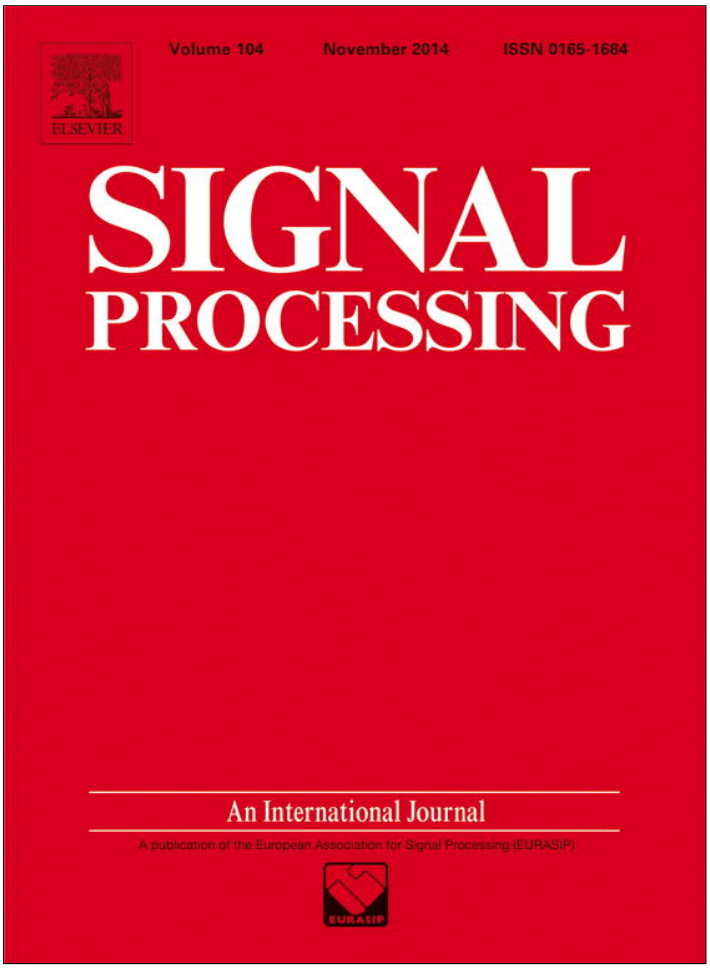
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.




This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/authorsrights

Author's personal copy

Signal Processing 104 (2014) 1–14

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

# Wiener variable step size and gradient spectral variance smoothing for double-talk-robust acoustic echo cancellation and acoustic feedback cancellation ☆

Jose M. Gil-Cacho [a,*], Toon van Waterschoot [a], Marc Moonen [a], Søren Holdt Jensen [b]

[a] KU Leuven, Department of Electrical Engineering ESAT-STADIUS, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[b] Department of Electronic Systems Aalborg University, Fredrik Bajers Vej 7, DK-9220 Aalborg, Denmark

ARTICLE INFO

*Article history:*
Received 13 June 2013
Received in revised form
6 March 2014
Accepted 14 March 2014
Available online 3 April 2014

*Keywords:*
Acoustic echo cancellation
Acoustic feedback cancellation
Adaptive filtering
Wiener variable step size
Prediction error method
Gradient smoothing
Double-talk

ABSTRACT

Double-talk (DT)-robust acoustic echo cancellation (AEC) and acoustic feedback cancellation (AFC) are needed in speech communication systems, e.g., in hands-free communication systems and hearing aids. In this paper, we derive a practical and computationally efficient algorithm based on the frequency-domain adaptive filter prediction error method using row operations (FDAF-PEM-AFROW) for DT-robust AEC and AFC. The proposed algorithm features two main modifications: (a) the Wiener variable step size (WVSS) and (b) the gradient spectral variance smoothing (GSVS). In AEC simulations, the WVSS-GSVS-FDAF-PEM-AFROW algorithm obtains outstanding robustness and smooth adaptation in highly adverse scenarios such as in bursting DT at high levels, and in a change of acoustic path during continuous DT. Similarly, in AFC simulations, the algorithm outperforms state-of-the-art algorithms when using a low-order near-end speech model and in colored non-stationary noise.

© 2014 Elsevier B.V. All rights reserved.

## 1. Introduction

Acoustic echo and acoustic feedback in speech communication systems are two well-known problems, which are caused by the acoustic coupling between a loudspeaker and a microphone. On one hand, acoustic echo cancellation (AEC) is widely used in mobile and hands-free telephony [1] where the existence of echoes degrades

☆ This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven Research Council CoE PFV/10/002 (OPTEC), KU Leuven Research Council Bilateral Scientific Cooperation Project Tsinghua University 2012–2014, Concerted Research Action GOA-MaNet, the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P7/19 "Dynamical systems control and optimization" (DYSCO) 2012–2017 and IUAP P7/23 "Belgian network on stochastic modeling analysis design and optimization of communication systems" (BESTCOM) 2012–2017, Flemish Government iMinds 2013, Research Project FWO nr. G.0763.12 "Wireless Acoustic Sensor Networks for Extended Auditory Communication", Research Project FWO nr. G.091213 "Cross-layer optimization with real-time adaptive dynamic spectrum management for fourth generation broadband access networks", Research Project FWO nr. G.066213 'Objective mapping of cochlear implants', the FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech" (DREAMS), funded by the European Commission under Grant Agreement no. 316969, IWT Project "Signal processing and automatic fitting for next generation cochlear implants". EC-FP6 project "Core Signal Processing Training Program" (SIGNAL) and was supported by a Postdoctoral Fellowship of the Research Foundation Flanders (FWO-Vlaanderen, T. van Waterschoot). The scientific responsibility is assumed by its authors.

* Corresponding author.
*E-mail address:* pepegilcacholorenzo@gmail.com (J.M. Gil-Cacho).

http://dx.doi.org/10.1016/j.sigpro.2014.03.020
0165-1684/© 2014 Elsevier B.V. All rights reserved.

Author's personal copy



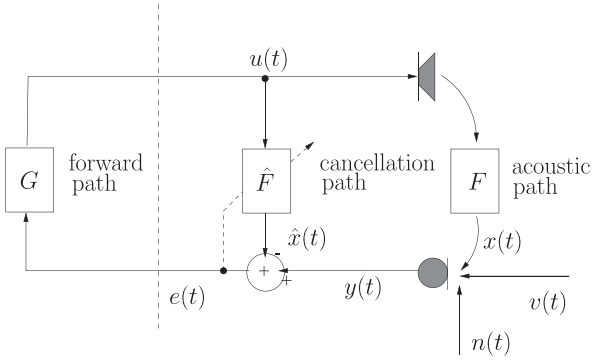



**Fig. 1.** Typical set-ups for AEC/AFC. The left part (forward path) only relates to AFC. The right part relates to both AEC and AFC.

the intelligibility and listening comfort. On the other hand, acoustic feedback limits the maximum amplification that can be applied, e.g., in a hearing aid, before howling due to instability appears [2,3]. The maximum attainable amplification may be too small to compensate for the hearing loss, which makes acoustic feedback cancellation (AFC) an important component in hearing aids. Fig. 1 shows a typical set-up for AEC and AFC. The goal of AEC and AFC is essential to identify a model of the echo or feedback path $F(q,t)$, i.e., the room impulse response (RIR), and to produce an estimate of the echo or feedback signal, which is then subtracted from the microphone signal $y(t)$. The microphone signal is given by $y(t)=x(t)+v(t)+n(t)=F(q,t)u(t)+v(t)+n(t)$ where $q$ denotes the time shift operator, e.g., $q^{-k}u(t)=u(t-k)$, $t$ is the discrete time variable, $x(t)$ is the echo or feedback signal, $u(t)$ is the loudspeaker signal, $v(t)$ is the near-end speech, and $n(t)$ is the near-end noise. In the sequel, we will use the term *near-end signal* to refer to $v(t)$ and/or $n(t)$ if, according to the context, there is no need to point out a difference. The operator $F(q,t)=f_0(t)+f_1(t)q^{-1}+\cdots+f_{n_F}(t)q^{-n_F}$ represents a linear time-varying model of the RIR between the loudspeaker and the microphone, where $n_F$ is the RIR model order. The aim of AEC or AFC is to obtain an estimate $\hat{F}(q,t)$ of the RIR model $F(q,t)$ by means of an adaptive filter, which is steered by the error signal $e(t)=[F(q,t)-\hat{F}(q,t)]u(t)+v(t)+n(t)$.

In AEC applications, the loudspeaker signal $u(t)$ is considered to be the signal coming from the far-end side, i.e., the far-end signal. The *echo-compensated* error signal $e(t)$ is then transmitted to the far-end side. On the other hand, in AFC applications, the forward path $G(q,t)$ maps the *feedback-compensated* error signal $e(t)$ to the loudspeaker signal, i.e., $u(t)=G(q,t)e(t)$. Typically, $G(q,t)$ consists of an amplifier with a (possibly time-varying) gain $K(t)$ cascaded with a linear equalization filter $J(q,t)$ such that $G(q,t)=J(q,t)K(t)$. AEC and AFC in principle look the same and share many common characteristics, however, different essential problems can be distinguished, as elaborated below.

### 1.1. Double-talk in acoustic echo cancellation

Practical AEC implementations rely on computationally-simple stochastic gradient algorithms, such as the normalized least mean squares (NLMS) algorithms, which may be very sensitive to the presence of a near-end signal [4]. Especially, near-end speech, in a so-called double-talk (DT) scenario, will affect adaptation in the AEC context by making the adaptive filter converge slowly or even diverge.

To tackle the DT problem, adaptive filters have been equipped with DT detectors (DTDs) to switch off during DT periods. Since the Geigel algorithm [5] was proposed, several other DTD algorithms have been specifically designed for AEC applications [6–9]. However, in general, a DTD takes some time before the onset of a DT period is detected. Moreover, in AFC scenarios, as will be seen in the next section, the near-end speech is continuously present and then the use of a DTD becomes futile. Therefore, DT-robust algorithms without the need for a DTD are called for. DT-robustness may be achieved based on three approaches, namely, (1) by using a postfilter to suppress or enhance residual echo, (2) by using a variable step size to slow down adaption during DT, and (3) by prefiltering the loudspeaker and microphone signal with a decorrelation filter to minimize the RIR estimation variance.

The first approach consists of using a *postfilter*, which interplays with the AEC to suppress residual echo (and also to reduce near-end signals) based on signal enhancement techniques [10–12]. On the other hand, in [13,14], the idea behind the postfilter design is the opposite, i.e., to enhance the residual echo in the adaptive filter loop. The postfilter design, in any case, is typically based on single-channel noise-reduction techniques, which carry a trade-off between residual echo/noise reduction and signal distortion.

The second approach to DT-robust AEC is to equip the (stochastic gradient) adaptive filter with a *variable step size* (VSS), which may be derived using information about the gradient vector or the near-end signal power. The first type of VSS algorithms relies on two properties of the gradient vector to control the step size [15–19]: (1) the property that the norm of the gradient vector will be large initially and converge to a small value, ideally zero, at steady state and (2) the property that the gradient vector direction will generally show a consistent trend during initial convergence, in contrast to a random trend around the optimal value during DT and at steady state. From this class of algorithms, the only one specifically designed for DT-robust AEC is the *projection-correlation* VSS (PC-VSS), which has been proposed in [20]. PC-VSS is a VSS algorithm based on the affine projection algorithm (APA) [21], where the adaptation rate is controlled by a measure of the correlation between instantaneous and long-term averages of the so-called *projection vectors*, i.e., gradient vectors in APA, which allows one to achieve robustness and to distinguish between an echo path change and DT. PC-VSS is chosen as one of the competing algorithms in this paper and hence further explanation will be given in Section 4.

