

Let *the SDW-MWF* [2] $\mathbf{w}_{0:M-1}[k] \in \mathbb{C}^{ML\times 1}$ be defined as the stacked filter

$$\mathbf{w}_{0:M-1}^H = \left[\begin{array}{cccc} \mathbf{w}_0^H[k] & \mathbf{w}_1^H[k] & ... & \mathbf{w}_{M-1}^H[k] \end{array} \right], \tag{4}$$

where $\mathbf{w}_i[k]$ is the L-dimensional filter applied to the reference signal $y_i[k]$, i.e.,

$$\mathbf{w}_i[k] = \left[\begin{array}{cccc} w_{i,0}[k] & w_{i,1}[k] & ... & w_{i,L-1}[k] \end{array} \right]^T. \tag{5}$$

Similarly, define the ML-dimensional stacked data vector $\mathbf{y}_{0:M-1}[k]$ as[3],

$$\mathbf{y}_{0:M-1}^H[k] = \left[\begin{array}{cccc} \mathbf{y}_0^H[k] & \mathbf{y}_1^H[k] & ... & \mathbf{y}_{M-1}^H[k] \end{array} \right], \tag{6}$$

$$\mathbf{y}_i[k] = \left[\begin{array}{cccc} y_i[k] & y_i[k-1] & ... & y_i[k-L+1] \end{array} \right]^T. \tag{7}$$

The goal of the SDW-MWF $\mathbf{w}_{0:M-1}[k]$ is to minimize the noise energy $\varepsilon_n^2 = \mathcal{E}\{|z^n[k]|^2\}$ in the output $z[k]$, defined as (see also Figure 1)

$$z[k] = y_0[k-\Delta] - \mathbf{w}_{0:M-1}^H[k]\mathbf{y}_{0:M-1}[k], \tag{8}$$

where Δ is an integer delay[4] with $\Delta \in [0,\, L-1]$, while keeping the energy ε_d^2 of the speech distortion in the output $z[k]$, i.e.,

$$\varepsilon_d^2 = \mathcal{E}\{|z^s[k] - y_0^s[k-\Delta]|^2\} = \mathcal{E}\{\left|\mathbf{w}_{0:M-1}^H[k]\mathbf{y}_{0:M-1}^s[k]\right|^2\}, \tag{9}$$

limited [24]. The SDW-MWF is obtained as the filter that minimizes the cost function $J(\mathbf{w}_{0:M-1}[k])$

$$\boxed{J(\mathbf{w}_{0:M-1}[k]) = \frac{1}{\mu}\underbrace{\mathcal{E}\{\left|\mathbf{w}_{0:M-1}^H[k]\mathbf{y}_{0:M-1}^s[k]\right|^2\}}_{\varepsilon_d^2} + \underbrace{\mathcal{E}\{\left|y_0^n[k-\Delta] - \mathbf{w}_{0:M-1}^H[k]\mathbf{y}_{0:M-1}^n[k]\right|^2\}}_{\varepsilon_n^2},} \tag{10}$$

i.e.,

$$\mathbf{w}_{0:M-1}[k] = \left(\frac{1}{\mu}\mathcal{E}\{\mathbf{y}_{0:M-1}^s\mathbf{y}_{0:M-1}^{s,H}\} + \mathcal{E}\{\mathbf{y}_{0:M-1}^n\mathbf{y}_{0:M-1}^{n,H}\}\right)^{-1}\mathcal{E}\{\mathbf{y}_{0:M-1}^n[k]y_0^{n,*}[k-\Delta]\}. \tag{11}$$

The parameter $\mu \in [0,\infty)$ trades off between noise reduction and speech distortion: the larger $\frac{1}{\mu}$, the smaller the amount of speech distortion.

[2]In a time-domain implementation, the filter $\mathbf{w}_{0:M-1}$ and its input signals $\mathbf{y}_{0:M-1}$ are real. Hence, $\mathbf{w}_{0:M-1}^H = \mathbf{w}_{0:M-1}^T$. In the sequel, the formulas are generalized to complex input signals so that they can also be applied to a subband implementation.

[3]In the sequel, the time index k in $\mathbf{y}_{0:M-1}[k]$ is omitted for reasons of conciseness, unless it enhances readability.

[4]The delay Δ is applied to the speech reference to make the filter $\mathbf{w}_{0:M-1}$ non-causal. Usually, it is set to $\lceil \frac{L}{2} \rceil$, where $\lceil x \rceil$ returns the smallest integer equal or larger than x [19].

B. Different parameter settings of SP-SDW-MWF

Depending on the setting of $\frac{1}{\mu}$ and the presence of the filter $\mathbf{w}_0[k]$ on the speech reference, the GSC, the (SDW-)MWF as well as in-between solutions such as the SDR-GSC, may be obtained [24]. Below, the different settings of the SP-SDW-MWF are briefly reviewed.

1) No filter $\mathbf{w}_0[k]$ on the speech reference (SDR-GSC): Without filter $\mathbf{w}_0[k]$, the SP-SDW-MWF $\mathbf{w}_{1:M-1}[k] \in \mathbb{C}^{(M-1)L\times 1}$ equals

$$\mathbf{w}_{1:M-1}[k] = \left(\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s_{1:M-1}\mathbf{y}^{s,H}_{1:M-1}\} + \mathcal{E}\{\mathbf{y}^n_{1:M-1}\mathbf{y}^{n,H}_{1:M-1}\}\right)^{-1} \mathcal{E}\{\mathbf{y}^n_{1:M-1}[k]y_0^{n,*}[k-\Delta]\}, \tag{12}$$

which corresponds to the so-called Speech Distortion Regularized GSC (SDR-GSC). The regularization term $\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s_{1:M-1}\mathbf{y}^{s,H}_{1:M-1}\}$ limits the speech distortion due to possible errors in the assumed signal model. For $\mu = 0$, $\mathbf{w}_{1:M-1}[k] = \mathbf{0}$ so $z[k]$ is equal to the output of the fixed beamformer $\mathbf{A}(z)$ delayed by Δ samples and no speech distortion occurs. For $\mu = \infty$, distortion is completely ignored, which corresponds to the GSC-solution. The SDR-GSC is an alternative technique for the QIC-GSC [15], [17] to decrease the sensitivity of the GSC to signal model errors. In the QIC-GSC, the ANC $\mathbf{w}[k] \in \mathbb{C}^{(M-1)L\times 1}$ minimizes the output noise power subject to the quadratic inequality constraint

$$\mathbf{w}^H[k]\mathbf{w}[k] \leq \beta^2 \tag{13}$$

where $\beta \in [0, \infty)$. In contrast to the QIC-GSC, the SDR-GSC puts more emphasis on speech distortion when the amount of speech leakage increases. As a result, the SDR-GSC outperforms the QIC-GSC for small model errors, while guaranteeing robustness against large model errors [24].

2) Filter $\mathbf{w}_0$ on the speech reference: Since the SP-SDW-MWF takes speech distortion explicitly into account in the design criterion, a filter $\mathbf{w}_0[k]$ on the speech reference can be added. In [24], it is shown that - in the absence of speech leakage and for infinite filter lengths- the SP-SDW-MWF corresponds to a cascade of an SDR-GSC and an SDW Single-channel Wiener post-filter [30], [31]. In the presence of speech leakage, the SP-SDW-MWF with $\mathbf{w}_0[k]$ contains extra filter operations that compensate for the performance degradation of the SDR-GSC due to the speech leakage: in contrast to the SDR-GSC (and the GSC), performance is not degraded by microphone mismatch. For a given maximum allowable speech distortion level, the SP-SDW-MWF with $\mathbf{w}_0[k]$ outperforms the SDR-GSC and the QIC-GSC [24].

C. Implementation of SP-SDW-MWF

In practice, the correlation matrix $\mathcal{E}\{\mathbf{y}^s_{0:M-1}\mathbf{y}^{s,H}_{0:M-1}\}$ is unknown. During periods of speech, the inputs $y_i[k]$ consist of speech + noise, i.e., $y_i[k] = y_i^s[k] + y_i^n[k]$, $i = 0, ..., M-1$. During periods of noise

(speech pauses), only the noise component is observed, i.e., $y_i[k] = y_i^n[k]$. The second order statistics of the noise are assumed to be sufficiently stationary so that they can be estimated during periods of noise. Assuming that the speech and noise signal are uncorrelated, $\mathcal{E}\{\mathbf{y}^s_{0:M-1}\mathbf{y}^{s,H}_{0:M-1}\}$ can then be estimated as

$$\mathcal{E}\{\mathbf{y}^s_{0:M-1}\mathbf{y}^{s,H}_{0:M-1}\} = \mathcal{E}\{\mathbf{y}_{0:M-1}[k]\mathbf{y}^H_{0:M-1}[k]\} - \mathcal{E}\{\mathbf{y}^n_{0:M-1}[k]\mathbf{y}^{n,H}_{0:M-1}[k]\}, \tag{14}$$

where $\mathcal{E}\{\mathbf{y}_{0:M-1}[k]\mathbf{y}^H_{0:M-1}[k]\}$ is estimated during periods of speech + noise and $\mathcal{E}\{\mathbf{y}^n_{0:M-1}[k]\mathbf{y}^{n,H}_{0:M-1}[k]\}$ during periods of noise only. Like for the GSC, a robust speech detection is thus needed. Using (14), the SP-SDW-MWF (11) can be re-written as:

$$\mathbf{w}_{0:M-1}[k]=\left(\frac{1}{\mu}\mathcal{E}\{\mathbf{y}_{0:M-1}\mathbf{y}^H_{0:M-1}\} + (1-\frac{1}{\mu})\mathcal{E}\{\mathbf{y}^n_{0:M-1}\mathbf{y}^{n,H}_{0:M-1}\}\right)^{-1}\mathcal{E}\{\mathbf{y}^n_{0:M-1}[k]y_0^{n,*}[k-\Delta]\}. \tag{15}$$

So far, only recursive *matrix-based* algorithms are available for implementing the SP-SDW-MWF [20]-[24]. In Section III, stochastic gradient algorithms are derived for the SP-SDW-MWF. To reduce computational complexity and speed up convergence, the algorithms are implemented in the frequency-domain.

D. Performance measures

To assess the performance of the different algorithms, the broadband intelligibility weighted Signal-to-Noise Ratio (SNR) improvement [32] is used, defined as

$$\Delta\text{SNR}_{\text{intellig}} = \sum_i I_i(\text{SNR}_{i,\text{out}} - \text{SNR}_{i,\text{in}}), \tag{16}$$

where the band importance function I_i expresses the importance of the i-th one-third octave band with center frequency f_i^c for intelligibility, $\text{SNR}_{i,\text{out}}$ is the output SNR (in dB) and $\text{SNR}_{i,\text{in}}$ is the input SNR (in dB) in the i-th one third octave band. The center frequencies f_i^c and the values I_i are defined in [33]. The intelligibility weighted signal-to-noise ratio reflects how much intelligibility is improved by the noise reduction algorithms, but does not take into account speech distortion.

