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 Waterschoot[a], Marc Moonen[a], Søren Holdt Jensen[b]

[a] *KU Leuven, Department of Electrical Engineering-ESAT, SCD-SISTA/iMinds Future Health Department, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium.*
[b] *Department of Electronic Systems Aalborg University, Fredrik Bajers Vej 7, DK-9220 Aalborg, Denmark.*

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 for AFC. The proposed algorithm features two main modifications: (a) the WIener variable Step sizE (WISE), and (b) the GRAdient Spectral variance Smoothing (GRASS). In AEC simulations, WISE-GRASS-FDAF-PEM-AFROW algorithm obtains improved 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.

Keywords: Acoustic echo cancellation, acoustic feedback cancellation, adaptive filtering, Wiener

☆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. Email: pepe.gilcacho@esat.kuleuven.be

variable step size, prediction error method, gradient smoothing, double-talk.

1. Introduction

Acoustic echo and acoustic feedback are two well-known problems appearing in speech communication systems, which are caused by the acoustic coupling between a loudspeaker and a microphone. On the one hand, acoustic echo cancellation (AEC) is widely used in mobile and hands-free telephony [1] where the existence of echoes degrades 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]. This maximum amplification may be too small to compensate for the hearing loss, which makes acoustic feedback cancellation (AFC) an important component in hearing aids. Figure 1 shows the typical set-up for AEC and AFC. The goal of AEC and AFC is essentially to identify a model for the echo or feedback path, i.e., the room impulse response (RIR), and to estimate the echo or feedback signal which is then subtracted from the microphone signal. 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. $F(q,t) = f_0(t) + f_1(t)q^{-1} + ... + f_{n_F}(t)q^{-n_F}$ represents a linear time-varying model of the RIR between the loudspeaker and the microphone, where n_F is the RIR model order. Therefore, 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 $e(t) = \left[F(q,t) - \hat{F}(q,t)\right] u(t) + v(t) + n(t)$.

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. In AFC applications, on the other hand, 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:

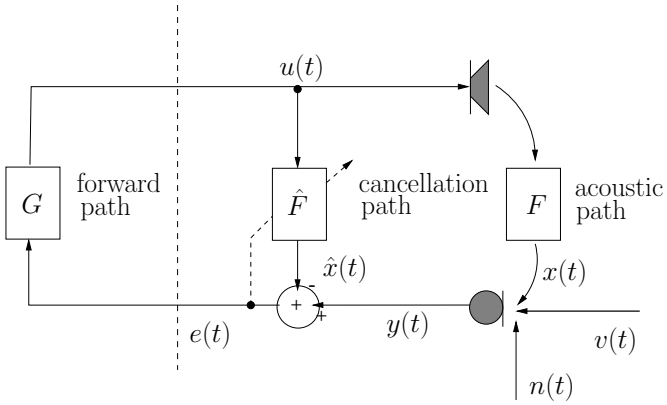


Figure 1: Typical set-ups for AEC/AFC. The left part (forward path) only relates to AFC. The right part relates to both AEC and AFC.

1.1. The problem of 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 the 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, e.g., [6]. However, in general, DTDs take some time until an onset of a DT period is detected. Moreover, in some AEC scenarios the near-end signal will be continuously present and then the use of a DTD becomes futile. This may be the case for example in a noisy teleconferencing or AFC application. 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* (VSS) to slow down adaption during DT, (3) by prefiltering the loudspeaker and microphone signal with a decorrelation filter to achieve a *minimum-variance* RIR estimate.

The first approach consists in the use of a postfilter which interplays with the AEC to suppress residual echo (and also to reduce near-end signals) based on signal enhancement techniques [7]-[9]. On the other hand, in [10], [11], the idea behind the postfilter design is the opposite, i.e.,

to the 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 VSS which may be derived using information about the gradient vector or information about the near-end signal power. The first type of VSS algorithms rely on two properties of the gradient vector to control the step size [12]-[16]: (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, (2) the property that the gradient vector direction will generally show a coherent 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 [17]. PC-VSS is a VSS algorithm based on the affine projection algorithm (AP) [18] 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 to achieve robustness and to distinguish between an RIR 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 4.1.