Recently, more effort has been spent to steer VSS algorithm design towards DT-robust AEC. Some of these recent algorithms are based on the *non-parametric* VSS (NPVSS) algorithm proposed in [22]. The NPVSS algorithm was developed in a system identification context, aiming to recover the system noise (i.e., near-end noise) from the error signal of the adaptive filter when updated by the NLMS algorithm. Inspired by this idea, several approaches have focused on applying the NPVSS algorithm to real AEC applications where the microphone signal also contains near-end speech. Consequently, different VSS-NLMS algorithms have been

Author's personal copy



successfully developed for DT-robust AEC, e.g., [23,24]. However, their convergence is slow in practice, and hence, an APA version of the VSS-NLMS algorithm in [23] has been proposed in [25] to increase the convergence speed. The resulting *practical* VSS affine projection algorithm (PVSS) [25] is chosen as one of the competing algorithms in this paper and will be further explained also in Section 4.

The third approach is to search for the optimal AEC solution in a minimum-variance linear estimation framework, rather than in a traditional least squares (LS) framework. The minimum-variance echo path estimate, which is also known as the *best linear unbiased estimate* (BLUE) [26], depends on the near-end signal characteristics, which are in practice unknown and time-varying [4,27]. The algorithms in [4,27] aim to whiten the near-end speech component in the microphone signal by using adaptive decorrelation filters that are estimated concurrently with the acoustic echo path. In order to achieve the BLUE, it is also necessary to add a scaled version of the near-end speech excitation signal variance to the denominator of the stochastic gradient update equation. The use of the prediction error method (PEM) approach [28] was proposed to jointly estimate the RIR and an autoregressive (AR) model of the near-end speech. Among the PEM-based algorithms proposed in [4,27], the PEM-based adaptive filtering using row operations (PEM-AFROW) [29] is particularly interesting because it efficiently uses the Levison–Durbin algorithm to estimate both the near-end speech AR model coefficients and the near-end speech excitation signal variance. Thus, the algorithms in [4,27] can be seen as belonging to a new family aiming at both reducing the correlation between the near-end speech and loudspeaker signal, and minimizing the RIR estimation variance.

### 1.2. Correlation in acoustic feedback cancellation

In the AFC set-up, the near-end speech will be continuously present, so using a DTD is pointless. However, the main problem in AFC is the correlation that exists between the near-end speech component in the microphone signal and the loudspeaker signal itself. This correlation problem, which is caused by the closed loop, makes standard adaptive filtering algorithms converge to a *biased* solution [2,30]. This means that the adaptive filter not only predicts and cancels the feedback component in the microphone signal, but also cancels part of the near-end speech. This generally results in a distorted *feedback-compensated* error signal. One approach to reduce the bias in the feedback path model identification is to prefilter the loudspeaker and microphone signal with the inverse near-end speech model, which is estimated jointly with the adaptive filter [2,30] using the PEM [28]. For a near-end speech signal, an AR model is commonly used [2] as this yields a simple finite impulse response (FIR) prefilter. However, the AR model fails to remove the speech periodicity, which causes the prefiltered loudspeaker signal to still be correlated with the prefiltered near-end speech signal during voiced speech segments. More advanced models using different cascaded near-end speech models have been proposed to remove the coloring and periodicity in voiced as well as unvoiced speech segments. The constrained pole-zero linear prediction (CPZLP) model [31], the pitch prediction model [3], and the sinusoidal model [32] are examples of alternative models used in recently proposed algorithms. However, the overall algorithm complexity typically increases significantly when using cascaded near-end speech models [33]. In [34], a transform-domain PEM-AFROW algorithm using DFT (DFT-PEM-AFROW) has been proposed to improve the performance of an AFC without the need for cascaded and computationally intensive near-end signal models. Significant improvement was achieved w.r.t. standard PEM-AFROW even using low-order AR models. PEM-AFROW and DFT-PEM-AFROW are chosen as competing algorithms for AFC. DFT-PEM-AFROW is also chosen as a competing algorithm for AEC.

### 1.3. Contributions and outline

In [35], we have proposed the use of the FDAF-PEM-AFROW framework to improve several VSS and variable regularization (VR) algorithms. The improvement is basically due to two aspects: (1) the *instantaneous pseudo-correlation* (IPC) [35] between the near-end signal and the far-end signal is heavily reduced when using FDAF-PEM-AFROW compared to the (time-domain) PEM-AFROW and (2) FDAF itself may be seen to minimize a BLUE criterion if a proper normalization factor is used during adaptation [36]. In this paper, we propose two modifications of the FDAF-PEM-AFROW algorithm for robust and smooth adaptation in both AFC and AEC with continuous and bursting DT without the need of a DTD. In particular, we propose the *Wiener variable step size* (WVSS) and the *gradient spectral variance smoothing* (GSVS) to be performed in FDAF-PEM-AFROW, leading to the WVSS-GSVS-FDAF-PEM-AFROW algorithm. The WVSS modification is implemented as a single-channel noise-reduction Wiener filter applied to the (prefiltered) microphone signal. The *Wiener filter gain* [12] is used as a VSS in the adaptive filter, rather than as a signal enhancement parameter. On the other hand, the GSVS modification aims at reducing the variance of the noisy gradient estimates based on time-recursive averaging of instantaneous gradients. Combining the WVSS and GSVS with the FDAF-PEM-AFROW algorithm consequently gathers the best characteristics we are seeking in an algorithm for both AEC and AFC, namely, decorrelation properties (PEM, FDAF), minimum variance (GSVS, FDAF, PEM), variable step size (WVSS), and computational efficiency (FDAF).

The outline of the paper is as follows. In Section 2, we briefly present the PEM, we provide a simple algorithm description and explain the choice of the near-end speech model. In Section 3, the proposed algorithm is presented with in-depth explanations about the novel algorithm modifications. The motivation for including these modifications is also justified for DT-robust AEC and AFC. In Section 4, computer simulation results are provided to verify the performance of the proposed algorithm compared to three competing algorithms, in particular, the PC-VSS [20], the PVSS [25], and the DFT-PEM-AFROW [34]. A description of the competing algorithms is provided together with a computational complexity analysis. The

Author's personal copy



Matlab files implementing all the algorithms and those generating the figures in Sections 3–4 can be found in [37]. Finally, Section 5 concludes the paper.

## 2. Prediction error method

The PEM-based AEC/AFC is shown in Fig. 2. It relies on a linear model for the near-end speech $v(t)$, which in Fig. 2 is specified as

$$v(t) = H(q,t)w(t), \tag{1}$$

where $H(q,t)$ contains the filter coefficients of the linear model and $w(t)$ represents the excitation signal, which is assumed to be white noise with time-dependent variance $\sigma_w^2(t)$, i.e.,

$$\mathcal{E}\{w(t)w(t-k)\} = \sigma_w^2(t)\delta(k), \tag{2}$$

where $\mathcal{E}\{\cdot\}$ is the expected value operator. The near-end noise $n(t)$ is also assumed to be a white noise signal for the time being. As outlined before, a minimum-variance echo path model (in AEC) or an unbiased feedback path model (in AFC) can be identified by first prefiltering the loudspeaker signal $u(t)$ and the microphone signal $y(t)$ with the inverse near-end speech model $H^{-1}(q,t)$ before feeding these signals to the adaptive filtering algorithm. As $H^{-1}(q,t)$ is obviously unknown, the near-end speech model and the echo/feedback path model have to be jointly identified using the PEM [28]. A common approach in PEM-based AEC/AFC is to model the near-end speech with an AR model, i.e.,

$$y(t) = x(t) + v(t) + n(t) \tag{3}$$

$$y(t) = F(q,t)u(t) + \frac{1}{A(q,t)}w(t) + n(t), \tag{4}$$

with $F(q,t)$ defined previously and $A(q,t)$ given as

$$A(q,t) = 1 + a_1(t)q^{-1} + \cdots + a_{n_A}(t)q^{-n_A}, \tag{5}$$

where $n_A$ is the AR model order.

The PEM gives an estimate of the models $F(q,t)$ and $A(q,t)$ by minimization of the prediction error criterion

$$\hat{\vartheta}(t) = \arg\min_{\vartheta(t)} \sum_{i=1}^{t} e_a^2[i,\vartheta(t)] \tag{6}$$

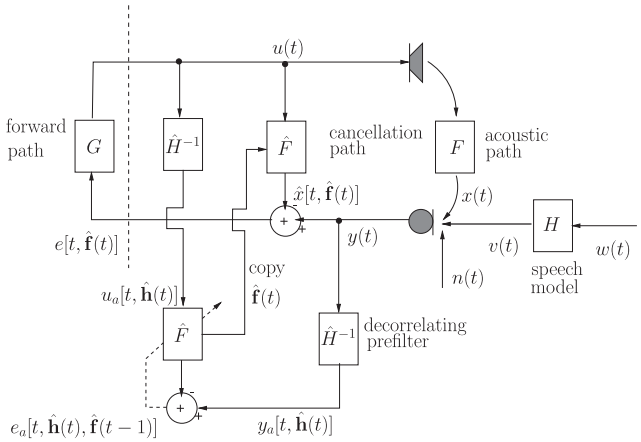



**Fig. 2.** AEC/AFC with prefiltering of the loudspeaker and microphone signals using the inverse $[H^{-1}(q,t)]$ of the near-end speech model $[H(q,t)]$.

where the prediction error is defined as

$$e_a[t,\vartheta(t)] = A(q,t)[y(t) - F(q,t)u(t)], \tag{7}$$

and the parameter vector $\vartheta(t) = [\mathbf{f}^T(t), \mathbf{a}^T(t)]^T$ contains the parameters of the echo or feedback path model and the near-end speech model, i.e.,

$$\mathbf{f}(t) = [f_0(t), f_1(t), \ldots, f_{n_F}(t)]^T, \tag{8}$$

$$\mathbf{a}(t) = [1, a_1(t), \ldots, a_{n_A}(t)]^T. \tag{9}$$

Note that throughout the paper, we assume a sufficient-order condition for the acoustic path model (i.e., $n_{\hat{F}} = n_F$). An additional assumption is that the near-end speech $v(t)$ is short-term stationary, which implies that the near-end speech model $A(q,t)$ does not need to be re-estimated at each time instant $t$. That is, instead of identifying the near-end speech model recursively, it can also be identified non-recursively on a batch of loudspeaker and microphone data. This is the idea behind the PEM-AFROW algorithm, which estimates $A(q,t)$ in a block-based manner using a block length that approximates the stationary interval of speech. The PEM-AFROW algorithm was originally developed in an AFC framework [29] and applied to a continuous-DT AEC scenario in [27]. It performs only row operations on the loudspeaker data matrix, hence the name PEM-AFROW, and both $\hat{\mathbf{a}}(t)$ and $\hat{\sigma}_w^2(t)$ are efficiently calculated using the Levinson–Durbin recursion. For a detailed description of the original PEM-AFROW algorithm the reader is referred to [29].

**Algorithm 1.1.** First part of the WVSS-GSVS-FDAF-PEM-AFROW algorithm, showing the FDAF-PEM-AFROW. Lines within brackets correspond to the data generation and are not part of the algorithm.