To measure the amount of speech distortion, we define the following intelligibility weighted spectral distortion measure

$$\text{SD}_{\text{intellig}} = \sum_i I_i \text{SD}_\text{i} \tag{17}$$

with SD_i the average spectral distortion (dB) in i-th one-third band, measured as

$$\text{SD}_i = \int_{2^{-1/6}f_i^c}^{2^{1/6}f_i^c} |10\log_{10} G^s(f)|\, df \Big/ \left[\left(2^{1/6} - 2^{-1/6}\right) f_i^c\right], \tag{18}$$

with $G^s(f)$ the power transfer function of speech from the input to the output of the noise reduction algorithm.

III. STOCHASTIC GRADIENT ALGORITHM

Starting from the cost function (10) of the SP-SDW-MWF, a Time-Domain Stochastic Gradient algorithm (TD-SG) is derived in Section III-A. To speed-up convergence and reduce computational complexity, the TD-SG algorithm is implemented in the frequency-domain, as described in Section III-B. For spectrally highly non-stationary noise, the proposed stochastic gradient algorithm is expected to suffer from a large excess error. Section III-C shows that - if the *long-term* spectral and the spatial characteristics of the noise vary slowly in time - the excess error can be reduced by low pass filtering the part of the gradient estimate that limits speech distortion, so that spectrally highly non-stationary noise can still be spatially suppressed.

A. Time-domain Stochastic Gradient algorithm (TD-SG)

A stochastic gradient algorithm approximates the steepest descent algorithm

$$\mathbf{w}[j+1] = \mathbf{w}[j] + \rho \left(-\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} \right)_{\mathbf{w}=\mathbf{w}[j]}, \tag{19}$$

where j denotes the iteration index and ρ equals the step size, using an instantaneous gradient estimate. Replacing the iteration index j by a time index k and leaving out the expectation values $\mathcal{E}\{.\}$ as is typically done in the derivation of a stochastic gradient algorithm, we obtain the following update equation[5] for the cost function (10):

$$\boxed{\mathbf{w}[k+1] = \mathbf{w}[k] + \rho \left\{ \mathbf{y}^n[k](y_0^{n,*}[k-\Delta] - \mathbf{y}^{n,H}[k]\mathbf{w}[k]) - \mathbf{r}[k] \right\},} \tag{20}$$

with

$$\boxed{\mathbf{r}[k] = \frac{1}{\mu}\mathbf{y}^s[k]\mathbf{y}^{s,H}[k]\mathbf{w}[k],} \tag{21}$$

where $\mathbf{w}[k]$, $\mathbf{y}[k] \in \mathbb{C}^{NL\times 1}$ and N denotes the number of input channels to the adaptive filter ($N = M$ if $\mathbf{w}_0$ is present, $N = M-1$ if $\mathbf{w}_0$ is absent) . For $\frac{1}{\mu} = 0$ and no filter $\mathbf{w}_0$ on the speech reference, equation (20) reduces to an LMS type update formula often used in GSC, which is then operated during periods of *noise only* (i.e., when $y_i[k] = y_i^n[k]$, $i = 0, ..., M-1$) . The additional term $\mathbf{r}[k]$ in the gradient estimate limits the speech distortion due to possible signal model errors.

Equation (21) requires knowledge of the correlation matrix $\mathbf{y}^s[k]\mathbf{y}^{s,H}[k]$ or $\mathcal{E}\{\mathbf{y}^s[k]\mathbf{y}^{s,H}[k]\}$ of the clean speech. In practice, this information is not available. To avoid the need for calibration, L-dimensional

[5]The update formula can be applied to both the SDR-GSC and the SP-SDW-MWF with $\mathbf{w}_0$. In the sequel, the subscripts $0:M-1$ (or $1:M-1$) in the adaptive filter $\mathbf{w}_{0:M-1}$ (or $\mathbf{w}_{1:M-1}$) and the input vector $\mathbf{y}_{0:M-1}$ (or $\mathbf{y}_{1:M-1}$) are omitted for the sake of conciseness.

speech + noise signal vectors $\mathbf{y}_i[k]$ of the channels $i = M-N, ..., M-1$ are stored in a circular buffer $\mathbf{B}_1 \in \mathbb{C}^{L_{buf_1} \times N}$ with $L_{buf_1} = lL$ ($l \in \mathbb{N}$) the buffer length per channel, during processing as in [26]. During periods of noise only (i.e., when $y_i[k] = y_i^n[k]$, $i = 0, ..., M-1$), the filter $\mathbf{w}[k]$ is updated using the following approximation of the term $\mathbf{r}[k]$ in (20):

$$\boxed{\mathbf{r}[k] = \frac{1}{\mu}\left(\mathbf{y}_{buf_1}[k]\mathbf{y}_{buf_1}^H[k] - \mathbf{y}^n[k]\mathbf{y}^{n,H}[k]\right)\mathbf{w}[k],} \tag{22}$$

where $\mathbf{y}_{buf_1}[k]$ is an NL-dimensional stacked speech + noise vector that is created by concatenating the speech + noise vectors $\mathbf{y}_{i,buf_1}[k]$, $i = M-N, ..., M-1$ from the buffer $\mathbf{B}_1$ as in (6).

Additional storage of L-dimensional noise only vectors $\mathbf{y}_i^n[k]$, $i = 0, \cdots, M-1$ in a second buffer $\mathbf{B}_2 \in \mathbb{C}^{L_{buf_2} \times M}$ with L_{buf_2} the buffer length per channel, allows to adapt $\mathbf{w}[k]$ also *during periods of speech + noise*, using:

$$\begin{aligned} \mathbf{w}[k+1] &= \mathbf{w}[k] + \rho\left\{\mathbf{y}_{buf_2}[k](y_{0,buf_2}^*[k-\Delta] - \mathbf{y}_{buf_2}^H[k]\mathbf{w}[k]) - \mathbf{r}[k]\right\}, \\ \mathbf{r}[k] &= \frac{1}{\mu}\left(\mathbf{y}[k]\mathbf{y}^H[k] - \mathbf{y}_{buf_2}[k]\mathbf{y}_{buf_2}^H[k]\right)\mathbf{w}[k]. \end{aligned} \tag{23}$$

In the sequel, we refer to (20), (22) and (23) as the TD-SG algorithm. The SG algorithm (20) with the true regularization term (21) is called *optimal* TD-SG algorithm.

Remark 1: For $\mu > 1$, an alternative stochastic gradient algorithm similar to [26] can be derived from the TD-SG algorithm by invoking some independence assumptions. Simulations, however, showed that these independence assumptions result in a significant performance degradation with respect to the TD-SG algorithm, while hardly reducing the computational complexity [27].

Below, we will - for reasons of conciseness - only consider the update procedure during noise only of the TD-SG algorithm to explain further refinements and modifications, hence, $\mathbf{y}[k] = \mathbf{y}^n[k]$ and $y_0[k] = y_0^n[k]$. The extension towards updating during speech + noise periods with the use of a second, noise buffer $\mathbf{B}_2$ is straightforward: e.g., equation (23) is found by replacing the noise-only, input vectors $\mathbf{y}^n[k]$ by $\mathbf{y}_{buf_2}[k]$ and the speech + noise vectors $\mathbf{y}_{buf_1}[k]$ by the input speech + noise vector $\mathbf{y}[k]$ in formulas (20)-(22).

Using[6]

$$\mathbf{w}_{opt} = \left(\frac{1}{\mu}\mathcal{E}\{\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H\} + (1-\frac{1}{\mu})\mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}\}\right)^{-1}\mathcal{E}\{\mathbf{y}^{n,H}[k]y_0^{n,*}[k-\Delta]\}, \tag{24}$$

and (20), (22), it can be shown that

$$\mathcal{E}\{\mathbf{w}[k+1] - \mathbf{w}_{opt}\} = \left(\mathbf{I}_{NL} - \rho\mathcal{E}\left\{\frac{1}{\mu}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H + (1-\frac{1}{\mu})\mathbf{y}^n\mathbf{y}^{n,H}\right\}\right)^{k+1}\mathcal{E}\{\mathbf{w}[0] - \mathbf{w}_{opt}\}. \tag{25}$$

[6]When the second order statistics of the noise are stationary, i.e., $\mathcal{E}\{\mathbf{y}_{buf_1}^n\mathbf{y}_{buf_1}^{n,H}\} = \mathcal{E}\{\mathbf{y}^n\mathbf{y}^{n,H}\}$, $\mathbf{w}_{opt}$ equals (11).

Hence, the TD-SG algorithm (20), (22) is *convergent in the mean* provided that the step size ρ is smaller than $\frac{2}{\lambda_{\max}}$ with $\lambda_{\max}$ the maximum eigenvalue of $\mathcal{E}\{\frac{1}{\mu}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H + (1-\frac{1}{\mu})\mathbf{y}^n\mathbf{y}^{n,H}\}$. The similarity of (20) with standard LMS lets us presume that setting

$$\rho \;<\; \frac{2}{\frac{1}{\mu}L\sum_{i=M-N}^{M-1}\mathcal{E}\{|y_{i,buf_1}[k]|^2\} + (1-\frac{1}{\mu})L\sum_{i=M-N}^{M-1}\mathcal{E}\{|y_i^n[k]|^2\}}, \tag{26}$$

guarantees *convergence in the mean square*. In the sequel, the step size ρ of (20)-(22) is therefore normalized as:

$$\rho = \frac{\rho'}{\frac{1}{\mu}\left|\mathbf{y}_{buf_1}^H\mathbf{y}_{buf_1}[k] - \mathbf{y}^{n,H}\mathbf{y}^n[k]\right| + \mathbf{y}^{n,H}\mathbf{y}^n[k] + \delta}, \tag{27}$$

where the step size $\rho' \in [0,2)$ and δ is a very small constant. The absolute value $\left|\mathbf{y}_{buf_1}^H\mathbf{y}_{buf_1} - \mathbf{y}^{n,H}\mathbf{y}^n\right|$ has been inserted to guarantee a positive valued estimate of the clean speech energy $\mathbf{y}^{s,H}\mathbf{y}^s[k]$.