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 [19]. 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 with 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 successfully developed for DT-robust AEC, e.g., [20], [21]. Their convergence, however, is slow in practice and hence, an APA version of the VSS-NLMS algorithm in [21] has been proposed in [22] to increase the convergence speed. The resulting *practical* VSS affine projection algorithm (PVSS) [22] is chosen as one of the competing algorithms in this paper and will be further explained also in Section 4.1.

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 RIR estimate, which is also known as the *best linear unbiased estimate* (BLUE) [23], depends on the near-end signal characteristics, which are in practice unknown and time-varying [4], [24]. The algorithms in [4], [24] aim to whiten the near-end speech component in the microphone signal by using adaptive decorrelation filters that are estimated concurrently with the RIR. 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 [25] 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], [24], the PEM-based adaptive filtering using row operations (PEM-AFROW) [26] 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], [24] 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. The problem of 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 to converge to a *biased* solution [2], [27]. This means that the adaptive filter does not only predict and cancel the feedback component in the microphone signal, but also 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], [27] using the PEM [25]. 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 still to 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 [28], the pitch prediction model [3], and the sinusoidal model [29] 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 [30]. In [31], a Transform-Domain PEM-AFROW (TD-PEM-AFROW) algorithm 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 TD-PEM-AFROW are chosen as competing algorithms for AFC. TD-PEM-AFROW was also successfully applied to DT-robust AEC in [31] and it is, therefore, also chosen as a competing algorithm for AEC.

1.3. Contributions and outline

In [32], 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 short-term estimated correlation (STEC) between the near-end signal and the far-end signal is heavily reduced when using FDAF-PEM-AFROW compared to stochastic gradient and (time-domain) PEM-AFROW algorithms [32] and (2) FDAF itself may be seen to minimize a BLUE criterion if a proper normalization factor is used during adaptation [33]. In this paper we propose two modifications of the FDAF-PEM-AFROW algorithm for robust and smooth adaptation both in AFC and in AEC in continuous DT and bursting DT without the need of a DTD. In particular, we propose the WIener variable Step sizE (WISE) and the GRAdient Spectral variance Smoothing (GRASS) to be performed in FDAF-PEM-AFROW leading to the WISE-GRASS-FDAF-PEM-AFROW algorithm. The WISE modification is implemented as a single-channel noise reduction Wiener filter applied to the (prefiltered) microphone signal. The *Wiener filter gain* [9] is used as a VSS in the adaptive filter, rather than as a signal enhancement parameter. On the other hand, the GRASS modification aims at reducing the variance in the noisy gradient estimates based on time-recursive averaging of instantaneous gradients. The combination of the WISE and the GRASS into the FDAF-PEM-AFROW algorithm consequently gathers the best characteristics we are seeking for in an algorithm both for AEC and AFC, namely, decorrelation properties (PEM, FDAF), minimum variance (GRASS, FDAF, PEM), variable step size (WISE), 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 [17], the PVSS [22] and the TD-PEM-AFROW [31] algorithm. A description of the competing algorithms is provided together with a computational complexity analysis. The Matlab files implementing all the algorithms and those generating the figures in Section 3-4 can be found online [1]. Finally, Section 5 concludes the paper.

2. Prediction error method

The PEM-based AEC/AFC is shown in Figure 2.