1: Initialize: $K$, $k=0$, and $\hat{\mathbf{P}}_{U_a} = \hat{\mathbf{P}}_{X_a} = \hat{\mathbf{P}}_{D_a} = \hat{\mathbf{F}} = \nabla = \mathbf{0}_{M\times 1}$
2: $[\mathbf{U}(k) = \mathcal{F}\{\mathbf{u}(k)\}]$
3: $\hat{\mathbf{x}}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k) \odot \hat{\mathbf{F}}(k-1)\}$ (Echo estimation)
4: $[\mathbf{x}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k) \odot \mathbf{F}\}$(True echo signal simulation)$]$
5: $[\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k) + \mathbf{n}(k)$(Microphone signal simulation)$]$
6: $\mathbf{e}(k) = [\mathbf{y}(k) - \hat{\mathbf{x}}(k)]_{N+1:M}$ (Error signal)
7: $[\mathbf{u}(k+1) = K \odot \mathbf{e}(k)$(Loudspeaker signal simulation, only for AFC)$]$
8: **for** $k = 1, 2, \ldots$ **do**
9: $[\mathbf{U}(k) = \mathcal{F}\{\mathbf{u}(k)\}]$
10: $\hat{\mathbf{x}}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k) \odot \hat{\mathbf{F}}(k-1)\}$ (Echo estimation)
11: $[\mathbf{x}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k) \odot \mathbf{F}\}$ (True echo signal simulation)$]$
12: $[\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k) + \mathbf{n}(k)$ (Microphone signal simulation)$]$
13: $\mathbf{e}(k) = [\mathbf{y}(k) - \hat{\mathbf{x}}(k)]_{N+1:M}$
14: $[\mathbf{u}(k+1) = K \odot \mathbf{e}(k)$ (Loudspeaker signal simulation. Only for AFC)$]$
15: $\hat{\mathbf{a}}(k) = \mathrm{AR}\{[\mathbf{e}^T(k)\ \mathbf{e}^T(k-1)]_{1:P}^T, n_A\}$ (Order $n_A$ AR coefficients estimation)
16: **for** $m = 0, \ldots, M-1$ do (Decorrelation prefilter)
17: $u_a(m,k) = [u(kN+1+m), \ldots, u(kN+1+m-n_A)]\mathbf{a}(k)$
18: $y_a(m,k) = [y(kN+1+m), \ldots, y(kN+1+m-n_A)]\mathbf{a}(k)$
19: **end for**
20: $\mathbf{U}_a(k) = \mathcal{F}\{\mathbf{u}_a(k)\}$
21: $\hat{\mathbf{x}}_a(k) = \mathcal{F}^{-1}\{\mathbf{U}_a(k) \odot \hat{\mathbf{F}}(k-1)\}$
22: $\mathbf{e}_a(k) = [\mathbf{y}_a(k) - \hat{\mathbf{x}}_a(k)]_{N+1:M}$ [(Prediction) Error signal (7)]
23: $\mathbf{E}_a(k) = \mathcal{F}\{[\mathbf{0}_N \quad \mathbf{e}_a^T(k)]^T\}$
(Continued in Algorithm 1.2.)

## 3. WVSS-GSVS modifications to FDAF-PEM-AFROW

The WVSS-GSVS-FDAF-PEM-AFROW for AEC/AFC is given in Algorithm 1 (Parts 1.1 and 1.2). The FDAF

Author's personal copy



implementation corresponds to the overlap-save FDAF with gradient constraint and power normalization [38], where $\mathbf{u}(k)$, $\mathbf{v}(k)$, and $\mathbf{n}(k)$ are length-$M$ vectors, with $M=2N$ and $N=n_F+1$, satisfying the overlap-save condition $\mathbf{u}(k)=[u(kN-N+1),\ldots,u(kN+N)]^T$, where $k=0,1,\ldots,(L-N)/N$ is the block-time index, $L$ is the total length of the signals, $[\cdot]_{a:b}$ represents a range of samples within a vector, and $P$ is the block length used to estimate $A(q,t)$, with $N\le P\le 2N$. The subscript $m$ in $\omega_m$, e.g., $E_a(\omega_m,k)$, refers to a signal in the $m$th frequency bin of a block, $m=0,\ldots,M-1$, $\odot$ represents component-wise (Hadamard) multiplication, and a capital bold-face variable, e.g., $\mathbf{E}_a(k)$, denotes an $M$-dimensional vector of frequency components. The subscript $a$, e.g., $E_a(\omega_m,k)$, denotes a signal output from the decorrelating prefilter, as shown in Fig. 2. Lines within brackets correspond to the data generation and are not truly part of the algorithm. The specific WVSS-GSVS modifications within the FDAF-PEM-AFROW algorithm are shown in Algorithm 1.2. The proposed WVSS-GSVS-FDAF-PEM-AFROW weight update equation is given by

$$\boldsymbol{\phi}(k)=[\mathcal{F}^{-1}\{\boldsymbol{\mu}(k)\odot\mathbf{W}(k)\odot\boldsymbol{\Theta}(k)\}]_{1:N} \tag{10}$$

$$\hat{\mathbf{F}}(k)=\hat{\mathbf{F}}(k-1)+\mu_{\max}\mathcal{F}\{[\boldsymbol{\phi}^T(k)\ \mathbf{0}_N]^T\}, \tag{11}$$

where $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ denote the $M$-point discrete Fourier transform (DFT) and inverse DFT (IDFT), respectively, $\boldsymbol{\mu}(k)$ corresponds to the power normalization step typically included in an FDAF update, $\mu_{\max}$ sets the maximum allowed value of the step size, and $\mathbf{W}(k)$ together with $\boldsymbol{\Theta}(k)$ forms the WVSS-GSVS modifications that will be explained in detail in the following sections.

### 3.1. Wiener variable step size

The Wiener variable step size (WVSS) modification to the FDAF-PEM-AFROW algorithm introduces a frequency-domain variable step size. Basically, the goal is to slow down the adaptation in frequency bins where the *echo-to-near-end-signal ratio* (ENR) is low, and increase it in those frequency bins where the ENR is high. If we recall that the near-end signal consists of the near-end speech and near-end noise, then we can calculate ENR $(t)=\sigma_x^2(t)/[\sigma_v^2(t)+\sigma_n^2(t)]$, where $\sigma_x^2(t)$, $\sigma_v^2(t)$, and $\sigma_n^2(t)$ are the variance of the echo, the near-end speech and the near-end noise, respectively. In the *near-end-signal-free* case, i.e. $v(t)=n(t)=0$, the microphone signal consists only of the echo, so one could apply the maximum step size in each frequency bin. Once a near-end signal is present in the microphone signal, and especially so if it is colored and non-stationary, the step sizes in the different frequency bins should be reduced accordingly.

The concept for deriving the WVSS is to apply a single-channel frequency-domain noise-reduction Wiener filter to the microphone signal $y(t)$. This may also be seen as an *echo-enhancement* filter, however, we do not explicitly use the output of the filter itself, but rather, use the Wiener filter gain as a variable step size in the adaptive filer. Thus, the step size in each frequency bin is varied by the gain of the Wiener filter at that frequency bin.

**Algorithm 1.2.** Second part of the WVSS-GSVS-FDAF-PEM-AFROW algorithm showing the WVSS-GSVS modifications.

**for** $m=0,\ldots,M-1$ **do**
  $\hat{P}_{U_a}(\omega_m,k)=\lambda_0\hat{P}_{U_a}(\omega_m,k-1)+(1-\lambda_0)|U_a(\omega_m,k)|^2$ (Recursive power estimation)
  $\mu(\omega_m,k)=[\hat{P}_{U_a}(\omega_m,k)+\delta]^{-1}$ (Power normalization)
  $\theta(\omega_m,k)=E_a(\omega_m,k)U_a^*(\omega_m,k)$ (Gradient estimation)
  $\nabla(\omega_m,k)=\lambda_3\nabla(\omega_m,k-1)+(1-\lambda_3)|\theta(\omega_m,k)|^2$
  $\alpha(\omega_m,k)=\angle\theta(\omega_m,k)$ (Phase estimation)
  $\Theta(\omega_m,k)=\sqrt{\nabla(\omega_m,k)}e^{[j\alpha(\omega_m,k)]}$ (GSVS)
  $\hat{P}_{X_a}(\omega_m,k)=\lambda_1\hat{P}_{X_a}(\omega_m,k-1)+(1-\lambda_1)|\hat{X}_a(\omega_m,k)|^2$
  $\hat{P}_{D_a}(\omega_m,k)=\lambda_2\hat{P}_{D_a}(\omega_m,k-1)+(1-\lambda_2)|E_a(\omega_m,k)|^2$
  $W(\omega_m,k)=\dfrac{\sqrt{\hat{P}_{X_a}(\omega_m,k)}}{\sqrt{\hat{P}_{X_a}(\omega_m,k)}+\hat{P}_{D_a}(\omega_m,k)}$ (WVSS)
**end for**
$\boldsymbol{\phi}(k)=[\mathcal{F}^{-1}\{\boldsymbol{\mu}(k)\odot\mathbf{W}(k)\odot\boldsymbol{\Theta}(k)\}]_{1:N}$
$\hat{\mathbf{F}}(k)=\hat{\mathbf{F}}(k-1)+\mu_{\max}\mathcal{F}\{[\boldsymbol{\phi}^T(k)\ \mathbf{0}_N]^T\}$
**end for**

We assume that the signals are wide-sense stationary, that we have access to the full record of samples, and that disjoint frequency bins can be considered uncorrelated [39]. So, without loss of generality, we may consider a single frequency bin and work with the $m$-dependency. The frequency-domain microphone signal (3) is

$$\begin{aligned}Y(\omega_m,k)&=X(\omega_m,k)+V(\omega_m,k)+N(\omega_m,k)\\&=X(\omega_m,k)+D(\omega_m,k)\end{aligned}$$

where $X(\omega_m,k)$ is the *desired* signal and $D(\omega_m,k)$ is the *noise* signal, which is to be removed from the microphone signal. An estimate of the desired signal may be obtained as

$$\hat{X}(\omega_m,k)=W_0(\omega_m,k)Y(\omega_m,k), \tag{12}$$

which gives the (theoretical) frequency-domain Wiener filter gain as

$$W_0(\omega_m,k)=\frac{P_{XY}(\omega_m,k)}{P_Y(\omega_m,k)} \tag{13}$$

where $P_Y(\omega_m,k)=\mathcal{E}\{Y(\omega_m,k)Y^*(\omega_m,k)\}$ and $P_{XY}=\mathcal{E}\{X(\omega_m,k)Y^*(\omega_m,k)\}$ are the power spectral density (PSD) of $Y(\omega_m,k)$ and the cross-power spectral density (CPSD) of $X(\omega_m,k)$ and $Y(\omega_m,k)$, respectively, and the upper asterisk $(\cdot)^*$ denotes complex conjugation. A common assumption in single-channel noise-reduction algorithms is that the desired signal $X(\omega_m,k)$ is uncorrelated with the noise component $D(\omega_m,k)$, so that the numerator of the Wiener filter results in $P_{XY}(\omega_m,k)=P_X(\omega_m,k)$ and the denominator becomes $P_Y(\omega_m,k)=P_X(\omega_m,k)+P_D(\omega_m,k)$, which gives the (theoretical) frequency-domain Wiener filter gain as

$$W_0(\omega_m,k)=\frac{P_X(\omega_m,k)}{P_X(\omega_m,k)+P_D(\omega_m,k)}. \tag{14}$$

In order to obtain (14), we have assumed that $X(\omega_m,k)$ and $D(\omega_m,k)$ are uncorrelated. Usually, the far-end speech and the near-end speech are indeed statistically uncorrelated, which however does not imply that the IPC between these two signals is zero. The (rather strong) assumption,

Author's personal copy



of the near-end signal being uncorrelated with the loudspeaker signal is also adopted in most AEC applications. On the other hand, in AFC applications, this assumption is clearly violated, as explained in Section 1. In [35], we have shown that the IPC between the far-end and the near-end speech may be very large. Besides, the practical computation of (14) is generally based on the time-recursive estimate of the PSDs using short-time Fourier transforms (SFTs) of $X(\omega_m,k)$ and $D(\omega_m,k)$. All this means that, in real AEC and AFC, the practical computation of (14) would produce misleading gain values because the assumption made is potentially not true when considering IPC.