However, the estimate (22) (and (23)) of the regularization term (21) is quite bad due to large differences between the rank-one noise correlation matrices $\mathbf{y}^n[k]\mathbf{y}^{n,H}[k]$ and $\mathbf{y}^n[k']\mathbf{y}^{n,H}[k']$ at different time instants k and k'. Consequently, the instantaneous gradient estimate in the TD-SG update equation (20)-(22) is -compared to the optimal TD-SG algorithm (20)-(21)- additionally perturbed by

$$\frac{1}{\mu}\left(\mathbf{y}^n[k]\mathbf{y}^{n,H}[k] - \mathbf{y}_{buf_1}^n[k]\mathbf{y}_{buf_1}^{n,H}[k]\right)\mathbf{w}[k], \tag{28}$$

for $\mu \neq \infty$. Hence, for $\mu \neq \infty$, the TD-SG algorithm, i.e., (20), (22) and (23), suffers from a larger residual excess error than the optimal TD-SG algorithm (20), (21). The additional excess error grows for decreasing μ, increasing step size ρ and increasing vector length NL of the vector $\mathbf{y}[k]$. It is expected to be especially large for highly non-stationary noise, e.g., multi-talker babble noise. The gradient estimate can be improved by replacing the instantaneous estimate

$$\mathbf{y}_{buf_1}[k]\mathbf{y}_{buf_1}^H[k] - \mathbf{y}^n[k]\mathbf{y}^{n,H}[k] \tag{29}$$

in (22) with a time-average, resulting in

$$\mathbf{r}[k] = \frac{1}{\mu}\frac{1}{K}\left(\sum_{l=k-K+1}^{k}\mathbf{y}_{buf_1}[l]\mathbf{y}_{buf_1}^H[l] - \sum_{l=k-K+1}^{k}\mathbf{y}^n[l]\mathbf{y}^{n,H}[l]\right)\mathbf{w}[k]. \tag{30}$$

However, this would require expensive matrix operations.

A frequency-domain implementation implicitly performs this averaging (30) over $K = L$ samples, while offering a reduced computational complexity and a faster convergence compared to its time-domain equivalent. Therefore, the TD-SG algorithm is implemented in the frequency-domain as is described in Section III-B. To improve the performance in spectrally highly non-stationary noise, Section III-C proposes an approximation of (30) without matrix operations that allows for averaging over more than L samples.

B. Frequency-Domain Stochastic Gradient algorithm (FD-SG)

Similarly as frequency-domain LMS [34], the Frequency-Domain Stochastic Gradient algorithm (FD-SG) is derived from a *block-based TD-SG implementation* with block length $K = L$ by computing the linear convolutions and correlations in the DFT-domain. The block-based TD-SG algorithm (and hence, the FD-SG algorithm) intrinsically performs the averaging (30) over $K = L$ samples:

$$\begin{aligned}\mathbf{w}[(k+1)L] &= \mathbf{w}[kL] + \frac{\rho}{L}\left[\sum_{i=0}^{L-1}\mathbf{y}^n[kL+i]\left(y_0^{n,*}[kL+i-\Delta] - \mathbf{y}^{n,H}[kL+i]\mathbf{w}[kL]\right)\right. \\ &\left. -\frac{1}{\mu}\sum_{i=0}^{L-1}\left(\mathbf{y}_{buf_1}[kL+i]\mathbf{y}_{buf_1}^H[kL+i] - \mathbf{y}^n[kL+i]\mathbf{y}^{n,H}[kL+i]\right)\mathbf{w}[kL]\right]. \qquad (31)\end{aligned}$$

The gradient and hence also (29) are averaged over L terms prior to making adjustments to $\mathbf{w}$. This goes at the expense of a reduced (i.e., by a factor L) update rate [34], [35]. For stationary signals, the block-based algorithm requires an L times larger step size than the nonblock-based algorithm to obtain the same convergence rate and misadjustment. The block-based algorithm (31) is convergent in the mean provided that the step size is smaller than $\frac{2}{\lambda_{\max}}$ with $\lambda_{\max}$ the maximum eigenvalue of $\mathcal{E}\{\frac{1}{\mu}\mathbf{y}_{buf_1}\mathbf{y}_{buf_1}^H + (1-\frac{1}{\mu})\mathbf{y}^n\mathbf{y}^{n,H}\}$. As the stability bound for the step size of the block-based algorithm is the same as for the nonblock-based algorithm, the block-based algorithm is more likely to run unstable when a fast convergence rate is required [35], [34].

The block-based implementation (31) is computationally more efficient, especially for large filter lengths, when it is implemented in the *frequency-domain*: the linear convolutions and correlations are then efficiently realized by FFT algorithms based on overlap-save or overlap-add [34]. In this paper, we concentrate on an FD-SG implementation based on overlap-save. Let the $2L$-dimensional $\mathbf{W}_i[k]$ be derived from the corresponding time-domain quantity $\mathbf{w}_i[k]$, $i = M-N, \cdots, M-1$, as

$$\begin{aligned}\mathbf{W}_i[k] &= \left[\begin{array}{cccc} W_{i,0}[k] & W_{i,1}[k] & \cdots & W_{i,2L-1}\ [k]\end{array}\right]^T \\ &= \mathbf{F}\left[\begin{array}{cccc} \mathbf{w}_i^H[k] & 0 & \cdots & 0\end{array}\right]^T, \qquad (32)\end{aligned}$$

and let - during periods of noise - the $2L \times 2L$-dimensional matrices $\mathbf{Y}_i^n[k]$ and $\mathbf{Y}_i[k]$ be derived from $\mathbf{y}_i^n[kL-1]$,$\mathbf{y}_i^n[kL+L-1]$ and $\mathbf{y}_{buf_1,i}[kL-1]$, $\mathbf{y}_{buf_1,i}[kL+L-1]$, as

$$\mathbf{Y}_i^n[k] = \text{diag}\left\{\mathbf{F}\left[\begin{array}{ccc} y_i^n[kL-L] & ... & y_i^n[kL+L-1]\end{array}\right]^T\right\}, \qquad (33)$$

$$\mathbf{Y}_i[k] = \text{diag}\left\{\mathbf{F}\left[\begin{array}{ccc} y_{buf_1,i}[kL-L] & ... & y_{buf_1,i}[kL+L-1]\end{array}\right]^T\right\}, \qquad (34)$$

and

$$\mathbf{d}[k] = \left[\begin{array}{cccc} y_0^n[kL-\Delta] & ... & y_0^n[kL-\Delta+L-1] \end{array} \right]^T, \tag{35}$$

then the FD-SG implementation of (31) becomes

$$\boxed{\begin{aligned} \mathbf{W}_i[k+1] &= \mathbf{W}_i[k] + \mathbf{F}\mathbf{g}\mathbf{F}^{-1}\boldsymbol{\rho}[k]\left(\mathbf{Y}_i^{n,H}[k]\mathbf{E}[k] - \mathbf{R}_i[k]\right), \\ \mathbf{R}_i[k] &= \frac{1}{\mu}\left(\mathbf{Y}_i^H[k]\mathbf{E}_2[k] - \mathbf{Y}_i^{n,H}[k]\mathbf{E}_1[k]\right), \end{aligned}} \tag{36}$$

with

$$\mathbf{E}[k] = \mathbf{F}\mathbf{k}^T\left(\mathbf{d}[k] - \mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j^n[k]\mathbf{W}_j[k]\right), \tag{37}$$

$$\mathbf{E}_1[k] = \mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j^n[k]\mathbf{W}_j[k], \tag{38}$$

$$\mathbf{E}_2[k] = \mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1}\sum_{j=M-N}^{M-1}\mathbf{Y}_j[k]\mathbf{W}_j[k], \tag{39}$$

where the matrices $\mathbf{g}$ and $\mathbf{k}$ are defined as

$$\mathbf{g} = \left[\begin{array}{cc} \mathbf{I}_L & \mathbf{0}_L \\ \mathbf{0}_L & \mathbf{0}_L \end{array} \right]; \ \mathbf{k} = \left[\begin{array}{cc} \mathbf{0}_L & \mathbf{I}_L \end{array} \right]. \tag{40}$$

The time-varying matrix $\boldsymbol{\rho}[k]$ in (36) is a diagonal matrix that contains the step sizes $\rho_m[k]$, $m = 0, \cdots, 2L-1$ of the different frequency bins m, i.e.,

$$\boldsymbol{\rho}[k] = \text{diag } \{\rho_0[k], ..., \rho_{2L-1}[k]\}. \tag{41}$$

Each step size $\rho_m[k]$ is normalized using the noise power and the speech + noise power in that bin:

$$\rho_m[k] = \frac{2\rho'}{L}P_m^{-1}[k], \tag{42}$$

$$P_m[k] = \gamma P_m[k-1] + (1-\gamma)\left(\sum_{j=M-N}^{M-1}\left|Y_{j,m}^n[k]\right|^2 + \frac{1}{\mu}\left|\sum_{j=M-N}^{M-1}\left(|Y_{j,m}[k]|^2 - \left|Y_{j,m}^n[k]\right|^2\right)\right|\right), \tag{43}$$

where $\rho' \in [0,2)$, γ is a forgetting factor and $Y_{j,m}^n[k]$, $Y_{j,m}[k]$ are the m-th diagonal element of $\mathbf{Y}_j^n[k]$ and $\mathbf{Y}_j[k]$, respectively.

In the FD-SG implementation, each frequency bin has its own optimal step size $\rho_m[k]$, resulting in faster convergence compared to a time-domain implementation while (upon convergence) not degrading the steady-state excess MSE [34]: the different bins have a more uniform convergence rate and excess

error that are less dependent on the eigenvalue spread of the correlation matrix of the input signals and hence, on the power spectra of the input signals. In a *time-domain implementation*, one common step size ρ is (implicitly) used for the different frequency bins. The convergence rate and excess errors in each bin then depend on the power spectra of the input signals: in frequency bins with little power this common step size will be smaller than in the frequency-domain approach, resulting in slower convergence and less excess error in this bin; in frequency bins with large power on the other hand, this common step size will be larger than in the frequency-domain approach, resulting in faster convergence but larger excess error. The improvement $\Delta\text{SNR}_{\text{intellig}}$ (16) of the algorithms for different signal model errors (e.g., microphone mismatch) is determined by the SNR-improvements and hence, the excess errors in each frequency bin. In a time-domain implementation, the power spectra of the input signals not only determine the convergence rate but also the improvement $\Delta\text{SNR}_{\text{intellig}}$: signal model errors modify the power spectra of the noise references and hence, the convergence rates and excess errors in the different frequency bins, so that comparing the improvement $\Delta\text{SNR}_{\text{intellig}}$ for different signal model errors becomes difficult. Frequency-domain implementations that optimize the step size for each frequency bin, are more appropriate to evaluate the improvement $\Delta\text{SNR}_{\text{intellig}}$ (16) of the algorithms for different signal model errors: the change in power spectra caused by signal model errors has a smaller effect on the convergence rates and the excess errors in the different frequency bins so that differences in $\Delta\text{SNR}_{\text{intellig}}$ are mostly due to the signal model errors and not to different convergence rates.

The straightforward extension of (33)-(43) to include updating during speech + noise can be found in the tabulated Algorithm 1 of Section III-C, which summarizes the complete FD-SG algorithm (including the improvement in Section III-C). Note that N FFT operations can be saved in the FD-SG algorithm, by storing the FFT-transformed speech + noise and noise-only vectors in the buffers[7] $\tilde{\mathbf{B}}_1 \in \mathbb{C}^{L_{buf_1} \times N}$ and $\tilde{\mathbf{B}}_2 \in \mathbb{C}^{L_{buf_2} \times N}$, respectively, instead of storing the time-domain vectors: then, the FD-SG (36) requires $(3N + 4)$ FFTs of length $2L$. When adaptation is also done during speech + noise periods, also the time-domain vector $\mathbf{y}_0^n[k - \Delta]$ should be stored in an additional buffer $\tilde{\mathbf{B}}_{2,0} \in \mathbb{R}^{L_{buf_2} \times 1}$ during periods of noise-only, which -for $N = M$- results in an additional storage of L_{buf_2} words.