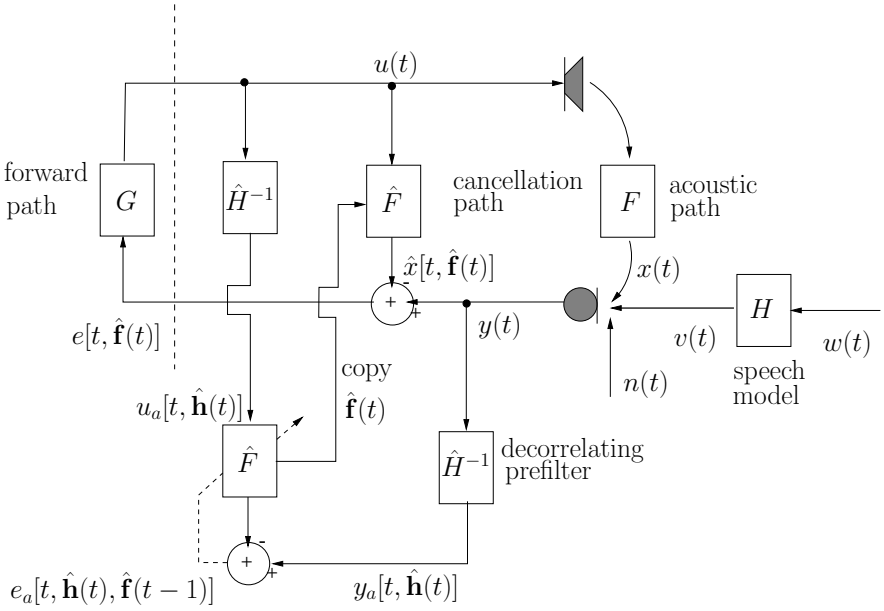


Figure 2: AEC/AFC with prefiltering of the loudspeaker and microphone signal using the inverse $[H^{-1}(q,t)]$ of the near-end speech model $[H(q,t)]$.

[1] *http : //homes.esat.kuleuven.be/ ∼ pepe*

It relies on a linear model for the near-end speech $v(t)$, which in Figure 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 a white noise signal 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 RIR 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 [25]. A common approach in PEM-based AEC/AFC is to model the near-end speech with an AR model i.e.,

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

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

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

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

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

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

where the prediction error is defined as

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

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

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

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

Algorithm 1.1. First part of the WISE-GRASS-FDAF-PEM-AFROW algorithm for AEC/AFC.

The first part shows the FDAF-PEM-AFROW, whereas the second part shows the WISE-GRASS.

Initialize: K, $k = 0$ and $\hat{\mathbf{P}}_{U_a} = \hat{\mathbf{P}}_{X_a} = \hat{\mathbf{P}}_{D_a} = \hat{\mathbf{F}} = \boldsymbol{\nabla} = \mathbf{0}_{M\times 1}$
$[\mathbf{U}(k) = \mathcal{F}\{\mathbf{u}(k)\}]$
$\hat{\mathbf{x}}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k)\cdot\hat{\mathbf{F}}(k-1)\}$ (Echo estimation)
$\left[\mathbf{x}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k)\cdot\mathbf{F}\}\right.$ (True echo signal simulation)$\left.\right]$
$[\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k) + \mathbf{n}(k)$ (Microphone signal simulation)$]$
$\mathbf{e}(k) = [\mathbf{y}(k) - \hat{\mathbf{x}}(k)]_{N+1:M}$ (Error signal)
$[\mathbf{u}(k+1) = K\cdot\mathbf{e}(k)$ (Loudspeaker signal simulation. Only for AFC)$]$
for $k = 1, 2, ...$ **do**
    $[\mathbf{U}(k) = \mathcal{F}\{\mathbf{u}(k)\}]$
    $\hat{\mathbf{x}}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k)\cdot\hat{\mathbf{F}}(k-1)\}$ (Echo estimation)
    $\left[\mathbf{x}(k) = \mathcal{F}^{-1}\{\mathbf{U}(k)\cdot\mathbf{F}\}\right.$ (True echo signal simulation)$\left.\right]$
    $[\mathbf{y}(k) = \mathbf{x}(k) + \mathbf{v}(k) + \mathbf{n}(k)$ (Microphone signal simulation)$]$
    $\mathbf{e}(k) = [\mathbf{y}(k) - \hat{\mathbf{x}}(k)]_{N+1:M}$
    $[\mathbf{u}(k+1) = K\cdot\mathbf{e}(k)$ (Loudspeaker signal simulation. Only for AFC)$]$
    $\hat{\mathbf{a}}(k) = \mathrm{AR}\{\left[\mathbf{e}^T(k) \quad \mathbf{e}^T(k-1)\right]_{1:P}^T, n_A\}$ (Order n_A AR coefficients estimation)
    for $m = 0, ..., M-1$ **do** (Decorrelation prefilter)
        $u_a(m,k) = [u(kN+1+m), ..., u(kN+1+m-n_A)]\mathbf{a}(k)$
        $y_a(m,k) = [y(kN+1+m), ..., y(kN+1+m-n_A)]\mathbf{a}(k)$
    end for
    $\mathbf{U}_a(k) = \mathcal{F}\{\mathbf{u}_a(k)\}$
    $\hat{\mathbf{x}}_a(k) = \mathcal{F}^{-1}\{\mathbf{U}_a(k)\cdot\hat{\mathbf{F}}(k-1)\}$
    $\mathbf{e}_a(k) = [\mathbf{y}_a(k) - \hat{\mathbf{x}}_a(k)]_{N+1:M}$ [(Prediction) Error signal (7)]
    $\mathbf{E}_a(k) = \mathcal{F}\{[\mathbf{0}_N \quad \mathbf{e}_a^T(k)]^T\}$