However, we have shown in [35] that the IPC between the far-end and the near-end speech can be significantly reduced by using the FDAF-PEM-AFROW framework. Consequently, we will consider deriving the Wiener filter gains using the filtered signals (see Fig. 2), so as to apply them in both real AEC and AFC applications. Then, the *desired* signal is the filtered echo signal $x_a(t)$ and the filtered near-end signals, grouped in $d_a(t)=v_a(t)+n_a(t)$, are considered as the *noise* component in the microphone signal $y_a(t)$. The frequency-domain Wiener filter using whitened signals will moreover result in a filter similar to (14), since $P_{X_aY_a}(\omega_m,k)=\mathcal{E}\{A(\omega_m,k)X(\omega_m,k)Y^*(\omega_m,k)A^*(\omega_m,k)\}=|A(\omega_m,k)|^2P_{XY}$, and $P_{Y_a}(\omega_m,k)=\mathcal{E}\{A(\omega_m,k)Y(\omega_m,k)Y^*(\omega_m,k)A^*(\omega_m,k)\}=|A(\omega_m,k)|^2P_Y$, so that

$$W_0(\omega_m,k)=\frac{P_{X_aY_a}(\omega_m,k)}{P_{Y_a}(\omega_m,k)}=\frac{P_{X_a}(\omega_m,k)}{P_{X_a}(\omega_m,k)+P_{D_a}(\omega_m,k)}. \tag{15}$$

We will instead compute time-recursive estimates of the PSDs and use available estimates of the desired signal $\hat{X}_a(\omega_m,k)$ and of the noise component $E_a(\omega_m,k)$ in the microphone signal. The Wiener filter gains are thus efficiently estimated as

$m=0,\ldots,M-1$

$$\hat{P}_{X_a}(\omega_m,k)=\lambda_1\hat{P}_{X_a}(\omega_m,k-1)+(1-\lambda_1)|\hat{X}_a(\omega_m,k)|^2 \tag{16}$$

$$\hat{P}_{D_a}(\omega_m,k)=\lambda_2\hat{P}_{D_a}(\omega_m,k-1)+(1-\lambda_2)|E_a(\omega_m,k)|^2. \tag{17}$$

Finally, using (16) and (17), we can write

$$W(\omega_m,k)=\frac{\hat{P}_{X_a}(\omega_m,k)}{\hat{P}_{X_a}}(\omega_m,k)+\hat{P}_{D_a}(\omega_m,k) \tag{18}$$

where $\hat{P}_{X_a}(\omega_m,k)$ is an estimate of the PSD of $X_a(\omega_m,k)$, which is calculated by (16) using the output of the adaptive filter $\hat{X}_a(\omega_m,k)=U_a(\omega_m,k)\hat{F}(\omega_m,k-1)$, and where $\hat{P}_{D_a}(\omega_m,k)$ is an estimate of the PSD of $D_a(\omega_m,k)$, which is calculated by (17) using the prediction-error signal $E_a(\omega_m,k)$.

An interpretation of the Wiener filter gain $W(\omega_m,k)$ can be given by writing (18) as

$$W(\omega_m,k)=\frac{\hat{\mathrm{ENR}}(\omega_m,k)}{\hat{\mathrm{ENR}}}(\omega_m,k)+1\,, \tag{19}$$

where $\hat{\mathrm{ENR}}(\omega_m,k)=\hat{P}_{X_a}(\omega_m,k)/\hat{P}_{D_a}(\omega_m,k)=\hat{P}_{X_a}(\omega_m,k)/[\hat{P}_{V_a}(\omega_m,k)+\hat{P}_{N_a}(\omega_m,k)]$ is an estimate of the ENR. The Wiener filter gain is a real positive number in the range $0\le W(\omega_m,k)\le 1$, and is used as a variable step size in (10). A maximum value for the step size is also adopted, so that the effective step size is $\mu_{\max}W(\omega_m,k)$. Let us now consider the two limiting cases: (1) an "echo-only" microphone signal, $\hat{\mathrm{ENR}}(\omega_m,k)=\infty$ and (2) a "near-end-only" microphone signal, $\hat{\mathrm{ENR}}(\omega_m,k)=0$. In the first case, the Wiener filter gain $W(\omega_m,k)=1$, so the filter would apply the maximum step size in the $m$th frequency bin. In the second case, the Wiener filter gain $W(\omega_m,k)=0$, so the adaptation is suspended in the $m$th frequency bin. Between these two extreme cases, the Wiener filter gain reduces the step size in proportion to an estimate of the ENR in each frequency bin.

### 3.2. Gradient spectral variance smoothing

The gradient spectral variance smoothing (GSVS) is included to reduce the variance of the gradient estimation, in particular, the effect of sudden high-amplitude near-end signal samples. Although step-size control has been considered in Section 3.1, it may not be sufficiently fast to react on certain occasions, e.g., in DT bursts. In FDAF (-PEM-AFROW), an estimate of the gradient $\theta(\omega_m,k)=E_a(\omega_m,k)U_a^*(\omega_m,k)$ is calculated for every block of samples. This noisy estimate can be separated into two components as

$$\theta(\omega_m,k)=\theta_0(\omega_m,k)+\theta_{D_a}(\omega_m,k), \tag{20}$$

where $\theta_0(\omega_m,k)=[X(\omega_m,k)-\hat{X}(\omega_m,k)]U_a^*(\omega_m,k)$ is the *true gradient* and $\theta_{D_a}(\omega_m,k)=D_a(\omega_m,k)U_a^*(\omega_m,k)$ is the *gradient noise*.

The concept for deriving GSVS is that of applying averaged periodograms, which are typically used for estimating the PSD of a signal [40]. This concept would be sufficiently justified if the RIR does not change and the algorithm has converged to an optimum, therefore $\theta_0(\omega_m,k)=0$ by definition. In practice, of course, the true gradient $\theta_0(\omega_m,k)$ is time-varying. However, we assume that the true gradient varies slowly, so a low-pass filter (LPF) with a low cut-off frequency would be more beneficial than a simple average. Therefore, the realization of the GSVS is based on a time-recursive averaging of the gradient estimate, i.e., a LPF with a low cut-off frequency that will effectively filter out the gradient noise and reduce the variance in the gradient estimation. This is another way to obtain a first-order infinite impulse response (IIR) filtering, i.e.,

$$\nabla(\omega_m,k)=\lambda_3\nabla(\omega_m,k-1)+(1-\lambda_3)|\theta(\omega_m,k)|^2 \tag{21}$$

where $\lambda_3$ is the pole of the low-pass IIR filter, and

$$\theta(\omega_m,k)=E_a(\omega_m,k)U_a^*(\omega_m,k). \tag{22}$$

Finally, the phase is applied to form the GSVS as

$$\Theta(\omega_m,k)=\sqrt{\nabla(\omega_m,k)}e^{[j\alpha(\omega_m,k)]} \tag{23}$$

where

$$\alpha(\omega_m,k)=\angle\theta(\omega_m,k). \tag{24}$$

A practical simulation of GSVS for $m=15$, i.e., the 15th frequency bin, with $k=1,\ldots,200$ and $\lambda_3=0.95$ is shown in Fig. 3. For this simulation, we have used two speech signals, $u(t)$ and $v(t)$, and a noise signal $n(t)$, all sampled at 8 kHz. The AR coefficients of the near-end signal model are estimated using $d(t)=v(t)+n(t)$ and $n_A=55$. The signals $u(t)$ and $d(t)$ are then filtered, as in Algorithm 1.1

Author's personal copy



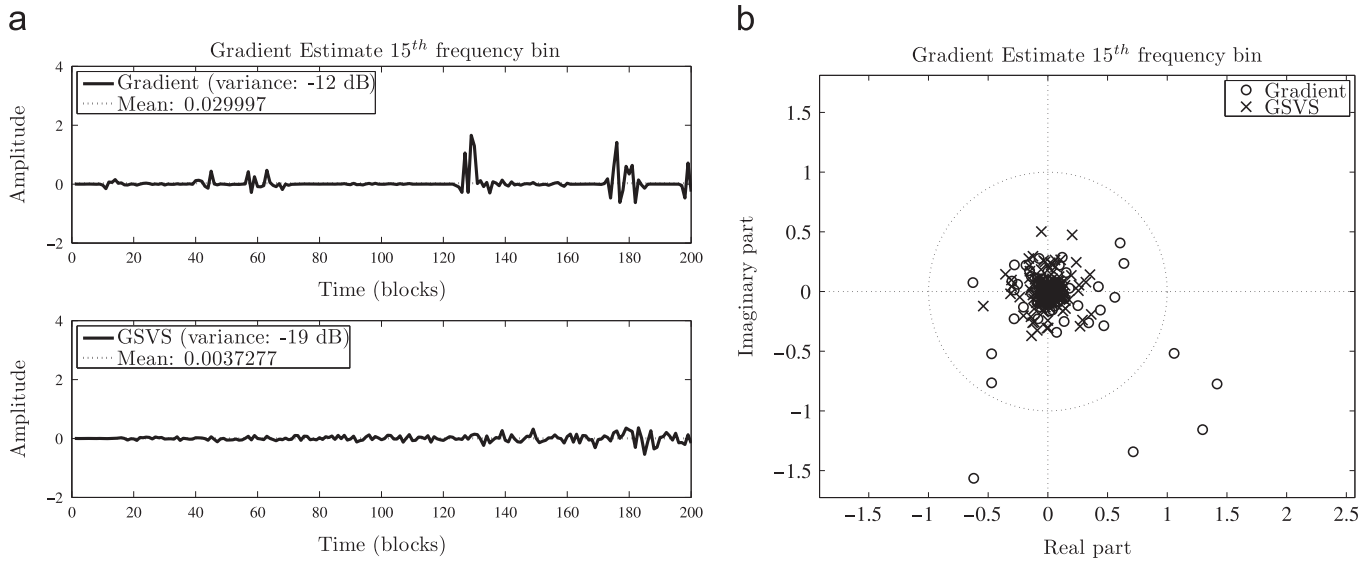



**Fig. 3.** Instantaneous gradient and GSVS estimates for $m=15$ and $k=1,\ldots,200$. (a) Time representation corresponding to the real part of $\theta(\omega_m,k)$ (upper) and $\Theta(\omega_m,k)$ (lower), respectively, showing the gradient estimate (solid line) and its mean value (dotted line). (b) Complex representation of $\theta(\omega_m,k)$ (circles) and $\Theta(\omega_m,k)$ (crosses).

lines 16–20, before feeding them to the overlap-save-type recursion. The gradient-constraint-type of calculations are performed, as in Algorithm 1.2 lines 34–35, to obtain both the noisy gradient and the GSVS gradient. We have assumed that the true gradient has converged to an optimum, and thus it is, by definition, equal to zero. In Fig. 3(a), a time-domain representation is shown, where the upper and lower figures correspond to the real part of $\theta(\omega_m,k)$ and $\Theta(\omega_m,k)$, respectively. Fig. 3(b) shows a complex representation, where the circles (○) represent the complex value of $\theta(\omega_m,k)$ and the crosses ( × ) represent the complex value of $\Theta(\omega_m,k)$. In both representations, the variance of the estimate is shown to be reduced by about 7 dB, the mean value is closer to zero, and high-amplitude samples are clearly smoothed.

## 4. Evaluation

Simulations are performed using speech signals sampled at 8 kHz. Two types of near-end noise $n(t)$ are used in the simulations, namely, white Gaussian noise (WGN) and *speech babble*, which is one of the most challenging noise types in signal enhancement applications [41]. We define two measures to determine the relative signal levels in the simulations: the *signal-to-echo ratio* $(\mathrm{SER}=10\log_{10}\sigma_x^2/\sigma_v^2$ and the *signal-to-noise ratio* $(\mathrm{SNR}=10\log_{10}\sigma_x^2/\sigma_n^2$, where $\sigma_x^2$, $\sigma_v^2$, and $\sigma_n^2$ are the variances of the echo, the near-end speech, and the near-end noise, respectively.