*Remark 2: In (*36*), a common trade-off parameter μ is used in all frequency bins. Alternatively, a different μ can be used per frequency bin. E.g., for the SDR-GSC, μ could be set to ∞ at those frequencies where the GSC is sufficiently robust, e.g., for small-sized arrays at high frequencies [13]. In this case,*

[7]Since the input signals are real, half of the FFT components are complex-conjugated. Hence, only half of the complex FFT components have to be stored in memory.

only a few low-frequency components of the regularization terms $\mathbf{R}_i[k]$, $i = M-N, \cdots, M-1$ *have to be computed, reducing the computational complexity.*

C. Improvement of stochastic gradient algorithm

For spectrally stationary noise, the limited (i.e., $K = L$) averaging of (29) by the block-based[8] and FD-SG implementation proposed in Section III-B may offer a reasonable estimate of the *short-term* speech correlation matrix $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$. However, in practical scenarios, the speech and the noise signals are often spectrally highly non-stationary (e.g., multi-talker babble noise) while their *long-term* spectral and spatial characteristics such as the positions of the sources usually vary more slowly in time. For these scenarios, a reliable estimate of the *long-term* speech correlation matrix $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ that captures the spatial rather than the *short-term* spectral characteristics can still be obtained by averaging (29) over $K \gg NL$ samples. Spectrally highly non-stationary noise can then still be spatially suppressed by using an estimate of the *long-term* speech correlation matrix in the regularization term $\mathbf{r}[k]$, i.e., by setting $K \gg NL$ in (30). The limited averaging ($K = L$) in the FD-SG algorithm does not suffice to make such an estimate. In this section, we propose a cheap method to incorporate a long-term averaging ($K \gg NL$) of (30) in the FD-SG algorithm, i.e., by low pass filtering the part of the gradient estimate that takes speech distortion into account, i.e., the term $\mathbf{R}_i[k]$ in (36). The averaging method is first explained for the TD-SG algorithm of Section III-A and then translated to the FD-SG implementation.

1) TD-SG with Low Pass Filtering (LPF): Assume that the *long-term* spectral and spatial characteristics of the noise are quasi-stationary during at least K speech + noise samples and K noise samples. A reliable estimate of the long-term speech correlation matrix $\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ is then obtained by (30) with $K \gg NL$. To avoid expensive matrix computations, we approximate (30) by

$$\mathbf{r}[k] = \frac{1}{\mu}\frac{1}{K}\sum_{l=k-K+1}^{k}\left(\mathbf{y}_{buf_1}[l]\mathbf{y}_{buf_1}^H[l] - \mathbf{y}^n[l]\mathbf{y}^{n,H}[l]\right)\mathbf{w}[l]. \tag{44}$$

Since the filter coefficients $\mathbf{w}$ of a stochastic gradient algorithm vary slowly in time, (44) appears a good approximation of (30), especially for small step sizes ρ' and/or small K. The averaging operation (44) is performed by low pass filtering $\mathbf{r}[k]$, i.e.,

$$\boxed{\mathbf{r}[k] = \tilde{\lambda}\mathbf{r}[k-1] + (1-\tilde{\lambda})\frac{1}{\mu}\left(\mathbf{y}_{buf_1}[k]\mathbf{y}_{buf_1}^H[k] - \mathbf{y}^n[k]\mathbf{y}^{n,H}[k]\right)\mathbf{w}[k],} \tag{45}$$

[8]A block-based algorithm with a large block length K with $K \gg L$ would result in a too slow convergence rate.

where $\tilde{\lambda} < 1$. This corresponds to an averaging window of about $K = \frac{1}{1-\tilde{\lambda}}$ samples. The normalized step size ρ in (27) is modified into

$$\rho = \frac{\rho'}{\zeta[k] + \mathbf{y}^{n,H}[k]\mathbf{y}^{n}[k] + \delta} \tag{46}$$

$$\zeta[k] = \tilde{\lambda}\zeta[k-1] + (1-\tilde{\lambda})\frac{1}{\mu}\left|\mathbf{y}^{H}_{buf_1}[k]\mathbf{y}_{buf_1}[k] - \mathbf{y}^{n,H}[k]\mathbf{y}^{n}[k]\right|. \tag{47}$$

In the sequel, we use the abbreviation LPF to refer to the low pass filtering.

2) FD-SG with LPF: Equation (45) can be extended to the FD-SG algorithm, as:

$$\boxed{\mathbf{R}_i[k] = \lambda\mathbf{R}_i[k-1] + (1-\lambda)\frac{1}{\mu}\left(\mathbf{Y}_i^{H}[k]\mathbf{E}_2[k] - \mathbf{Y}_i^{n,H}[k]\mathbf{E}_1[k]\right).} \tag{48}$$

To obtain the same time constant in the averaging operation as in the TD-SG with LPF, λ should equal $\tilde{\lambda}^L$. The step sizes $\rho_m[k] = \frac{2\rho'}{LP_m[k]}$ in the different frequency bins are now normalized using:

$$P_m[k] = \gamma P_m[k-1] + (1-\gamma)\left(P_{1,m}[k] + P_{2,m}[k]\right) \tag{49}$$

$$P_{1,m}[k] = \sum_{j=M-N}^{M-1}\left|Y^{n}_{j,m}[k]\right|^2 \tag{50}$$

$$P_{2,m}[k] = \lambda P_{2,m}[k-1] + (1-\lambda)\frac{1}{\mu}\left|\sum_{j=M-N}^{M-1}\left(|Y_{j,m}[k]|^2 - \left|Y^{n}_{j,m}[k]\right|^2\right)\right|. \tag{51}$$

Experiments in Section V show that the LPF (48) improves the performance of the FD-SG, especially for spectrally highly non-stationary noise. The term $\mathbf{w}^{H}[k]\mathbf{r}[k]$ with $\mathbf{r}[k]$ computed as (45) and hence, also $\sum_{i=M-N}^{M-1}\mathbf{W}_i^{H}[k]\mathbf{F}\mathbf{g}\mathbf{F}^{-1}\mathbf{R}_i[k]$ with $\mathbf{R}_i[k]$ calculated as in (48), is an approximate measure for the average speech distortion over the last $K = 1/(1-\tilde{\lambda})$ samples. Hence, the FD-SG adaptive filter with LPF reduces noise while keeping the average distortion low. The LPF reduces fluctuations in the filter weights $\mathbf{W}_i[k]$ caused by bad estimates of the short-term speech correlation matrix $\mathcal{E}\{\mathbf{y}^{s}[k]\mathbf{y}^{s,H}[k]\}$ and/or by the highly non-stationary *short-term* speech spectrum. Admittedly, due to the LPF, only slow changes in the nominal position of the desired speaker can be tracked. However, in hearing aid applications, we assume the nominal position of the desired speaker to be in front of the hearing aid so that the spatial characteristics of the speech are expected to vary slowly in time. Experiments in Section V show that the FD-SG with LPF implementation (with e.g., $\lambda = 0.999$) provides good (i.e., similar to GSC) tracking of a sudden change in noise scenario. During periods of noise, the update formula (32) is given by

$$\mathbf{W}_i[k+1] = \mathbf{W}_i[k] + \underbrace{\mathbf{F}\mathbf{g}\mathbf{F}^{-1}\boldsymbol{\rho}[k]\mathbf{Y}_i^{n,H}[k]\mathbf{E}[k]}_{\delta\mathbf{W}_{i,1}[k]} - \underbrace{\mathbf{F}\mathbf{g}\mathbf{F}^{-1}\boldsymbol{\rho}[k]\mathbf{R}_i[k]}_{\delta\mathbf{W}_{i,2}[k]}. \tag{52}$$

A sudden change in noise scenario has an immediate impact on the term $\delta\mathbf{W}_{i,1}[k]$: like for GSC, the noise in the new scenario is suppressed with a convergence time determined by the step size matrix $\boldsymbol{\rho}[k]$. The

amount of noise suppression and speech distortion is determined by the term $\delta\mathbf{W}_{i,2}[k]$, which converges more slowly than $\delta\mathbf{W}_{i,1}[k]$ due to the averaging (48): initially, the distortion measured before the change determines $\mathbf{R}_i[k]$ and hence, the amount of noise suppression and speech distortion. Experiments in Section V show that also in the transient phase after the change in noise scenario, the speech distortion of the FD-SG with LPF implementation of the SP-SDW-MWF with $\frac{1}{\mu} > 0$ is small compared to the speech distortion of the GSC (i.e., SDR-GSC with $\frac{1}{\mu} = 0$).

The complete FD-SG algorithm with LPF is summarized in the tabulated Algorithm 1. Compared to the FD-SG algorithm of Section III-B, the FD-SG with LPF requires one extra $2L-$point FFT and about $8NL$ extra real multiply-accumulates (MACs) per L samples and additional memory storage of a $2NL \times 1$ real data vector.

IV. Computational complexity of different stochastic gradient algorithms

Table I summarizes the computational complexity (expressed in number of real multiply-accumulates[9] (MACs), divisions (D), square roots (Sq) and absolute values (Abs)) of the time-domain and frequency-domain SG implementations of the SP-SDW-MWF. We assume that one complex multiplication is equivalent to 4 real multiplications and 2 real additions. A $2L$-point FFT of a real input vector requires $2L \log_2 2L$ real MACs (assuming the radix-2 FFT algorithm). Comparison is made with standard Normalized LMS (NLMS) based ANC and the NLMS based SPA [9], [15]. The update equation of the ANC corresponds to the update equations (20) and (36) of the TD-SG and FD-SG with the number of input channels N set to $M-1$ and the trade-off parameter $1/\mu$ set to 0. The NLMS based SPA is translated to the frequency-domain by the following equations:

$$\|\mathbf{w}[k]\|_2^2 \;=\; \mathbf{w}^T[k]\mathbf{w}[k] = \frac{1}{2L}\sum_{i=1}^{M-1}\mathbf{W}_i^H[k]\mathbf{W}_i[k]; \tag{53}$$

$$\text{If } \|\mathbf{w}[k]\|_2^2 \geq \beta^2 : \mathbf{W}_i[k] \leftarrow \beta\frac{\mathbf{W}_i[k]}{\|\mathbf{w}[k]\|_2}, \, i = 1, \cdots, M-1. \tag{54}$$

Table I indicates that the TD-SG SDR-GSC (i.e., the SP-SDW-MWF without filter $\mathbf{w}_0$ and hence, $N = M-1$) and the NLMS-based SPA are about twice as complex as the standard ANC. When applying an LPF to the regularization term, the TD-SG SDR-GSC algorithm has about three times the computational complexity of the ANC. The SP-SDW-MWF with extra filter $\mathbf{w}_0$ on the speech reference is a bit more complex. The increase in complexity of the frequency-domain implementations is smaller.