(... Continue in Algorithm 1.2 ...)

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 blocks that approximates the stationary interval of speech. The PEM-AFROW algorithm was originally developed in an AFC framework [26] and applied to a continuous-DT AEC scenario in [24]. 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 [26].

3. Proposed WISE-GRASS Modifications in FDAF-PEM-AFROW for AEC/AFC

The WISE-GRASS-FDAF-PEM-AFROW for AEC/AFC is given in Algorithm 1 (parts 1.1 and 1.2). The FDAF implementation corresponds to the overlap-save FDAF with gradient constraint and power normalization [34], 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 in FDAF, e.g., $\mathbf{u}(k) = [u(kN - N + 1), ...u(kN + N)]^T$, where $k = 0, 1, ...\frac{L-N}{N}$ is the block-time index, L is the total length of the signals, and finally, P is the block length used to estimate $A(q, t)$ with $N \leq P \leq 2N$. The upper asterisk $(\cdot)^*$ denotes complex conjugation and $[\cdot]_{a:b}$ represents a range of samples within a vector. The subscript m in ω_m, e.g., $E_a(\omega_m, k)$, refers to a signal in the m-th frequency bin of a block, $m = 0, ..., M-1$. On the other hand, 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 Figure 2. Lines $2, 4, 5, 7$ and $9, 11, 12, 14$ correspond to the data generation and are bracketed to enphasyzed that these are not truly part of the algorithm. The specific WISE-GRASS modifications within the FDAF-PEM-AFROW algorithm are shown in Algorithm 1.2. The proposed WISE-GRASS-FDAF-PEM-AFROW weight update equation is given by

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

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

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, μ_{max} sets the maximum allowed value of the step size, and $\mathbf{W}(k)$ together with $\boldsymbol{\Theta}(k)$ form the WISE-GRASS modifications which will be explained in detail in the following sections.

3.1. WIener variable Step sizE

The *WIener variable Step sizE* (WISE) modification into 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 $ENR(t) = \dfrac{\sigma_x^2(t)}{\sigma_v^2(t) + \sigma_n^2(t)}$, where $\sigma_x^2(t)$, $\sigma_v^2(t)$ and

Algorithm 1.2. Second part of the WISE-GRASS-FDAF-PEM-AFROW algorithm for AEC/AFC.

The second part explains the WISE-GRASS.

for $m = 0, ..., M-1$ **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)

$\mu(\omega_m, k) = \left[\hat{P}_{U_a}(\omega_m, k) + \delta\right]^{-1}$ (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)

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

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

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

$$W(\omega_m, k) = \frac{\sqrt{\hat{P}_{X_a}(\omega_m, k)}}{\sqrt{\hat{P}_{X_a}(\omega_m, k) + \hat{P}_{D_a}(\omega_m, k)}} \text{ (WISE)}$$

end for

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

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

end for

$\sigma_n^2(t)$ 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 WISE is that of applying 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 we use the Wiener filter gain as a variable step size in the adaptive filer. The step size in each frequency bin is then varied by the gain of the Wiener filter at the frequency bin.