In the AEC simulations, the far-end (or loudspeaker) signal $u(t)$ is a female speech signal and the near-end speech is a male speech signal. The microphone signal consists of three concatenated segments of speech: the first and third 12-s segments correspond to a single-talk situation, i.e., $y(t)=x(t)+n(t)$, while the second 13-s segment corresponds to a DT situation, i.e., $y(t)=x(t)+n(t)+v(t)$. The AR model order in the AEC is $n_A=1$, and the APA order is $Q=4$. There are two RIRs used in the AEC simulations, namely, $\mathbf{f}_1$ and $\mathbf{f}_2$, which are shown in Fig. 4. Both 500-tap impulse responses have been measured in a room that is acoustically conditioned and prepared to have low reverberation time, but is not completely anechoic. In the AEC simulations, the WGN and speech babble noise types are set at different SNRs: 30 and 20 dB, respectively. Several SER values are used for the simulations: from mild (15 dB) to highly adverse (5 dB) DT conditions.

In the AFC simulations, the near-end speech is the same female speech signal as in the AEC simulations. Two different AR model orders are chosen as in [4]: $n_A=12$, which is common in speech coding for formant prediction, and $n_A=55$, which is high enough to capture all near-end signal dynamics. The forward path gain $K(t)$ is set 3 dB below the *maximum stable gain* (MSG) without feedback cancellation (details later in Section 4.2). A measured 100-tap acoustic impulse response was obtained from a real hearing aid and used to simulate the feedback path.

In both AEC and AFC, the window length $P$ was chosen to be 20 ms (160 samples), which corresponds to the average time interval in which speech is considered stationary.

### 4.1. Competing algorithms and tuning parameters

The AEC simulations are performed comparing four algorithms, namely, PC-VSS [20], PVSS [25], DFT-PEM-AFROW [34], and the proposed WVSS-GSVS-FDAF-PEM-AFROW. The PC-VSS algorithm belongs to the class of gradient-based VSS algorithms, and is claimed to feature the appealing ability to distinguish between echo path changes and DT. The PC-VSS algorithm is the result of improving the algorithm given in [15] to be specifically suited for AEC in DT situations. In the PC-VSS algorithm, the adaptation rate is controlled by a measure of the correlation between instantaneous and long-term averages of the so-called projection vectors, i.e., gradient vectors in APA. It appears that PC-VSS outperforms the algorithms given in [15,42] in DT situations. Moreover,

Author's personal copy



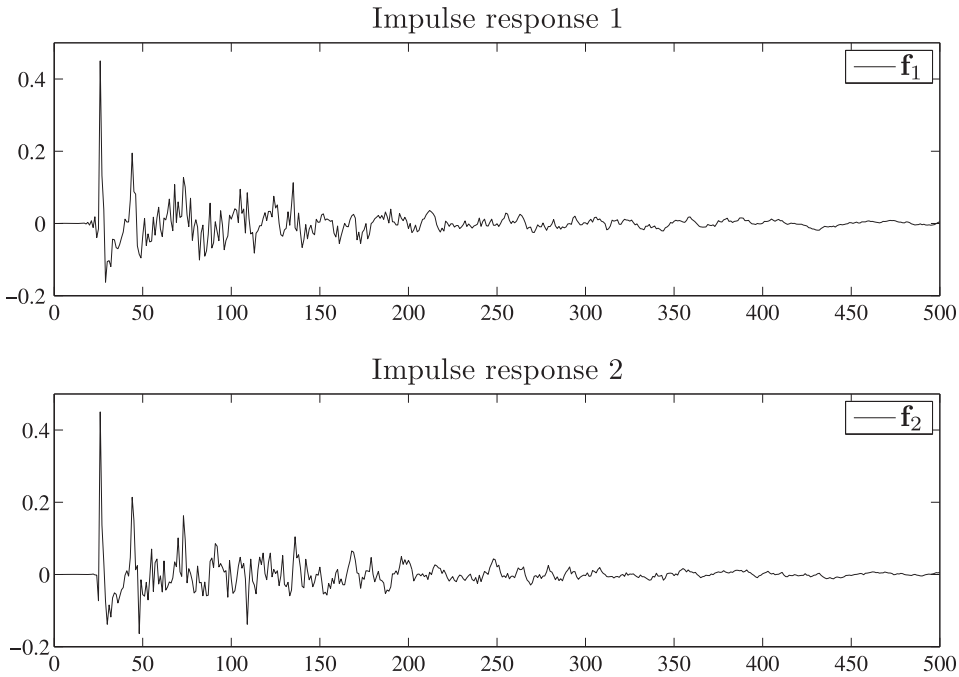



**Fig. 4.** Two RIRs used in the AEC simulations, namely, $\mathbf{f}_1$ and $\mathbf{f}_2$. Both 500-tap impulse responses have been measured in a room that is acoustically conditioned and prepared to have low reverberation time, but is not completely anechoic.

it does not rely on any signal or system model, so it is easy to control in practice. The PVSS algorithm [25] stems from the so-called NPVSS proposed in [22], and takes into account near-end signal power variations. PVSS is effectively used in DT situations, where it is claimed to be easy to control in practice, and has been shown to outperform the algorithms proposed in [23,43,44]. The DFT-PEM-AFROW algorithm has been investigated in terms of DT robustness in AEC and general improvement in AFC. It has been shown that the combination of prewhitening the input and microphone signals together with transform-domain filter adaptation leads to an algorithm that solves the problem of decorrelation in a very efficient manner. DFT-PEM-AFROW is very robust in DT situations and boosts the performance of even the simplest AFC, i.e., using only an AR model for the near-end signal.

All of the algorithms presented in this paper can be found in [37] along with the scripts generating every figure in this section. The tuning parameters in PVSS and PC-VSS are chosen according to the specifications given in [25] and [20], respectively. The parameters of DFT-PEM-AFROW are chosen to have an initial convergence curve similar to that of PVSS and PC-VSS. In the proposed WVSS-GSVS-FDAF-PEM-AFROW algorithm, the following parameter values are chosen to have similar initial convergence properties as the other algorithms: $\mu_{\max}=0.03$, $\delta=2.5e^{-6}$ , $\lambda_0=0.99$, $\lambda_1=0.1$, $\lambda_2=0.9$. The different $\lambda_1$ and $\lambda_2$ values in the time-recursive averaging for power estimation basically aim for having a longer averaging window for the near-end signal and a shorter averaging window for the echo signal. Note that, for higher robustness to near-end signals, a higher value of $\lambda_2$ may be chosen; however, convergence would then be slower.

AFC simulations are performed to compare the original PEM-AFROW [29], DFT-PEM-AFROW [34], and the proposed WVSS-GSVS-FDAF-PEM-AFROW. The parameters are tuned to have similar initial performance curves. In all three algorithms, the following common values are applied: the forward path $G(q,t)$ consists of a delay of 80 samples and a fixed gain $K(t)=K$, $\forall t$, resulting in a 3-dB MSG without AFC, and a window of $P=160$ samples is used for estimating the AR model. For WVSS-GSVS-FDAF-PEM-AFROW, $\lambda_0=\lambda_2=0.99$, $\lambda_1=0.1$, and $\mu_{\max}=0.025$.

The general computational complexity of the different algorithms used in the AEC simulations is given in Table 1, where is evaluated for $N=500$ and APA order $Q=4$. The computational complexity of WVSS-GSVS-FDAF-PEM-AFROW is the lowest and that of DFT-PEM-AFROW is the highest. The drawback of most of the FDAF-based algorithms is their inherent delay of $N$ samples. In the AFC application, this delay, or part of it, may be "absorbed" by the forward path.

### 4.2. Performance measures

The performance measure for AEC simulations is the mean-square deviation (MSD) or "misadjustment". The MSD between the estimated echo path $\hat{\mathbf{f}}(t)$ and the true echo path $\mathbf{f}_1$ or $\mathbf{f}_2$ represents the accuracy of the estimation and is defined as,

$$\mathrm{MSD}(t)=10\log_{10}\frac{\|\hat{\mathbf{f}}(t)-\mathbf{f}_{1,2}\|_2^2}{\|\mathbf{f}_{1,2}\|_2^2}, \tag{25}$$

where $\mathbf{f}_{1,2}$ is either $\mathbf{f}_1$ or $\mathbf{f}_2$. The performance measure for AFC is the maximum stable gain (MSG). The achievable amplification before instability occurs is measured by the MSG, which is derived from the Nyquist stability criterion

Author's personal copy



**Table 1**
Complexity comparison by the number of FLOPS per recursion evaluated for $N=500$ and $Q=4$. Each FFT/IFFT is $3M\log_2 M$ FLOPS and the phase calculation (i.e., $=\arctan(\text{imag}/\text{real})$) is an $M$ complexity operation.

| Algorithm | Number of computations for $N=500$, $Q=4$ | Total |
|---|---|---|
| PC-VSS | $2NQ+7Q^2+4N+12$ | 6124 |
| PVSS | $2NQ+7Q^2+3Q+6+Q+4Q+2$ | 4152 |
| PEM-AFROW | $\left(8+\frac{4P+2n_A+1}{P}\right)N+\frac{1}{P}n_A^2+\left(4+\frac{4P+2}{P}\right)n_A+\frac{P-1}{P}+10$ | 5180 |
| DFT-PEM-AFROW | $\left(8+\frac{4P+2n_A+1}{P}\right)N+\frac{1}{P}n_A^2+\left(4+\frac{4P+2}{P}\right)n_A+\frac{P-1}{P}$ $+10+6N\log_2 N$ | 33 725 |
| WVSS-GSVS-FDAF-PEM-AFROW | $\frac{18M\log_2 M+25M+3N+n_A^2+4Mn_A+4+M}{N}$ | 859 |

[30], and is defined as

$$\text{MSG}(t) = -20\log_{10}\{\max_{\omega\in\phi_{2\pi}}|J(\omega,t)[F(\omega)-\hat{F}(\omega,t)]|\}, \quad (26)$$

where $\phi_{2\pi}$ denotes the set of frequencies at which the loop phase is a multiple of $2\pi$ [i.e., the feedback signal $x(t)$ is in phase with the near-end speech $v(t)$] and $J(\omega,t)$ denotes the forward path before the amplifier, so that $G(\omega,t)=J(\omega,t)K(t)$.

### 4.3. Simulation results for DT-robust AEC

In the first set of simulations, the noise $n(t)$ is WGN at 30-dB SNR. The PVSS and PC-VSS algorithms are first tuned, as suggested in [25] and [20], respectively, to obtain the best performance both in terms of convergence rate and final MSD. The DFT-PEM-AFROW and WVSS-GSVS-FDAF-PEM-AFROW algorithms are tuned to have an initial convergence rate similar to that of PVSS and PC-VSS. This set of parameter values remains unchanged throughout the simulations. All (sub)figures in this section consist of an upper part showing the microphone signal, consisting of the echo (dark blue) and near-end signal (light green), and the lower part showing the AEC performance in terms of MSD on the same time scale. When a change of RIR occurs, the first part of the echo signal (generated by $\mathbf{f}_1$) will be in a lighter color than the second part (generated by $\mathbf{f}_2$). Plotting the time-domain representation of these signals allows us to distinguish between the amplitude and start/end points of both the echo and near-end signals.

#### 4.3.1. Bursting DT

Figs. 5 and 6 show the AEC performance when bursting DT (from 12.5 s to 25 s) occurs at two different SERs (15 dB and 5 dB) immersed in WGN at 30-dB SNR (Fig. 5) and immersed in speech babble at 20-dB SNR (Fig. 6).

More specifically, in Fig. 5(a) and (b), a DT burst is shown at 15-dB and 5-dB SER, respectively, both immersed in WGN at 30-dB SNR. It can be seen that in WGN, PC-VSS and PVSS achieve some improvement in the final MSD compared to the other two algorithms during single talk. However, during DT, WVSS-GSVS-FDAF-PEM-AFROW outperforms the other three algorithms: in Fig. 5(b) by 10 dB in the case of PVSS and DFT-PEM-AFROW, and by around 15 dB in the PC-VSS case; in Fig. 5(b), these differences are increased since WVSS-GSVS-FDAF-PEM-AFROW outperforms DFT-PEM-AFROW by 10–15 dB, PVSS by 5–10 dB, and PC-VSS by 10–20 dB.