[9]Counted as the number of multiply-accumulates, additions and multiplications.

As an illustration, Figure 2 plots the computational complexity (expressed in number of Mega operations[10] per second (Mops)) of the TD-SG and FD-SG with LPF as a function of L for $M = 3$ and a sampling frequency $f_s = 16$ kHz. Comparison is made with the NLMS based ANC and the NLMS based SPA. For small M, the computational complexity of the FD-NLMS based SPA is not significantly different from the cost of the FD-NLMS ANC. For $L > 8$, the frequency-domain implementations result in a significantly lower computational cost compared to their time-domain equivalents. The computational cost of the FD-SG with LPF is comparable to the complexity of the FD-NLMS based SPA, making it a good alternative to the SPA for implementation in hearing aids.

Remark 3: In Table I and Figure 2*, the computational complexity of time-domain and frequency-domain NLMS ANC and NLMS based SPA represents the computational complexity when the adaptive filter is only updated during noise only periods. If the adaptive filter is also updated during speech + noise periods using data from a noise buffer, the time-domain implementations require* $(M-1)L$ *additional MACs per sample and the frequency-domain implementations require* 2 *additional FFT and* $(4L(M-1)-2(M-1)+L)$ *MACs per* L *samples to compute the output.*
The computational complexity of NLMS based SPA in Table I and Figure 2 *represents the worst-case complexity, i.e., assuming that the filter* $\mathbf{w}$ *has to be scaled down after each update.*

V. Experimental results

In Section III, a low-cost, stochastic gradient implementation of the SP-SDW-MWF has been derived: to speed up convergence and reduce computational complexity, the algorithm is implemented in the frequency-domain. In spectrally highly non-stationary noise, the FD-SG algorithm is expected to suffer from a large excess error, especially for a large step size ρ' and for a large value of $1/\mu$. To reduce the excess error, Section III-C suggests to apply a low pass filter to the regularization term in the FD-SG algorithm that limits speech distortion.

In this section, we analyze the impact of the low pass filter on the FD-SG algorithm and compare the performance of the FD-SG with LPF algorithm and the FD-NLMS based SPA based on experimental results for a hearing aid application.

A. Set-up

A BTE hearing aid with three omnidirectional microphones (Knowles FG-3452) in endfire configuration, has been mounted on a dummy head in an office room. The interspacing d between the first and

[10]Counted as the number of MACs, divisions, square roots and absolute values.

the second microphone is about $d = 1$ cm and the interspacing between the second and third microphone about 1.5 cm. The reverberation time $T_{60\,\mathrm{dB}}$ is about 700 ms. The desired speech signal and the noise signals are uncorrelated. The desired speech source consists of sentences spoken by a male speaker from the HINT-database [36]. Both the speech and the noise signal have a level of 70 dB SPL at the center of the head. The desired speech source and noise sources are positioned at a distance of 1 meter from the head: the speech source in front of the head, the noise sources at an angle θ with respect to the speech source. For evaluation purposes, the speech and noise signal have been recorded separately. The sampling frequency f_s equals 16 kHz.

The microphone signals are pre-whitened prior to processing to improve performance [37], and the output is accordingly de-whitened. In the experiments, the microphones have been calibrated by means of recordings of an anechoic stationary speech weighted noise signal positioned at 0° measured while the microphone array was mounted on the head. A delay-and-sum beamformer is used as a fixed beamformer, since -in case of small microphone interspacing - it is robust to model errors. The blocking matrix $\mathbf{B}$ subtracts pairwisely the time aligned calibrated microphone signals.

The performance of the FD stochastic gradient algorithms is evaluated for a filter length $L = 32$ taps per channel, $\rho' = 0.8$ and $\gamma = 0.95$. For these values of ρ' and L, a good convergence behavior and noise reduction performance has been obtained: for larger L, the performance did not improve significantly. As indicated in Section III-A, the approximation (22) in the FD-SG algorithm causes an additional excess error that increases with ρ' and L, hence, ρ' and L should not be chosen too large. In the experiments, speech+noise and noise periods were marked manually. To exclude the effect of the spatial pre-processor, the performance measures (16) and (17) are calculated with respect to the output of the fixed beamformer. In some experiments, a microphone gain mismatch $\upsilon_2 = 4$ dB is applied to the second microphone to illustrate the sensitivity of the algorithms to errors in the assumed signal model. Among the different possible signal model errors, microphone mismatch was found to be especially harmful to the performance of the GSC in a hearing aid application [13], [14]. In hearing aids, microphones are rarely matched in gain and phase. In [38], gain and phase differences between microphone characteristics of up to 6 dB and 10°, respectively, have been reported.

B. Comparison of the FD-SG algorithm and the FD-SG with LPF

In this section, the FD-SG (i.e., Algorithm 1 with $\lambda = 0$) and the FD-SG with LPF (i.e., Algorithm 1 with $\lambda \neq 0$) implementation of the SP-SDW-MWF are compared.

In **Figure 3** (a) and (b), we compare the performance in terms of $\Delta\mathrm{SNR}_{\mathrm{intellig}}$ (16) and $\mathrm{SD}_{\mathrm{intellig}}$ (17)

of the FD-SG and the FD-SG with LPF (with exponential weighting factor $\lambda = 0.999$) implementation of the SDR-GSC (i.e., SP-SDW-MWF without $\mathbf{w}_0$) for different values of the trade-off parameter[11] $1/\mu$. Figure 3(a) depicts the performance for a stationary (i.e., speech-like noise) noise source at 90°, while Figure 3(b) plots the performance measures for spectrally non-stationary (i.e., multi-talker babble) noise at 90°. In this experiment, no model errors such as microphone mismatch have been applied (i.e., $\upsilon_2 = 0$ dB). To analyze the impact of the approximation (22) of the true regularization term (21) in the FD-SG algorithms, the performance of the FD implementation of the optimal TD-SG algorithm (20)-(21), which uses the clean speech, is depicted too. This algorithm, which we consider in some sense an optimal solution, is referred to as optimal FD-SG algorithm.

As indicated in Section III, the approximation (22) in the FD-SG algorithm indeed results in a performance degradation with respect to the optimal FD-SG, especially for large $\frac{1}{\mu}$ and spectrally highly non-stationary noise. For stationary speech-like noise, the FD-SG algorithm does not suffer a lot from the approximation, but for spectrally highly non-stationary noise, such as multi-talker babble, the limited averaging of approximation (22) in the FD implementation (cf. Section III-B) does not suffice to preserve the large noise reduction performance of the optimal FD-SG. The performance of the FD-SG in highly non-stationary noise is significantly improved by applying a low pass filter (48) with $\lambda = 0.999$ to the regularization term. Alternatively, the performance of the FD-SG algorithm could be improved by decreasing the step size ρ'. Decreasing the step size, however, compromises the convergence speed.

Figure 3 demonstrated that the noise reduction performance of the FD-SG SDR-GSC in multi-talker babble noise is significantly improved by applying a low pass filter to the regularization term $\mathbf{R}_i[k]$. In Figure 3, the exponential weighting λ of the LPF was set to 0.999, which corresponds to an averaging over $\frac{1}{1-\lambda} = 1000$ frames (1 frame corresponds to 32 samples or 2 ms). In **Figure 4**, we study the influence of the forgetting factor λ on the performance of the SDR-GSC and the SP-SDW-MWF with $\mathbf{w}_0$ ($\frac{1}{\mu} = 0.5$) for the babble noise scenario. Figure 4 plots $\Delta\text{SNR}_{\text{intellig}}$ and $\text{SD}_{\text{intellig}}$ obtained with the FD-SG with LPF implementation of the SDR-GSC and SP-SDW-MWF with $\mathbf{w}_0$ as a function of $\frac{1}{1-\lambda}$. Performance clearly improves for increasing λ, especially for λ up to 0.999. The SP-SDW-MWF with $\mathbf{w}_0$ suffers from a larger excess error -and hence worse $\Delta\text{SNR}_{\text{intellig}}$- compared to the SDR-GSC. This is due to the larger dimensions of $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ and the relatively large contribution of the speech reference in $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$.

The LPF reduces fluctuations in the filter coefficients caused by bad estimates of the short-term speech

[11]Note that the SDR-GSC reduces to the GSC if $1/\mu = 0$, since then, $\mathbf{R}_i[k] = 0$ (see also (36)).

correlation matrix $\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ and/or the highly non-stationary *short-term* spectrum of speech. Below, we analyze the influence of the LPF on the tracking performance of the FD-SG algorithm. As an illustration, **Figure 5** (a) and (b) compare the tracking performance of the FD-SG algorithm with LPF ($\lambda = 0.999$) and the FD-SG algorithm without LPF (i.e., $\lambda = 0$), when the noise source position suddenly changes from 150° to 60° and from 60° to 150°. The changes in noise position occur during periods of noise only at sample number 1.6×10^5 and sample number 2.4×10^5, respectively. Figure 5(a) depicts the results for the SDR-GSC (with $1/\mu = 0.5$), Figure 5(b) considers the SP-SDW-MWF with $\mathbf{w}_0$ (with $1/\mu = 0.5$). For comparison, the result of the GSC (i.e., the SDR-GSC with $1/\mu = 0$) is also shown in Figure 5(a). In the experiment, a gain mismatch υ_2 of 4 dB was applied to the second microphone. To avoid fast changes in the residual noise energy ε_n^2 and speech distortion energy ε_d^2 due to highly non-stationary short-term speech and noise spectra, the desired and interfering source in this experiment are stationary, speech-like noise: half of the samples contains speech + noise, the other half contains noise only. The upper figures depict the residual noise energy ε_n^2 (defined in (10)) as a function of time (expressed in number of samples at $f_s = 16$ kHz), the lower figures plot the speech distortion energy ε_d^2 during speech + noise periods as a function of the number of speech + noise samples. The speech and noise energy ε_d^2 and ε_n^2 have been normalized with the speech and noise energy in the speech reference $y_0[k]$, respectively.

When a sudden change in noise scenario occurs, the FD-SG with LPF implementation of the SDR-GSC and the SP-SDW-MWF with $\mathbf{w}_0$ suppress the noise with a convergence time similar to the FD-NLMS GSC, while the speech distortion ε_d^2 is smaller than for GSC, also in the transient phase after the change. In addition, the LPF improves the noise reduction performance of the FD-SG in the transient phase. After the change, the instantaneous estimate (22) of the true regularization term (21) is bad: e.g., from sample number 2.4×10^5 until 2.5×10^5, only noise is observed so that the data in the speech + noise buffer are not immediately updated. As a result, the FD-SG implementation with $\lambda = 0$ suffers from a bad noise reduction performance. In the FD-SG with LPF, the regularization term is averaged over time by the LPF, which reduces the negative impact of inaccurate estimates of the instantaneous speech correlation matrix on the regularization term and hence, on the performance.