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 [35]. 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. A common assumption in single-channel noise reduction 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 neglected the terms corresponding to the expected value of the cross-product of $X(\omega_m, k)$ and $D(\omega_m, k)$ assuming they are uncorrelated. Usually, the near-end speech and the far-end speech are indeed *statistically* uncorrelated, which however does not imply that the STEC between these two signals is zero. The (rather strong) assumption, of the near-end speech being uncorrelated with the loudspeaker signal, is also adopted in most AEC applications. In AFC applications, on the other hand, this assumption is clearly violated as explained in Section 1.2. In [32], we have shown that the STEC between the near-end speech and the loudspeaker signal may be very large. Besides, the practical computation of (14) is generally based on time-recursive estimates of the PSD using short-time Fourier transform (STFT) of $X(\omega_m, k)$ and $D(\omega_m, k)$. All this means that, in real AEC and AFC applications, the practical computation of (14) would clearly lead to misleading gain values.

In [32], however, we have shown that the STEC between the far-end and near-end speech can be significantly reduced by using the FDAF-PEM-AFROW framework. Consequently, we will consider deriving the Wiener filter gains (14) using the filtered signals (see Figure 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 (theoretical) frequency-domain Wiener filter

using whitened signals will moreover result in a similar filter as the theoretical one in (14), since $P_{X_aY_a}(\omega_m,k) = \mathcal{E}\{A(\omega_m,k)X(\omega_m,k)Y^*(\omega_m,k)A^*(\omega_m,k)\} = |A(\omega_m,k)|^2 P_{XY}$, and $P_{Y_a}(\omega_m,k) = \mathcal{E}\{A(\omega_m,k)Y(\omega_m,k)Y^*(\omega_m,k)A^*(\omega_m,k)\} = |A(\omega_m,k)|^2 P_Y$, 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 compute time-recursive estimates of the PSDs instead and use available estimates of the desired signal, i.e., $\hat{X}_a(\omega_m,k)$, and of the noise component in the microphone signal, i.e., $E_a(\omega_m,k)$. The Wiener filter gains are thus efficiently estimated as follows

$$m = 0, ..., M-1$$

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

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

Finally, using (16)-(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 in (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 in (17) using the prediction-error signal $E_a(\omega_m,k)$.

An interpretation of the Wiener filter gain $W(\omega_m,k)$ in terms of the $E\hat{N}R(\omega_m,k)$ can be given by writing (18) as

$$W(\omega_m,k) = \frac{E\hat{N}R(\omega_m,k)}{E\hat{N}R(\omega_m,k)+1} \tag{19}$$

where $E\hat{N}R(\omega_m,k) = \hat{P}_{X_a}(\omega_m,k)/\hat{P}_{D_a}(\omega_m,k) = \hat{P}_{X_a}(\omega_m,k)/\left[\hat{P}_{V_a}(\omega_m,k)+\hat{P}_{N_a}(\omega_m,k)\right]$. The Wiener filter gain is a real positive number in the range $0 \leq W(\omega_m,k) \leq 1$ which 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 of (1) an 'echo-only' microphone signal, i.e., $E\hat{N}R(\omega_m,k) = \infty$ and (2) a 'near-end-only' microphone signal, i.e., $E\hat{N}R(\omega_m,k) = 0$. 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. In 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* (GRASS) is included to reduce the variance in the gradient estimation, in particular, the effect on the gradient estimation of sudden high-amplitude near-end signal samples. Although the step size control has been considered in Section 3.1, it may not be sufficiently fast to react in 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 may be separated into two components as

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

where $\theta_0(\omega_m, k) = \left[X(\omega_m, k) - \hat{X}(\omega_m, k)\right] {U_a}^*(\omega_m, k)$ 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 the GRASS is that of applying averaged periodograms, which are typically used for estimating the PSD of a signal [36]. The periodogram is defined here as the squared magnitude of the gradient. An estimate of the true gradient PSD, at a given frequency ω_m, is obtained as

$$\hat{P}_{\hat{\theta}_0}(\omega_m, k) = \frac{1}{K} \sum_{j=0}^{K-1} |\theta(\omega_m, k-j)|^2. \tag{21}$$

It is shown in [36] that the variance of (21) is

$$\text{var}\left\{\hat{P}_{\hat{\theta}_0}(\omega_m, k)\right\} \approx \frac{1}{K} |\theta_0(\omega_m, k)|^2 \tag{22}$$

so that it is proportional to the square of the true gradient's PSD divided by K and hence, as $K \to \infty$, (22) approaches zero. 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. Therefore the concept of averaging starts to lose its sense. However we will assume that the true gradient varies slowly, therefore, a low-pass filter (LPF) with a low cut-off frequency would be more beneficial than (21).