Fig. 6(a) and (b) shows two scenarios, where two DT bursts at 15-dB and 5-dB SER occur after 12.5 s, and the near-end noise is speech babble at 20-dB SNR in both scenarios. In these adverse scenarios, the improved performance of WVSS-GSVS-FDAF-PEM-AFROW compared to the other three algorithms is demonstrated. The convergence of PVSS is seriously degraded in this scenario. As for the PC-VSS convergence, although it is the same as with WGN during the first second, the MSD during single-talk is much higher than with WGN and it is not recovered after DT. During DT, the MSD of WVSS-GSVS-FDAF-PEM-AFROW remains at $-20$ dB, being insensitive to DT. The other three algorithms perform as in the WGN case during DT, which highlights the great difference w.r.t. the WVSS-GSVS-FDAF-PEM-AFROW performance. These differences are clearly visible in the case of DT at 5-dB SER. After DT, WVSS-GSVS-FDAF-PEM-AFROW restores the low MSD value ($-20$ dB). The PC-VSS algorithm seems to have serious problems in speech babble since its MSD is almost 10 dB higher than that for the other algorithms.

#### 4.3.2. RIR change during bursting DT

Fig. 7 shows a scenario with an abrupt change of the RIR when the near-end noise is speech babble noise at 20-dB SNR. More specifically, in Fig. 7(a) the AEC performance is shown, where the change of RIR occurs during DT at 15-dB SER and in Fig. 7(b) the SER is 5 dB. The poorest performance of PC-VSS and PVSS becomes apparent now. On the other hand, DFT-PEM-AFROW, although quite affected by DT, still show a descending trend in their MSD curve, which implies that they are converging during DT. In both Fig. 7(a) and (b), it is observed that the performance of WVSS-GSVS-FDAF-PEM-AFROW is surprisingly almost constant. This fact highlights the robustness of WVSS-GSVS-FDAF-PEM-AFROW in an adverse scenario. Indeed, its performance is barely degraded, which demonstrates the robustness of WVSS-GSVS-FDAF-PEM-AFROW as compared to the other three algorithms.

#### 4.3.3. WVSS evolution

Fig. 8 shows the WVSS evolution, i.e., $W(\omega_m,k)$ for $m=0,\ldots,499$ and $k=1,\ldots,(L-N)/N$, in two different scenarios: in Fig. 8(a), speech babble at 40-dB SNR and bursting DT at 5-dB SER, and in Fig. 8(b), WGN both for the loudspeaker signal and for the near-end noise at 20-dB SNR, with an abrupt change of the RIR ($\mathbf{f}_2=0.5\mathbf{f}_1$) at 17.5 s.

Author's personal copy



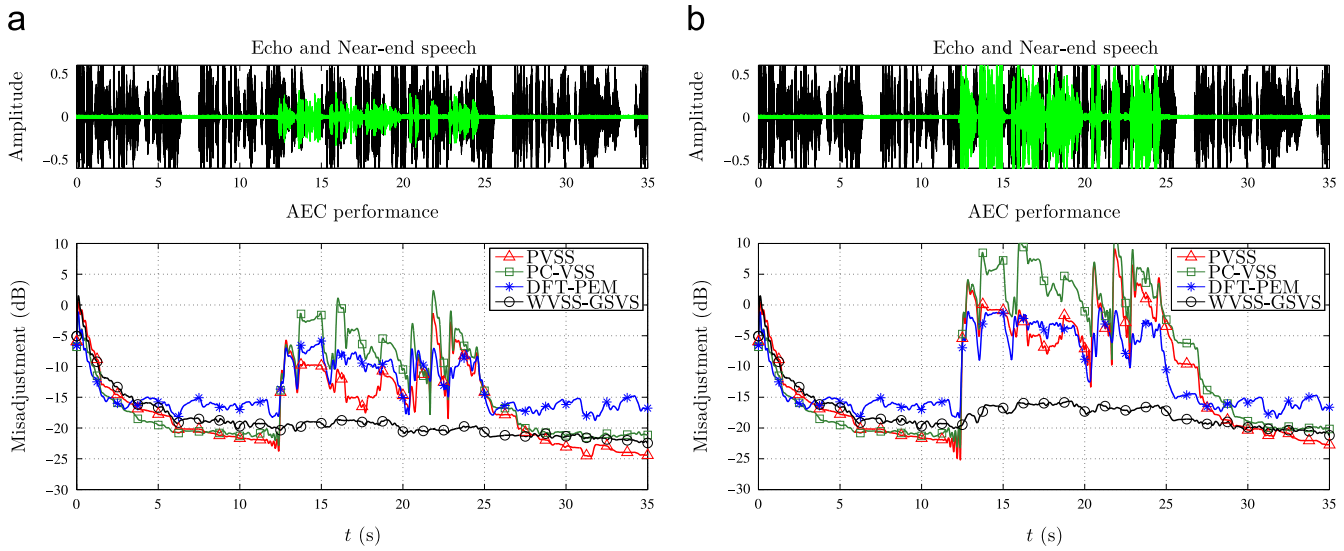



**Fig. 5.** AEC performance in WGN at 30-dB SNR. Bursting DT occurs at different SER. (a) SER = 15 dB. (b) SER = 5 dB.

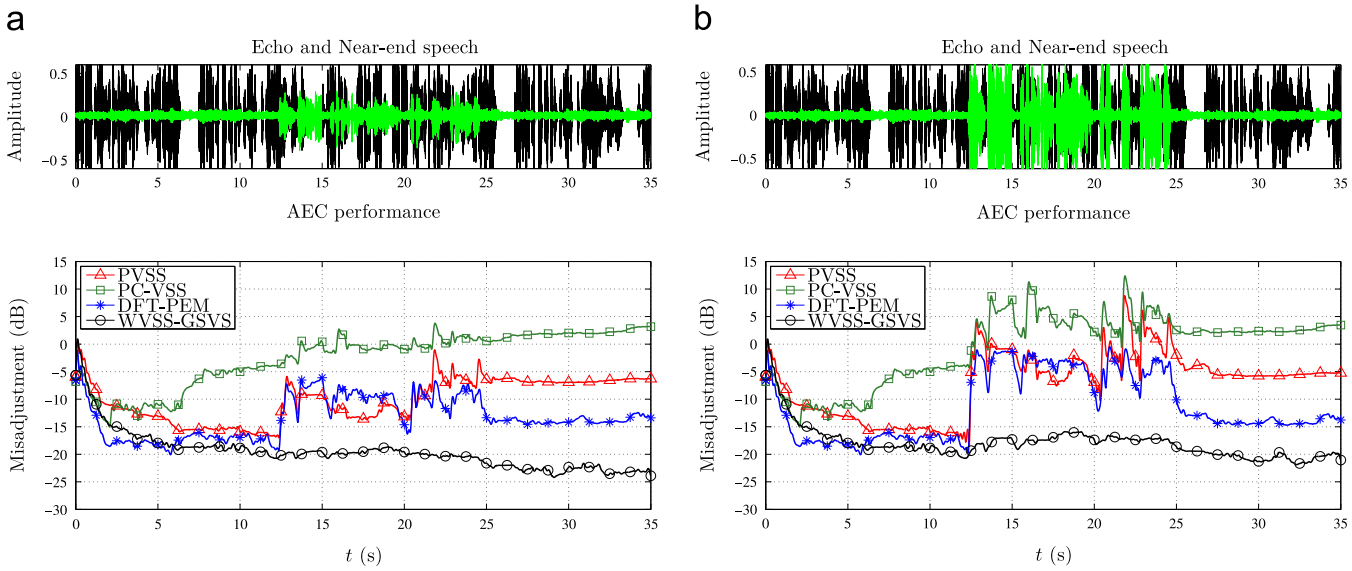



**Fig. 6.** AEC performance in speech babble at 20-dB SNR. Bursting DT occurs at different SER. (a) SER = 15 dB. (b) SER = 5 dB.

Fig. 8(a) displays the (dense) echo signal spectrogram and the near-end signal spectrogram, which clearly shows the near-end activity between 12.5 and 25 s, and at the bottom, the WVSS evolution shows a drastic step size reduction during DT. At $t = 20$ s, where the PSD of the near-end signal is low in every frequency bin, the step size is increased in those frequency bins where the SER and SNR is appropriate: step sizes in frequency bins 0–50 are not increased because the echo PSD is also low in those frequency bins and, thus, the Wiener filter gain should be low. In Fig. 8(b), it is shown how the step size follows the echo spectrum "weighted" by the near-end signal PSD.

#### 4.3.4. *WVSS-only, GSVS-only, and WVSS-GSVS with FDAF-PEM-AFROW*

We can shed some light on why WVSS-GSVS-FDAF-PEM-AFROW is so robust and yet performs so well by the comparisons in Fig. 9. A simulation of FDAF-PEM-AFROW using WVSS-only, GSVS-only, and the proposed combination WVSS-GSVS is shown in two adverse scenarios: speech babble at 10-dB and 20-dB SNR, both with DT at −5-dB SER and with an abrupt change of the RIR. In FDAF and for stochastic gradient algorithms in general, the excess MSE depends on the step size and noise variance [1,45]. The GSVS-only FDAF-PEM-AFROW, although robust and smooth, has a fixed step size, and thus, the final MSD is much higher. However, it is shown how WVSS-GSVS-FDAF-PEM-AFROW always obtains the best result compared to WVSS-only and GSVS-only. It is interesting to note how WVSS-GSVS-FDAF-PEM-AFROW transitions from GSVS-only FDAF-PEM-AFROW to WVSS-only FDAF-PEM-AFROW at around 6 s in Fig. 9(a) and around 8 s in Fig. 9(b).

### 4.4. *AFC*

Three AFC scenarios are shown in Fig. 10 to compare the performance of WVSS-GSVS-FDAF-PEM-AFROW (squares), DFT-PEM-AFROW (stars), and PEM-AFROW

Author's personal copy



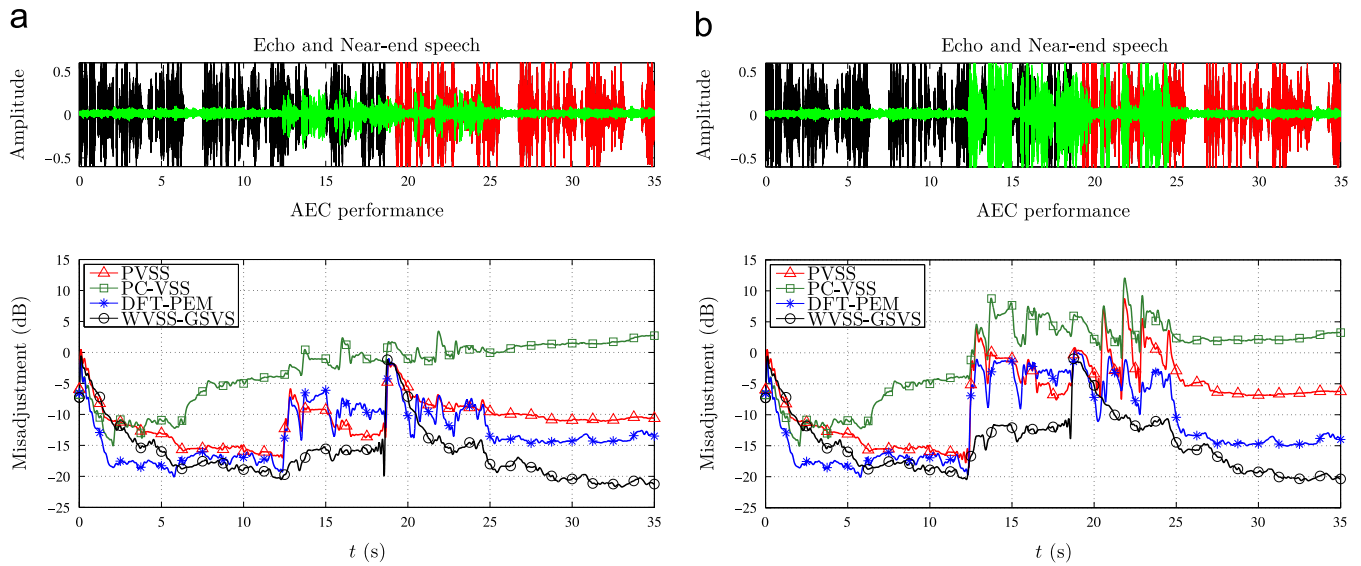



**Fig. 7.** AEC performance in speech babble at 20-dB SNR with an abrupt change of RIR after 17 s that occurs during bursting DT. (a) SER = 15 dB. (b) SER = 5 dB.