Remark 4: Because of the LPF, a wrong speech correlation estimate at time instant k has an influence on future regularization terms that exponentially decreases in time. This explains the further, slight decrease in the residual noise energy ε_n^2 in Figure 5(a) from sample number 1.7×10^5 until 1.9×10^5 after the initial convergence.

C. Comparison with FD-NLMS based SPA

The experiments above demonstrated that the FD-SG with LPF algorithm outperforms the FD-SG algorithm and offers a good tracking performance. In this section, we investigate whether this FD-SG with LPF implementation of the SP-SDW-MWF preserves the benefit of the exact SP-SDW-MWF over the QIC-GSC by comparing it with the FD-NLMS based SPA. For a fair comparison, the FD-NLMS based SPA is -like the FD-SG algorithm- also adapted during speech + noise using data from a noise buffer.

Figure 6 and **Figure 7** compare the performance of the FD-SG SP-SDW-MWF with LPF ($\lambda = 0.999$) and the FD-NLMS based SPA (53)-(54) in a multiple noise source scenario. The noise scenario consists of 3 multi-talker babble noise sources positioned at angles $75^\circ, 180^\circ, 240^\circ$ with respect to the desired source at 0°. To assess the sensitivity of the algorithms against errors in the assumed signal model, the influence of a gain mismatch $\upsilon_2 = 4$ dB of the second microphone on the performance is depicted too. In **Figure 6**, the improvement $\Delta\text{SNR}_{\text{intellig}}$ and the distortion $\text{SD}_{\text{intellig}}$ of the SDR-GSC and SP-SDW-MWF with $\mathbf{w}_0$ is depicted as a function of the trade off factor $\frac{1}{\mu}$. **Figure 7** shows the results of the FD-NLMS based SPA of (53)-(54) for different QIC values β^2. The FD-NLMS based GSC corresponds to the FD-SG SDR-GSC with $\frac{1}{\mu} = 0$ or the FD-NLMS based SPA with $\beta^2 = \infty$.

In this noise scenario, the GSC (i.e., the SDR-GSC with $\frac{1}{\mu} = 0$ or the QIC-GSC with $\beta^2 = \infty$) still offers good noise suppression for a mismatch of 4 dB, at the expense of a large distortion. Both, the SPA and the stochastic gradient based SP-SDW-MWF increase the robustness of the GSC: distortion decreases with increasing $\frac{1}{\mu}$ and decreasing β^2. The SPA is more conservative than the SDR-GSC: the constraint value β^2 should be chosen so that the maximum permissible speech distortion is not exceeded for the largest model error, e.g., 5 dB $\text{SD}_{\text{intellig}}$ for a gain mismatch up to 4 dB. This goes at the expense of less noise reduction in case of smaller model errors (e.g., $\Delta\text{SNR}_{\text{intellig}} = 4\,\text{dB}$ for $\beta^2 = 0.4$). The SDR-GSC on the other hand only puts emphasis on speech distortion if required, i.e., when the amount of speech leakage is large, so that a better noise reduction is obtained for small model errors (e.g., $\Delta\text{SNR}_{\text{intellig}}$ between 4 dB and 7.4 dB for $\frac{1}{\mu} = 0.5$). The SP-SDW-MWF offers more noise suppression at even larger model errors: the performance of the SP-SDW-MWF with $\mathbf{w}_0$ is -compared to the SDR-GSC and the SPA- less affected by microphone mismatch. In the absence of model errors, the SP-SDW-MWF with $\mathbf{w}_0$ achieves a slightly worse performance than the SDR-GSC. With $\mathbf{w}_0$, the estimate of $\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ is less accurate due to the larger dimensions of $\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ and the large contribution of the speech reference in $\frac{1}{\mu}\mathcal{E}\{\mathbf{y}^s\mathbf{y}^{s,H}\}$ (see also Figure 4).

In short, the proposed FD-SG with LPF implementation of the SP-SDW-MWF preserves the benefit of the exact SP-SDW-MWF over the QIC-GSC.

VI. Conclusions

In this paper, we derived a frequency-domain stochastic gradient (FD-SG) implementation for the SP-SDW-MWF and compared its performance to the NLMS-based SPA [15] for QIC-GSC. Starting from the cost function of the SP-SDW-MWF, an TD-SG algorithm has been derived in Section III. To speed-up convergence and reduce computational complexity, the algorithm has been implemented in the frequency-domain. Experiments in Section V indicated that the FD-SG algorithm suffers from a reasonably large excess error when the noise is spectrally non-stationary. In Section III-C and Section V, we showed that -for spatially slowly time-varying noise scenarios- the excess error of the proposed FD-SG algorithm is reduced by low pass filtering the part of the gradient estimate that limits speech distortion so that spectrally highly non-stationary noise can still be successfully suppressed. The low pass filtering reduces fluctuations in the filter coefficients caused by bad estimates of the short-term speech correlation matrix and/or by the highly non-stationary short-term speech spectrum, while providing good (i.e., similar to GSC) tracking of a sudden change in the noise scenario. Finally, experiments in Section V demonstrated that the proposed FD-SG preserves the benefit of the SP-SDW-MWF over QIC-GSC. The comparable computational cost and the better noise reduction performance of the proposed algorithm make it a good alternative to the SPA for implementation in hearing aids.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable suggestions and comments.

References

[1] R. W. Stadler and W. M. Rabinowitz, "On the potential of fixed arrays for hearing aids," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1332–1342, Sept. 1993.

[2] P. M. Peterson, *Adaptive array processing for multiple microphone hearing aids*, Ph.D. thesis, Dept. Elect. Eng. and Comp. Sci., M.I.T, Cambridge, MA, 1989, available as Res. Lab. Elect. Techn. Rept. 541.

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[5] J. E. Greenberg and P. M. Zurek, "Evaluation of an Adaptive Beamforming Method for Hearing Aids," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1662–1676, Mar. 1992.

[6] M. Kompis and N. Dillier, "Noise Reduction for Hearing Aids: Combining Directional Microphones with an Adaptive Beamformer," *J. Acoust. Soc. Amer.*, vol. 96, no. 3, pp. 1910–1913, Sept. 1994.

[7] J. Vanden Berghe and J. Wouters, "An adaptive noise canceller for hearing aids using two nearby microphones," *J. Acoust. Soc. Amer.*, vol. 103, pp. 3621–3626, June 1998.

[8] M. W. Hoffman and K. M. Buckley, "Robust time-domain processing of broadband acoustic data," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 193–203, May 1995.

[9] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Signal Processing*, vol. 47, pp. 2677–2683, 1999.

[10] F. Luo, J. Yang, C. Pavlovic, and A. Nehorai, "Adaptive Null-Forming Scheme in Digital Hearing Aids," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1583–1590, July 2002.

[11] 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, Apr. 1990, vol. 2, pp. 833–836.

[12] W. Herbordt and W. Kellermann, "Frequency-domain integration of acoustic echo cancellation and a generalized sidelobe canceller with improved robustness," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 123–132, Mar.-Apr. 2002.

[13] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of GSVD based optimal Filtering and generalized Sidelobe Canceller for Hearing Aid Applications," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2001, pp. 31–34.

[14] A. Spriet, M. Moonen, and J. Wouters, "Robustness Analysis of Multi-channel Wiener Filtering and Generalized Sidelobe Cancellation for Multi-microphone Noise Reduction in Hearing Aid Applications," To appear in *IEEE Trans. Speech, Audio Processing*, May 2005. Available at ftp://ftp.esat.kuleuven.ac.be/pub/sista/spriet/reports/02-81.pdf.

[15] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust Adaptive Beamforming," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[16] N. K. Jablon, "Adaptive beamforming with the Generalized Sidelobe Canceller in the presence of array imperfections," *IEEE Trans. Antennas Propag.*, vol. 34, pp. 996–1012, Aug. 1986.

[17] Z. Tian, K.L. Bell, and H.L. Van Trees, "A Recursive Least Squares Implementation for LCMP Beamforming Under Quadratic Constraint," *IEEE Trans. Signal Processing*, vol. 49, no. 6, pp. 1138–1145, June 2001.

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

[19] S. Doclo and M. Moonen, "GSVD-Based Optimal Filtering for Single and Multimicrophone Speech Enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

[20] S. Doclo and M. Moonen, "Multi-microphone noise reduction using recursive GSVD-based optimal filtering with ANC postprocessing stage," To appear in *IEEE Trans. Speech, Audio Processing*. Available at ftp://ftp.esat.kuleuven.ac.be/pub/sista/doclo/reports/02-04.ps.gz.

[21] G. Rombouts and M. Moonen, "QRD-based unconstrained optimal filtering for acoustic noise reduction," *Signal Processing*, vol. 83, no. 9, pp. 1889–1904, Sept. 2003.

[22] A. Spriet, M. Moonen, and J. Wouters, "A multi-channel subband generalized singular value decomposition approach to speech enhancement," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 149–158, Mar.-Apr. 2002.

[23] A. Spriet, M. Moonen, and J. Wouters, "A multichannel subband GSVD based approach for speech enhancement in hearing aids," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Darmstadt, Germany, Sept. 2001, pp. 187–191.

[24] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," in *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[25] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," *IEEE Trans. Speech, Audio Processing*, vol. 7, no. 3, pp. 241–252, May 1999.

[26] D. A. Florêncio and H. S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, May 2001.

[27] A. Spriet, M. Moonen, and J. Wouters, "Stochastic gradient based implementation of spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction in hearing aids," Tech. Rep. ESAT-SISTA/TR 03-47a, ESAT/SISTA, K.U. Leuven (Belgium), 2003, available at ftp://ftp.esat.kuleuven.ac.be/pub/sista/spriet/reports/03-47a.pdf.

[28] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Trans. Signal Processing*, vol. 51, no. 10, pp. 2511–2526, Oct. 2003.

[29] S. Nordebo, I. Claesson, and S. Nordholm, "Adaptive beamforming: spatial filter designed blocking matrix," *IEEE Journal of Oceanic Engineering*, vol. 19, no. 4, pp. 583–590, Oct. 1994.

[30] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[31] K. U. Simmer, J. Bitzer, and C. Marro, *Post-filtering techniques*, chapter 3 in "Microphone Arrays: Signal Processing Techniques and Applications" (Brandstein, M. S. and Ward, D. B., Eds.), pp. 39–57, Springer-Verlag, May 2001.

[32] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *J. Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[33] Acoustical Society of America, "ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index," June 1997.

[34] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, pp. 14–37, Jan. 1992.

[35] G. A. Clark, S. K. Mitra, and S. R. Parker, "Block implementation of adaptive digital filters," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 29, no. 3, pp. 744–752, 1981.

[36] M. Nilsson, S. D. Soli, and A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099, Feb. 1994.

[37] M. J. Link and K. M. Buckley, "Prewhitening for intelligibility gain in hearing aid arrays," *J. Acoust. Soc. Amer.*, vol. 93, no. 4, pp. 2139–2140, Apr. 1993.