The realization of the GRASS is based on a time-recursive averaging of the gradient estimate (23)-(24). This is another way to obtain a first order infinite impulse response (IIR) filtering where λ_3 is the pole of the low-pass IIR filter, and it is given as

$$\theta(\omega_m, k) = E_a(\omega_m, k){U_a}^*(\omega_m, k) \tag{23}$$

$$\nabla(\omega_m, k) = \lambda_3 \nabla(\omega_m, k-1) + (1-\lambda_3)|\theta(\omega_m, k)|^2 \tag{24}$$

In fact, a LPF with a low cut-off frequency, will effectively filter out the gradient noise and reduce the variance in the gradient estimation. Note that (21), is a special case of a low-pass FIR filter, i.e., $\hat{P}_{\hat{\theta}_0}(\omega_m, k) = \sum_{j=0}^{K-1} b_k |\theta(\omega_m, k-j)|^2$ with weights $b_j = 1/K, \forall j$.

Finally, the phase $\alpha(\omega_m, k)$, that is obtained from the gradient estimate in (25), is applied in (26) to form the GRASS, i.e., $\Theta(\omega_m, k)$, as

$$\alpha(\omega_m, k) = \angle\theta(\omega_m, k) \tag{25}$$

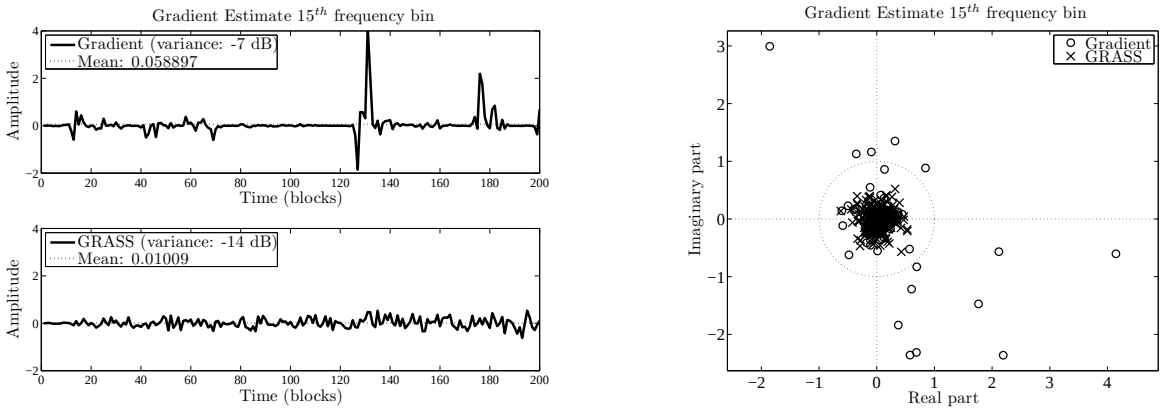
$$\Theta(\omega_m, k) = \sqrt{\nabla(\omega_m, k)} e^{[j\alpha(\omega_m, k)]} \tag{26}$$

A practical simulation of the GRASS for $m = 15$, i.e., the 15-th frequency bin, with $k = 1, ..., 200$ and $\lambda_3 = 0.95$ is shown in Figure 3. For this simulation we have used two speech signals, $u(t)$ and $v(t)$, and a noise signal $n(t)$ 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 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 GRASS gradient. We have assumed that the true gradient has converged to an optimum, and thus it is, by definition, equal to zero. In Figure 3a 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. Besides, in Figure 3b a complex representation is shown where the circles ($\circ$) represent the complex value of $\theta(\omega_m, k)$ and the crosses ($\times$) represent the complex value of $\Theta(\omega_m, k)$. In both representations, the variance in the estimation is shown to be reduced by 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, a white Gaussian noise (WGN) and a *speech babble* noise which is one of the most challenging noise types [37] in signal enhancement applications. We define two measures which determine the relative signal levels in the simulations: the *Signal-to-Echo Ratio* (SER)$= 10\log_{10}\frac{\sigma_x^2}{\sigma_v^2}$ and the *Signal-to-Noise Ratio* (SNR)$= 10\log_{10}\frac{\sigma_x^2}{\sigma_n^2}$, where σ_x^2, σ_v^2 and σ_n^2 are the variances of the echo, the near-end speech and near-end noise.