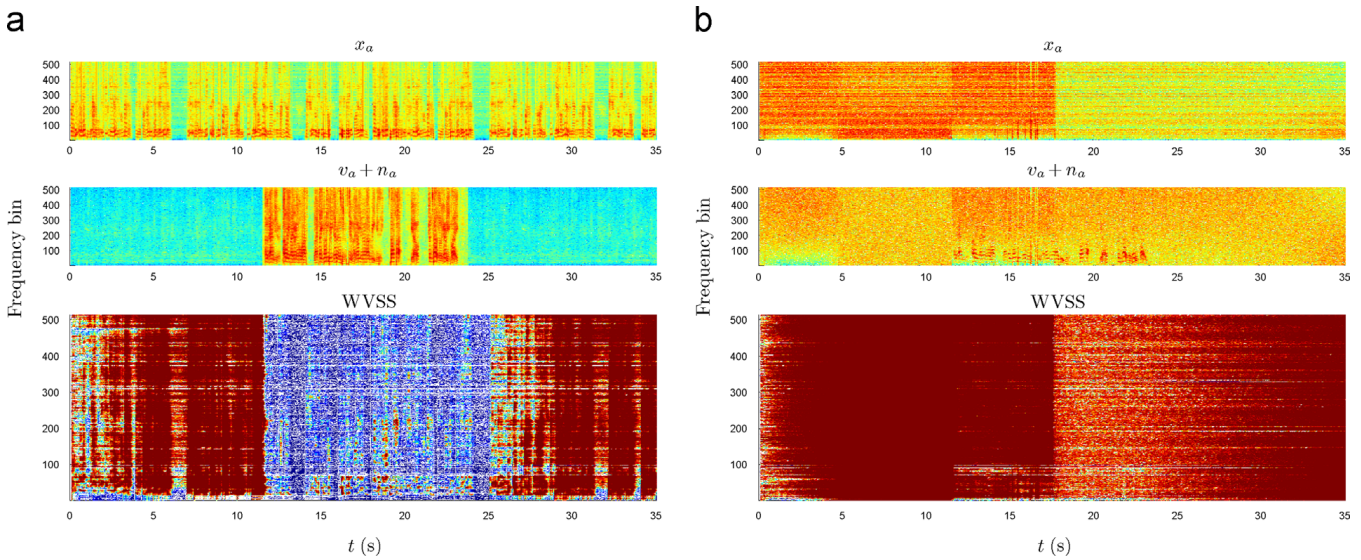



**Fig. 8.** WVSS evolution in different scenarios (bottom) together with the spectrograms of the echo signal (top) and the near-end signal (middle). (a) Speech babble at 40-dB SNR and bursting DT at 5-dB SER. (b) Both the near-end noise and the loudspeaker signal are WGN at 20-dB SNR with abrupt change of RIR at 17.5 s.

(circles). The performance is given in terms of the MSG. The value of $W(\omega_m, k) = 1$ for $k = 1, \ldots, 20$ because at start-up the estimated $\hat{P}_{X_a}(\omega_m, k)$, and therefore $W(\omega_m, k)$, are very low. The WVSS-GSVS-FDAF-PEM-AFROW algorithm needs some initial iterations before a significant feedback signal PSD can be estimated. More specifically, in Fig. 10(a), the MSG is shown when using a near-end signal model of $n_A = 55$ and WGN at 40-dB SNR. It can be seen that PEM-AFROW achieves 4–6.5 dB MSG improvement and DFT-PEM-AFROW achieves 2–4 dB of MSG improvement w.r.t. to PEM-AFROW. On the other hand, WVSS-GSVS-FDAF-PEM-AFROW outperforms the other two algorithms by 7–9 and 2–4 dB. The superior performance of WVSS-GSVS-FDAF-PEM-AFROW appears more clearly in the case of a near-end signal model of $n_A = 12$ and WGN at 40-dB SNR, as shown in Fig. 10(b). It can be seen that PEM-AFROW goes into the instability region and that DFT-PEM-AFROW outperforms PEM-AFROW by 6–7 dB when using low $n_A$ orders. WVSS-GSVS-FDAF-PEM-AFROW outperforms by far the other two algorithms, e.g., 10–15-dB improvement compared to PEM-AFROW. Moreover, it obtains almost the same performance as in the $n_A = 55$ case. In Fig. 10(c), a $n_A = 12$ AR model order is used and a speech babble at 40-dB SNR is chosen. In this case, both PEM-AFROW and DFT-PEM-AFROW remain at around 5 dB above the instability region. On the other hand, WVSS-GSVS-FDAF-PEM-AFROW maintains its superior performance of 6–9 dB w.r.t. the other two algorithms. In this last scenario, Fig. 10(d) shows the WVSS evolution in terms of its instantaneous values (in each frequency bin) as time evolves. It is clearly seen that some frequency bins usually have a smaller step size (e.g., frequency bins 5–35 and 85–100) than others (e.g., 35–60 and 70–85).

Author's personal copy



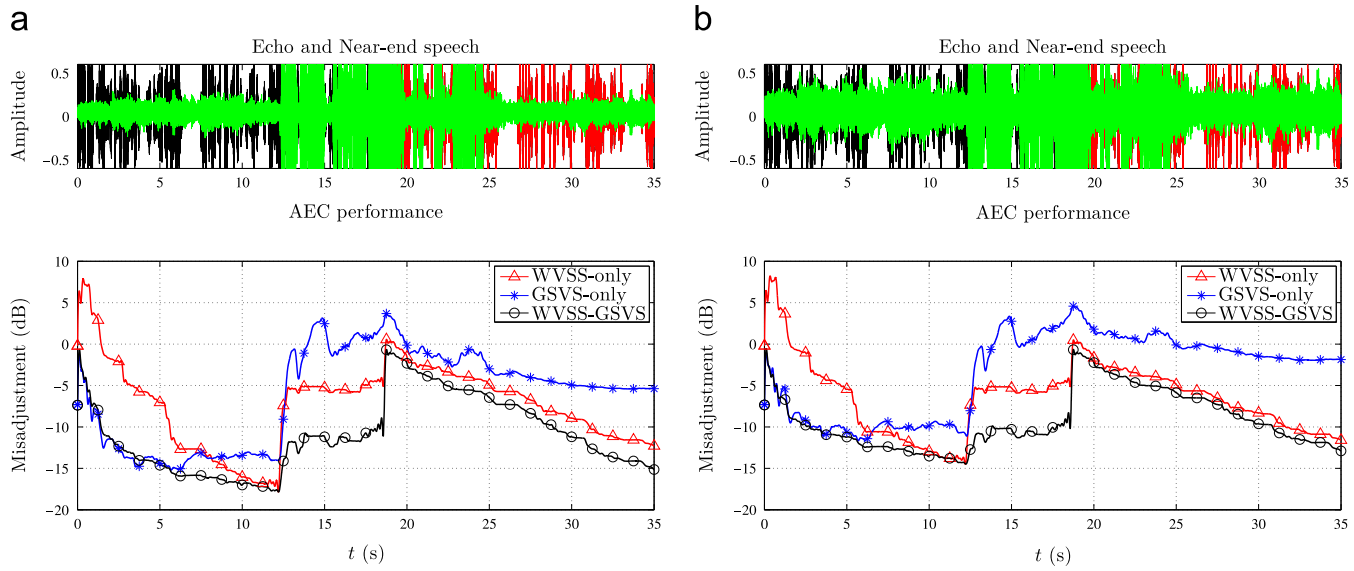



**Fig. 9.** Comparison of WVSS-only, GSVS-only, and the combination of the two WVSS-GSVS used in FDAF-PEM-AFROW for bursting DT at −5-dB SER and RIR change with continuous speech babble. (a) Speech babble at 10-dB SNR. (b) Speech babble at 20-dB SNR.

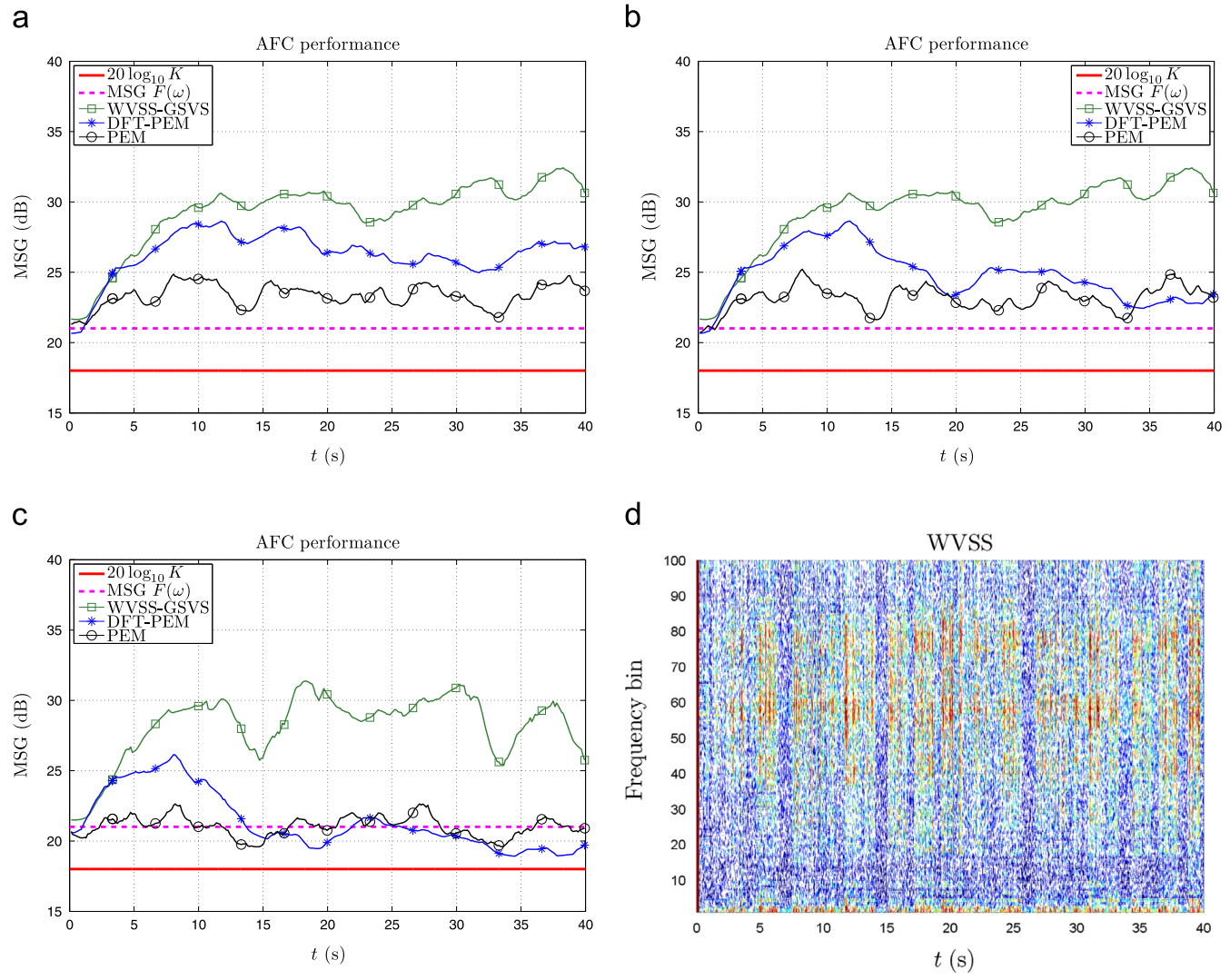



**Fig. 10.** AFC performance in terms of MSG(t) of three algorithms: PEM-AFROW, DFT-PEM-AFROW, and WVSS-GSVS-FDAF-PEM-AFROW. (a) $n_A=55$ and WGN at 40-dB SNR. (b) $n_A=12$ and WGN at 40-dB SNR. (c) $n_A=12$ and speech babble at 40-dB SNR. (d) WVSS evolution in the same scenario as (c).