[38] L. B. Jensen, "Hearing Aid with adaptive Matching of Input Transducers," U.S. patent 0041696, Apr. 2002.



Ann Spriet (S'01) was born in Izegem, Belgium, in 1976. She received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1999. From 1999 until 2003, she was a Research Assistant of the Fonds voor Wetenschappelijk Onderzoek - Vlaanderen. Currently, she is a postdoctoral researcher at the Electrical Engineering Department (ESAT) of the Katholieke Universiteit Leuven, in the framework of a collaboration with Katholieke Universiteit Leuven - Laboratory Experimental ORL. Her research interests are in digital signal processing for speech enhancement in hearing aids and cochlear implants. She received the 2001 Lapperre Award and a Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2003.



Marc Moonen (M'94) received the Electrical Engineering degree and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1986 and 1990, respectively. Since 2000 he has been an Associate Professor with the Electrical Engineering Department, Katholieke Universiteit Leuven, where he is currently heading a research team of sixteen Ph.D. candidates and postdocs, working in the area of signal processing for digital communications, wireless communications, DSL, and audio signal processing. Dr. Moonen received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with P. Vandaele) and was a 1997 "Laureate of the Belgium Royal Academy of Science." He was Chairman of the IEEE Benelux Signal Processing Chapter (1998-2002), and is currently a EURASIP AdCom Member (European Association for Signal, Speech, and Signal Processing, 2000). He is Editor-in Chief for the *EURASIP Journal on Applied Signal Processing* (2003), and a member of the editorial board of *Integration, the VLSI Journal, IEEE Transactions on Circuits and Systems II, EURASIP Journal on Wireless Communications and Networking, and IEEE Signal Processing Magazine*.



Jan Wouters was born in Leuven, Belgium, in 1960. He received the physics degree and the Ph.D. degree in sciences / physics from the Katholieke Universiteit Leuven, Leuven, Belgium, in 1982 and 1989, respectively. From 1989 till 1992 he has been a Research Fellow with the Belgian NFWO (National Fund for Scientific Research) at the Institute of Nuclear Physics (UCL Louvain-la-Neuve and KU Leuven) and at NASA Goddard Space Flight Center (USA). From 1993 he is a Professor at the Neurosciences Department of the Katholieke Universiteit Leuven. His research activities are about audiology and the auditory system, signal processing for cochlear implants and hearing aids. Dr. Wouters received in 1989 an Award of the Flemish Ministry, in 1992 a Fullbright Award and a NATO Research Fellowship, the 1996 Flemish VVL Speech therapy - Audiology Award. He is member of the International Collegium for Rehabilitative Audiology, a Board Member of the NAG (Dutch Acoustical Society) and responsible for the Lab. Exp. ORL, author of 75 articles in international peer-review journals, reviewer for several international journals.

LIST OF FIGURES

1 Spatially Pre-processed SDW-MWF. 28

2 Computational complexity (expressed in Mops) of TD-SG and FD -SG algorithm with LPF as a function of filter length L per channel; $M = 3$. For comparison, the computational complexity of the standard NLMS ANC and SPA are depicted too. 28

3 Performance of the FD-SG and FD-SG with LPF ($\lambda = 0.999$) implementation of SDR-GSC with respect to the optimal FD-SG SDR-GSC for different values of $1/\mu$; (a) Stationary speech-like noise at 90° ; (b) Multi-talker babble noise at 90°. 29

4 Performance of the FD-SG with LPF implementation of the SDR-GSC and the SP-SDW-MWF with $\mathbf{w}_0$ ($\frac{1}{\mu} = 0.5$) as a function of $1/(1-\lambda)$ where λ is the exponential weighting factor of the LPF. Babble noise at 90°. 29

5 Tracking performance of the FD-SG with LPF based SP-SDW-MWF (with $\lambda = 0.999$) compared to the GSC and the FD-SG implementation with $\lambda = 0$. The upper figures show the normalized residual noise energy as a function of time. The lower figures plot the normalized speech distortion energy as a function of the number of speech+noise samples. The noise source position suddenly changes from 150° to 60° at sample number 1.6×10^5 and from 60° to 150° at sample number 2.4×10^5. (a) SDR-GSC ($\frac{1}{\mu} = 0.5$) and GSC; (b) SP-SDW-MWF with $\mathbf{w}_0$ ($\frac{1}{\mu} = 0.5$). 30

6 Performance of FD-SG SP-SDW-MWF with LPF ($\lambda = 0.999$) in a multiple noise source scenario. 30

7 Performance of FD-NLMS based SPA in a multiple noise source scenario. 31

LIST OF TABLES

I Computational complexity of Time-Domain (TD) and Frequency-Domain (FD) SG implementations of the SP-SDW-MWF (expressed in number of real MACs, divisions (D), absolute values (Abs) and square roots (Sq) per sample) 31

LIST OF ALGORITHMS

1 FD-SG with LPF based on overlap-save. 32

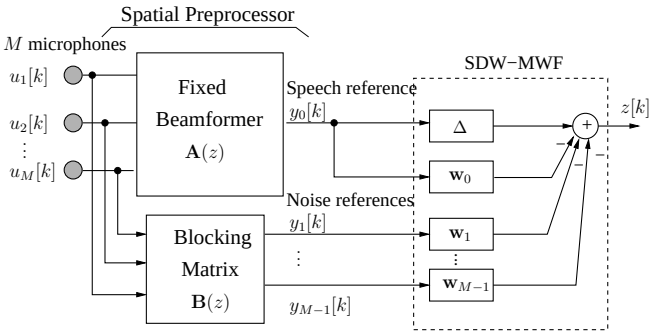


Fig. 1. Spatially Pre-processed SDW-MWF.

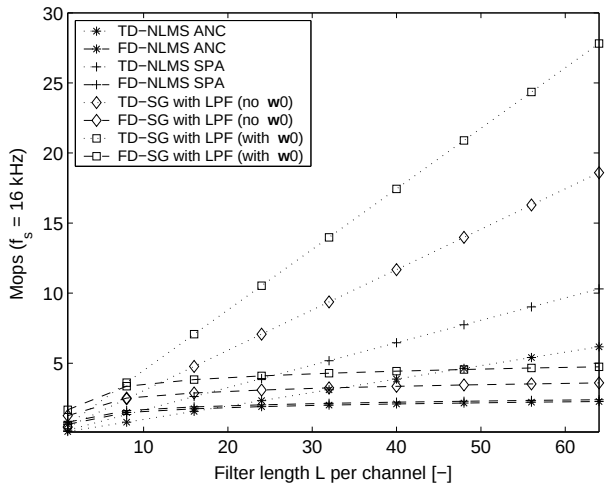


Fig. 2. Computational complexity (expressed in Mops) of TD-SG and FD -SG algorithm with LPF as a function of filter length L per channel; $M = 3$. For comparison, the computational complexity of the standard NLMS ANC and SPA are depicted too.

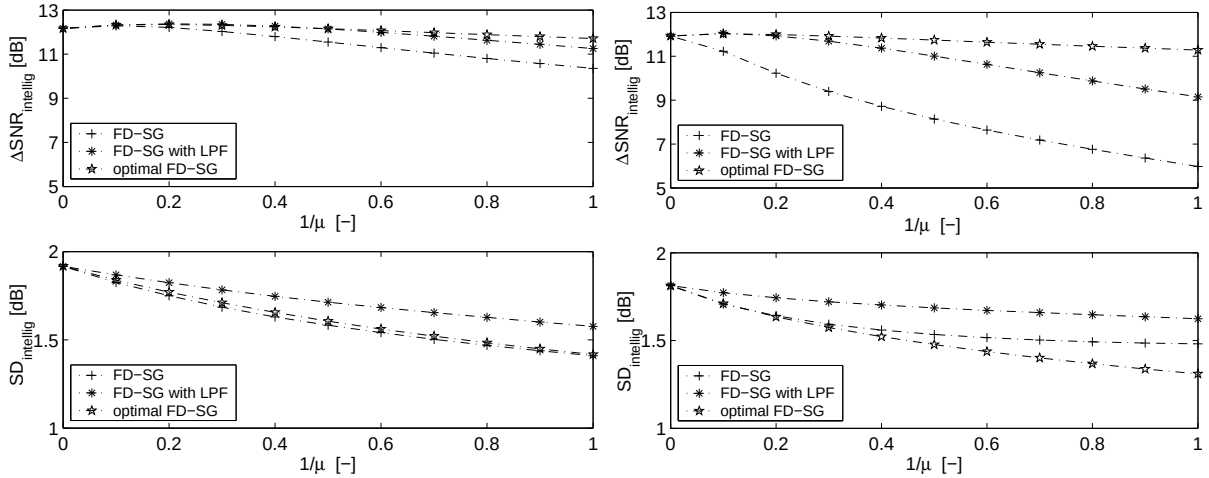


Fig. 3. Performance of the FD-SG and FD-SG with LPF ($\lambda = 0.999$) implementation of SDR-GSC with respect to the optimal FD-SG SDR-GSC for different values of $1/\mu$; (a) Stationary speech-like noise at 90° ; (b) Multi-talker babble noise at 90°.

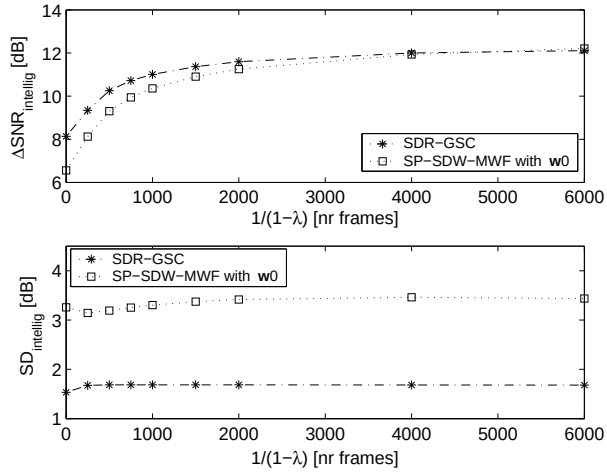
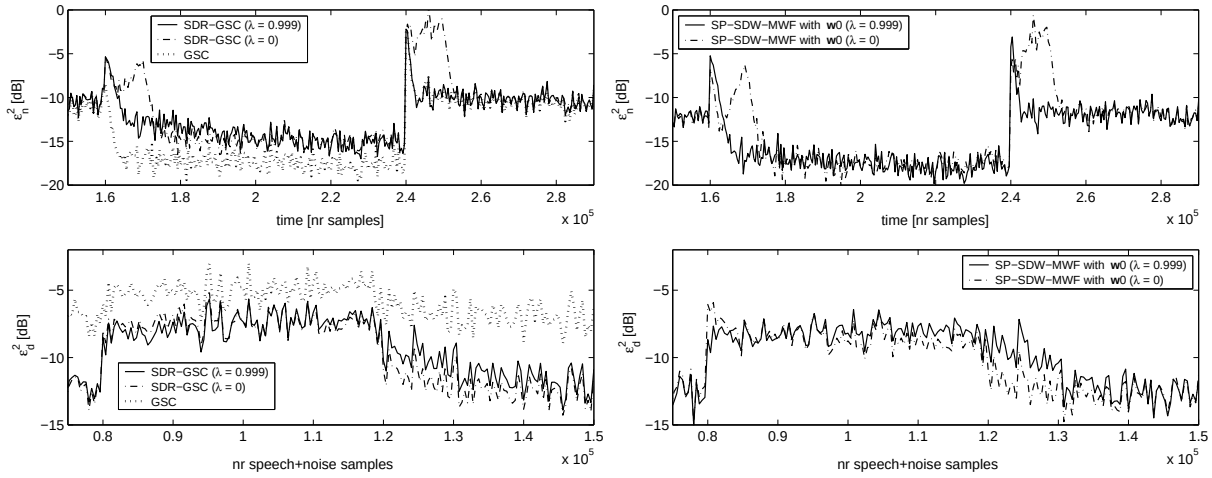


Fig. 4. Performance of the FD-SG with LPF implementation of the SDR-GSC and the SP-SDW-MWF with $\mathbf{w}_0$ ($\frac{1}{\mu} = 0.5$) as a function of $1/(1-\lambda)$ where λ is the exponential weighting factor of the LPF. Babble noise at 90°.