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



(a) (b)

Figure 3: Instantaneous gradient and GRASS estimates with $m = 15$ and $k = 1, ..., 200$. (a) Time representation: Upper and lower figures correspond to the real part of $\theta(\omega_m, k)$ and $\Theta(\omega_m, k)$, respectively. The solid line $(-)$ represents the gradient estimate and the dotted line $(\cdots)$ represents its mean value. (b) Complex representation: the circles $(\circ)$ represent the complex value of $\theta(\omega_m, k)$ and the crosses $(\times)$ represent the complex value of $\Theta(\omega_m, k)$.

segments of speech: the first and third 12 s segment 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 $n_A = 12$, following the indications given in [32], and the APA order is $Q = 4$. There are two RIRs used in the AEC simulations: an 80-taps impulse response $\mathbf{f}_1$ measured from a real mobile phone, and an artificial RIR $\mathbf{f}_2 = \mathbf{f}_1 + \bar{\mathbf{n}} \cdot \mathbf{f}_1$ simulating an abrupt RIR change where $\bar{\mathbf{n}}$ is one realization of a Gaussian random process. 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], [31]: $n_A = 12$ 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 *maxinum stable gain* (MSG) without feedback cancellation i.e., see (28) 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.

Algorithm	Computation	Total
PC-VSS	$2NQ + 7Q^2 + 4N + 12$	1084
PVSS	$2NQ + 7Q^2 + 3Q + 6 + Q + 4Q + 2$	792
PEM-AFROW	$\left(8 + \frac{4P + 2n_A + 1}{P}\right) N + \frac{1}{P} n_A^2 + \left(4 + \frac{4P + 2}{P}\right) n_A + \frac{P - 1}{P} + 10$	980
TD-PEM-AFROW	$\left(8 + \frac{4P + 2n_A + 1}{P}\right) N + \frac{1}{P} n_A^2 + \left(4 + \frac{4P + 2}{P}\right) n_A + \frac{P - 1}{P}$ $+10 + 6N \log_2 N$	4015
WISE-GRASS-FDAF-PEM-AFROW	$\frac{18M \log_2 M + 25M + 3N + n_A^2 + 4Mn_A + 4 + M}{N}$	416

Table 1: Complexity comparison by the number of FLOPS per recursion. One FFT/IFFT is $3M \log_2 M$ and the phase calculation is an M ($= \arctan(imag/real)$) complexity operation.

4.1. Competing algorithms and tuning parameters

AEC simulations are performed comparing four algorithms namely: the Projection Correlation VSS (PC-VSS) algorithm [17], the Practical VSS affine projection algorithm (PVSS) [22], the Transform-Domain PEM-AFROW (TD-PEM-AFROW) algorithm [31] and the proposed WISE-GRASS-FDAF-PEM-AFROW algorithm. PC-VSS belongs to the gradient-based VSS algorithms and it is claimed to feature the appealing characteristics of being able to distinguish between RIR changes and DT. PC-VSS is the result of improving the algorithm given in [12] to be specifically suited for AEC in DT situations. In PC-VSS, 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 [12] and [38] in DT situations. Moreover, it does not rely on any signal or system model so it is easy to control in practice. The second competing algorithm is the practical VSS affine projection algorithm (PVSS) [22] which stems from the so-called NPVSS proposed in [19]. PVSS takes into account the near-end signal power variations, it is effectively used in DT situations and, it is claimed to be easy to control in practice. PVSS has been shown to outperform the algorithms proposed in [21], [39], and [40]. The third competing algorithms, TDPEM-AFROW, has been investigated in terms of DT robustness in AEC and general improvement in AFC. It has been shown that the combination of a prewhitening of the input and microphone signal together with transform-domain filter adaptation, successfully leads to an algorithm that solves the problem of decorrelation in a very efficient manner. TD-PEM-AFROW is very robust in DT situations and boosts the performance of the simplest AFC (i.e., using only an AR model for the near-end signal).