Author's personal copy



## 5. Conclusion

In this paper, we have derived a practical, yet highly robust, algorithm based on the frequency-domain adaptive filter prediction error method using row operations (FDAF-PEM-AFROW) for DT-robust AEC and AFC. The proposed algorithm contains two modifications, namely, the Wiener variable step size (WVSS) and the gradient spectral variance smoothing (GSVS) to be performed in FDAF-PEM-AFROW, leading to the WVSS-GSVS-FDAF-PEM-AFROW algorithm. Simulations show that WVSS-GSVS-FDAF-PEM-AFROW outperforms other competing algorithms in adverse scenarios in both acoustic echo and feedback cancellation applications, where the near-end signals (i.e., speech and/or background noise) strongly affect the microphone signal. WVSS-GSVS-FDAF-PEM-AFROW obtains improved robustness and smooth adaptation in highly adverse scenarios, such as in bursting DT at high levels, and with a change of the acoustic path during continuous DT. The WVSS is implemented as a single-channel noise-reduction Wiener filter applied to the (prefiltered) microphone signal, where the Wiener filter gain is used as a VSS in the adaptive filter. On the other hand, the GSVS aims at reducing the variance in the noisy gradient estimates based on time-recursive averaging of gradient estimates. Combining the WVSS and the GSVS with the FDAF-PEM-AFROW algorithm consequently achieves all the characteristic we are seeking, namely, decorrelation properties (PEM, FDAF), minimum variance (GSVS, FDAF, PEM), variable step size (WVSS), and computational efficiency (FDAF).

## References


[1] S. Haykin, Adaptive Filter Theory, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[2] A. Spriet, I. Proudler, M. Moonen, J. Wouters, Adaptive feedback cancellation in hearing aids with linear prediction of the desired signal, IEEE Trans. Signal Process. 53 (10) (2005) 3749–3763.

[3] K. Ngo, T. van Waterschoot, M.G. Christensen, S.H.J.M. Moonen, J. Wouters, Prediction-error-method-based adaptive feedback cancellation in hearing aids using pitch estimation, in: Proceedings of the 18th European Signal Processing Conference (EUSIPCO'10), Aalborg, Denmark, August 2010, pp. 2422–2426.

[4] T. van Waterschoot, G. Rombouts, P. Verhoeve, M. Moonen, Double-talk-robust prediction error identification algorithms for acoustic echo cancellation, IEEE Trans. Signal Process. 55 (3) (2007) 846–858.

[5] D.L. Duttweiler, A twelve-channel digital echo canceler, IEEE Trans. Commun. 26 (5) (1978) 647–653.

[6] K. Ghose, V.U. Reddy, A double-talk detector for acoustic echo cancellation applications, Signal Process. 80 (8) (2000) 1459–1467.

[7] J. Benesty, D.R. Morgan, J.H. Cho, A new class of double-talk detectors based on cross-correlation, IEEE Trans. Speech Audio Process. 8 (2) (2000) 168–172.

[8] M. Kallinger, A. Mertins, K.-D. Kammeyer, Enhanced double-talk detection based on pseudo-coherence in stereo, in: Proceedings of the 2005 of the International Workshop on Acoustic Echo and Noise Control (IWAENC'05), Eindhoven, The Netherlands, September 2005, pp. 177–180.

[9] H.K. Jung, N.S. Kim, T. Kim, A new double-talk detector using echo path estimation, Speech Commun. 45 (1) (2005) 41–48.

[10] S. Gustafsson, F. Schwarz, A postfilter for improved stereo acoustic echo cancellation, in: 1999 International Workshop on Acoustic Echo Noise Control (IWAENC'99), Pocono Manor, Pennsylvania, September 1999, pp. 32–35.

[11] G. Enzner, P. Vary, Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones, Signal Process. 86 (6) (2006) 1140–1156.

[12] P. Loizou, Speech Enhancement: Theory and Practice, Taylor and Francis, Boca Raton, Florida, 2007.

[13] T.S. Wada, B.-H. Juang, Enhancement of residual echo cancellation for improved acoustic echo cancellation, in: Proceedings of the 15th European Signal Processing Conference (EUSIPCO'07), Poznan, Poland, September 2007, pp. 1620–1624.

[14] T.S. Wada, B.-H. Juang, Towards robust acoustic echo cancellation during double-talk and near-end background noise via enhancement of residual echo, in: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08), Las Vegas, USA, March 2008, pp. 253–256.

[15] V.J. Mathews, Z. Xie, A stochastic gradient adaptive filter with gradient adaptive step size, IEEE Trans. Signal Process. 41 (6) (1993) 2075–2087.

[16] W.-P. Ang, B. Farhang-Boroujeny, A new class of gradient adaptive step-size LMS algorithms, IEEE Trans. Signal Process. 49 (4) (2001) 805–810.

[17] H.-C. Shin, A.H. Sayed, W.-J. Song, Variable step-size NLMS and affine projection algorithms, IEEE Signal Process. Lett. 11 (2) (2004) 132–135.

[18] Y. Zhang, J.A. Chambers, W. Wang, P. Kendrick, T.J. Cox, A new variable step-size LMS algorithm with robustness to nonstationary noise, in: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07), Honolulu, Hawaii, USA, April 2007, pp. 1349–1352.

[19] Y. Zhang, N. Li, J.A. Chambers, Y. Hao, New gradient-based variable step size LMS algorithms, EURASIP J. Adv. Signal Process. 2008 (105) (2008) 1–9.

[20] T. Creasy, T. Aboulnasr, A projection-correlation algorithm for acoustic echo cancellation in the presence of double talk, in: Proceedings of 2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00), Istanbul, Turkey, June 2008, pp. 436–439.

[21] K. Ozeki, T. Umeda, An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, Electron. Commun. Jpn. 67 (5) (1984) 19–27.

[22] J. Benesty, H. Rey, L.R. Vega, S. Tressens, A nonparametric VSS NLMS algorithm, IEEE Signal Process. Lett. 13 (10) (2006) 581–584.

[23] C. Paleologu, S. Ciochină, J. Benesty, Double-talk robust VSS-NLMS algorithm for under-modeling acoustic echo cancellation, in: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08), Las Vegas, USA, March 2008, pp. 1141–1144.

[24] M.A. Iqbal, S.L. Grant, Novel variable step size NLMS algorithms for echo cancellation, in: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'08), Las Vegas, USA, March 2008, pp. 241–244.

[25] C. Paleologu, J. Benesty, S. Ciochină, A variable step-size affine projection algorithm designed for acoustic echo cancellation, IEEE Trans. Audio Speech Lang. Process. 16 (8) (2008) 1466–1478.

[26] S.M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, Upper Saddle River, New Jersey, 1993.

[27] T. van Waterschoot, M. Moonen, Double-talk robust acoustic echo cancellation with continuous near-end activity, in: Proceedings of the 13th European Signal Processing Conference (EUSIPCO'05), Antalya, Turkey, 2005, pp. 2517–2535.

[28] L. Ljung, System Identification: Theory for the user, Prentice Hall, Englewood Cliffs, New Jersey, 1987.

[29] G. Rombouts, T. van Waterschoot, K. Struyve, M. Moonen, Acoustic feedback cancellation for long acoustic paths using a nonstationary source model, IEEE Trans. Signal Process. 54 (9) (2006) 3426–3434.

[30] T. van Waterschoot, M. Moonen, Fifty years of acoustic feedback control: state of the art and future challenges, Proc. IEEE 99 (2) (2011) 288–327.

[31] T. van Waterschoot, M. Moonen, Adaptive feedback cancellation for audio applications, Signal Process. 89 (11) (2009) 2185–2201.

[32] K. Ngo, T. van Waterschoot, M.G. Christensen, M. Moonen, S.H. Jensen, J. Wouters, Adaptive feedback cancellation in hearing aids using a sinusoidal near-end signal model, in: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10), Dallas, USA, 2010, pp. 1878–1893.

[33] K. Ngo, T. van Waterschoot, M.G. Christensen, M. Moonen, S.H. Jensen, Improved prediction error filters for adaptive feedback cancellation in hearing aids, Signal Process. 93 (11) (2013) 3062–3075.

[34] J.M. Gil-Cacho, T. van Waterschoot, M. Moonen, S.H. Jensen, Transform domain prediction error method for improved acoustic echo and feedback cancellation, in: Proceedings of the 20th European Signal Processing International Conference (EUSIPCO'12), Bucharest, Rumania, August 2012, pp. 2422–2426.

[35] J.M. Gil-Cacho, T. van Waterschoot, M. Moonen, S.H. Jensen, A frequency-domain adaptive filtering (FDAF) prediction error

Author's personal copy



method (PEM) for double-talk-robust acoustic echo cancellation, IEEE Trans. Audio Speech Lang. Process, ESAT-STADIUS, November 2013, pp. 1–27 (online ⟨ftp://ftp.esat.kuleuven.be/pub/SISTA/pepegilcacho/reports/gilcacho13_13.pdf⟩).

[36] T. Trump, A frequency domain adaptive algorithm for colored measurement noise environment, in: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98), Seattle, USA, March 1998, pp. 1705–1708.

[37] J.M. Gil-Cacho, T. van Waterschoot, M. Moonen, S.H. Jensen, Matlab scripts for: Wiener variable step size and gradient spectral variance smoothing for double-talk-robust acoustic echo cancellation and acoustic feedback cancellation, in: Technical Report KULuven, ESAT-STADIUS, June 2013 (online ⟨http://homes.esat.kuleuven.be/pepe/abstract12-214_2.html⟩).

[38] J.J. Shynk, Frequency-domain and multirate adaptive filtering, IEEE Signal Process. Mag. 9 (1) (1992) 14–37.

[39] N.J. Bershad, P.L. Feintuch, Analysis of the frequency domain adaptive filter, Proc. IEEE 67 (12) (1979) 1658–1659.

[40] M.H. Hayes, Statistical Digital Signal Processing and Modeling, John Wiley & Sons, Inc., New York, NY, 1996.

[41] N. Krishnamurthy, J.H.L. Hansen, Babble noise: modeling, analysis, and applications, IEEE Trans. Audio Speech Lang. Process. 17 (7) (2009) 1394–1407.

[42] C. Rohrs, R. Younce, Double Talk Detector for Echo Canceler and Method, April 17, 1990. URL ⟨http://www.patentlens.net/patentlens/patent/US4918727/⟩.

[43] T. Gänsler, S.L. Gay, M.M. Sondhi, J. Benesty, Double-talk robust fast converging algorithms for network echo cancellation, IEEE Trans. Speech Audio Process. 8 (6) (2000) 656–663.

[44] H. Rey, L.R. Vega, S. Tressens, J. Benesty, Variable explicit regularization in affine projection algorithm: robustness issues and optimal choice, IEEE Trans. Signal Process. 55 (5) (2007) 2096–2108.

[45] B. Farhang-Boroujeny, Adaptive Filters: Theory and Applications, Wiley, Chichester, UK, 1998.