(a) (b)

Fig. 5. Tracking performance of the FD-SG with LPF based SP-SDW-MWF (with $\lambda = 0.999$) compared to the GSC and the FD-SG implementation with $\lambda = 0$. The upper figures show the normalized residual noise energy as a function of time. The lower figures plot the normalized speech distortion energy as a function of the number of speech+noise samples. The noise source position suddenly changes from 150° to 60° at sample number 1.6×10^5 and from 60° to 150° at sample number 2.4×10^5. (a) SDR-GSC ($\frac{1}{\mu} = 0.5$) and GSC; (b) SP-SDW-MWF with $\mathbf{w}_0$ ($\frac{1}{\mu} = 0.5$).

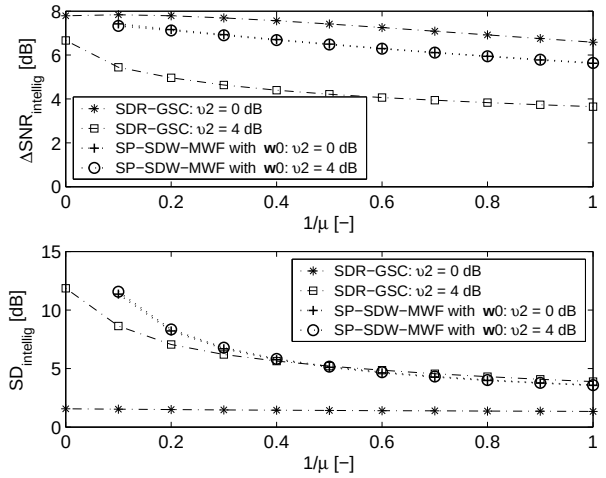


Fig. 6. Performance of FD-SG SP-SDW-MWF with LPF ($\lambda = 0.999$) in a multiple noise source scenario.

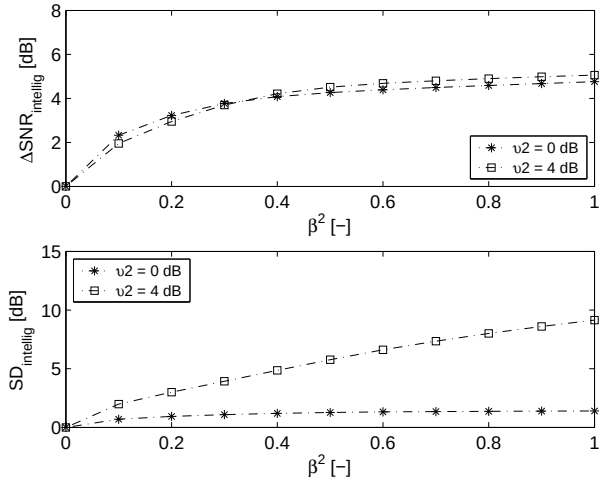


Fig. 7. Performance of FD-NLMS based SPA in a multiple noise source scenario.

TABLE I

COMPUTATIONAL COMPLEXITY OF TIME-DOMAIN (TD) AND FREQUENCY-DOMAIN (FD) SG IMPLEMENTATIONS OF THE SP-SDW-MWF (EXPRESSED IN NUMBER OF REAL MACS, DIVISIONS (D), ABSOLUTE VALUES (ABS) AND SQUARE ROOTS (SQ) PER SAMPLE)

Algorithm		**update formula**	**adaptation of step size**
TD	NLMS ANC	$(2(M-1)L+1)$ MAC	1 D $+(M-1)L$ MAC
	NLMS SPA	$(4(M-1)L+1)$ MAC $+1$ D $+1$ Sq	1 D $+(M-1)L$ MAC
	SG	$(4NL+5)$ MAC	1 D $+1$ Abs $+(2NL+2)$ MAC
	SG with LPF	$(7NL+4)$ MAC	1 D $+1$ Abs $+(2NL+4)$ MAC
FD	NLMS ANC	$(10M-7-\frac{4(M-1)}{L})+(6M-2)\log_2 2L$ MAC	1 D $+(2M+2)$ MAC
	NLMS SPA	$(14M-11-\frac{4(M-1)}{L})+(6M-2)\log_2 2L$ MAC $+1/L$ Sq $+1/L$ D	1 D $+(2M+2)$ MAC
	SG	$(18N+6-\frac{8N}{L})+(6N+8)\log_2 2L$ MAC	1 D $+1$ Abs $+(4N+4)$ MAC
	SG with LPF	$(26N+4-\frac{10N}{L})+(6N+10)\log_2 2L$ MAC	1 D$+1$ Abs$+(4N+6)$ MAC

Algorithm 1 FD-SG with LPF based on overlap-save.

Initialization and matrix definitions:

$$\mathbf{W}_i[0] = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^T,\ i = M-N,\ ...,\ M-1$$

$$P_m[0] = \delta_m,\ m = 0,\ ...,\ 2L-1$$

$$\tilde{\mathbf{B}}_1 \in \mathbb{C}^{L_{buf_1} \times N} = \text{speech+noise buffer}$$

$$\tilde{\mathbf{B}}_2 \in \mathbb{C}^{L_{buf_2} \times N},\ \tilde{\mathbf{B}}_{2,0} \in \mathbb{R}^{L_{buf_2} \times 1} = \text{noise buffers}$$

$$L_{buf_1} = v \cdot L,\ L_{buf_2} = w \cdot L \text{ with } v,\ w \in \mathbb{N}$$

For each new block of ML input samples:

If noise detected:

$$\mathbf{d}[k] = \begin{bmatrix} y_0[kL-\Delta] & \cdots & y_0[kL-\Delta+L-1] \end{bmatrix}^T$$

$$\mathbf{Y}_i^n[k] = \text{diag}\left\{ \mathbf{F} \begin{bmatrix} y_i[kL-L] & \cdots & y_i[kL+L-1] \end{bmatrix}^T \right\}$$

Store input data $\mathbf{Y}_i^n[k]$, $\mathbf{d}[k]$ in noise buffers $\tilde{\mathbf{B}}_2$, $\tilde{\mathbf{B}}_{2,0}$

Create $\mathbf{Y}_i[k]$ from data in speech+noise buffer $\tilde{\mathbf{B}}_1$

If speech detected:

$$\mathbf{Y}_i[k] = \text{diag}\left\{ \mathbf{F} \begin{bmatrix} y_i[kL-L] & \cdots & y_i[kL+L-1] \end{bmatrix}^T \right\}$$

Store input data $\mathbf{Y}_i[k]$ in speech + noise buffer $\tilde{\mathbf{B}}_1$

Create $\mathbf{Y}_i^n[k]$, $\mathbf{d}[k]$ from data in noise buffers $\tilde{\mathbf{B}}_2$, $\tilde{\mathbf{B}}_{2,0}$

Update formula:

$$\mathbf{W}_i[k+1] = \mathbf{W}_i[k] + \mathbf{F}\mathbf{g}\mathbf{F}^{-1}\boldsymbol{\rho}[k] \left\{ \mathbf{Y}_i^{n,H}[k]\mathbf{E}[k] - \mathbf{R}_i[k] \right\},$$

$$\mathbf{R}_i[k] = \lambda \mathbf{R}_i[k-1] + (1-\lambda)\frac{1}{\mu}\left(\mathbf{Y}_i^H[k]\mathbf{E}_2[k] - \mathbf{Y}_i^{n,H}[k]\mathbf{E}_1[k] \right),$$

with

$$\mathbf{E}_1[k] = \mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1} \sum_{j=M-N}^{M-1} \mathbf{Y}_j^n[k]\mathbf{W}_j[k] = \mathbf{F}\mathbf{k}^T\mathbf{e}_1[k]$$

$$\mathbf{E}[k] = \mathbf{F}\mathbf{k}^T\left(\mathbf{d}[k] - \mathbf{e}_1[k]\right) = \mathbf{F}\mathbf{k}^T\mathbf{e}[k]$$

$$\mathbf{E}_2[k] = \mathbf{F}\mathbf{k}^T\mathbf{k}\mathbf{F}^{-1} \sum_{j=M-N}^{M-1} \mathbf{Y}_j[k]\mathbf{W}_j[k] = \mathbf{F}\mathbf{k}^T\mathbf{e}_2[k]$$

and

$$\boldsymbol{\rho}[k] = \frac{2\rho'}{L}\text{diag}\ \left\{ P_0^{-1}[k],\ ...,\ P_{2L-1}^{-1}[k] \right\}$$

$$P_m[k] = \gamma P_m[k-1] + (1-\gamma)\left(P_{1,m}[k] + P_{2,m}[k]\right)$$

$$P_{1,m}[k] = \sum_{j=M-N}^{M-1} \left| Y_{j,m}^n[k] \right|^2$$

$$P_{2,m}[k] = \lambda P_{2,m}[k-1] + (1-\lambda)\frac{1}{\mu}\left| \sum_{j=M-N}^{M-1} \left(|Y_{j,m}[k]|^2 - \left|Y_{j,m}^n[k]\right|^2 \right) \right|$$

Output $\mathbf{z}[k] = \begin{bmatrix} z[kL] & \cdots & z[kL+L-1] \end{bmatrix}^T$ *:* $\mathbf{y}_0[k] = \begin{bmatrix} y_0[kL-\Delta] & \cdots & y_0[kL-\Delta+L-1] \end{bmatrix}^T$

- If noise detected: $\mathbf{z}[k] = \mathbf{y}_0[k] - \mathbf{e}_1[k]$
- If speech detected: $\mathbf{z}[k] = \mathbf{y}_0[k] - \mathbf{e}_2[k]$