All the algorithms presented in this paper can be downloaded [2] along with the scripts generating every figure in this section. The tuning parameters in both PVSS and PC-VSS are chosen according to the specifications given in [22] and [17] respectively. The parameters of TD-PEM-AFROW are chosen to have similar initial convergence curves as PVSS and PC-VSS. In the proposed WISE-GRASS-FDAF-PEM-AFROW algorithm, the following parameters 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 values of λ_1 and λ_2 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 λ_2 may be chosen; however, convergence would be slower.

AFC simulations are, on the other hand, performed comparing the original PEM-AFROW [26], TD-PEM-AFROW [31] and the proposed WISE-GRASS-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 $G(q,t)$ consist 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 WISE-GRASS-FDAF-PEM-AFROW, $\lambda_0 = \lambda_2 = 0.99$, $\lambda_1 = 0.1$, and $\mu_{\max} = 0.025$.

The computational complexity of the different algorithms used in the AEC simulations is given in Table 1 using $N = 80$ and APA order $Q = 4$. The computational complexity of WISE-GRASS-FDAF-PEM-AFROW is the lowest and that of TD-PEM-AFROWis the highest. The drawback of most of the FDAF-based algorithms is, however, 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 *misadjustment* (MSD). The MSD between the estimated RIR $\hat{\mathbf{f}}(t)$ and the true RIR $\mathbf{f}_1$ or $\mathbf{f}_2$ represents the accuracy of the estimation and is defined as,

$$\mathrm{MSD}(t) = 10\log_{10}\frac{\left\|\hat{\mathbf{f}}(t) - \mathbf{f}_{1,2}\right\|_2^2}{\|\mathbf{f}_{1,2}\|_2^2} \tag{27}$$

[2] *http : //homes.esat.kuleuven.be/ ∼ pepe*

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 [27] and it is defined as

$$\mathrm{MSG}(t) = -20\log_{10}[\max_{\omega\in\phi}|J(\omega,t)[F(\omega)-\hat{F}(\omega,t)]|] \tag{28}$$

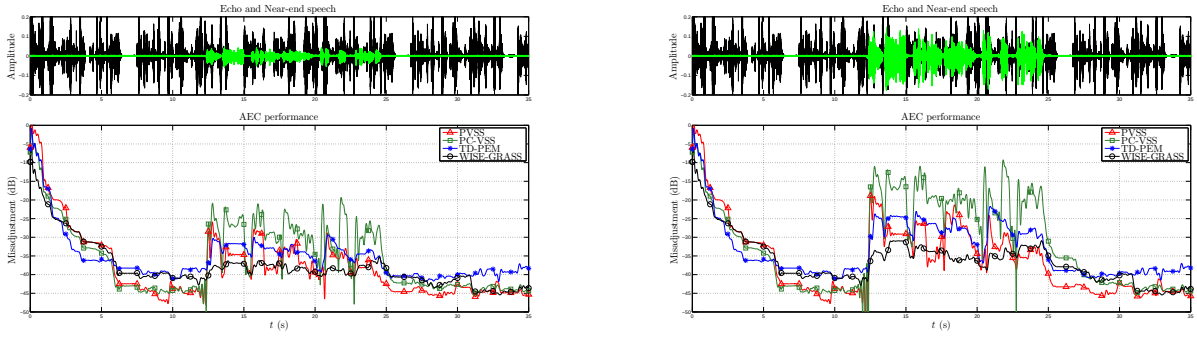
where ϕ denotes the set of frequencies at which the loop phase is a multiple of 2π (i.e., the feedback signal $x(t)$ is in phase with the near-end speech $v(t)$) and $J(\omega,t)$ denotes the forward path processing before the amplifier, i.e., $G(\omega,t) = J(\omega,t)K(t)$.

4.3. Simulation results for DT-robust AEC

In the first set of simulations the noise $n(t)$ is a WGN at 30 dB SNR. PVSS and PC-VSS are first tuned, as [22] and [17] respectively suggested, to obtain the best performance both in terms of convergence rate and final MSD. TD-PEM-AFROW and WISE-GRASS-FDAF-PEM-AFROW are tuned to have similar initial convergence rate as PVSS and PC-VSS. This set of parameters remains unchanged throughout the simulations. All figures in this section consist of an upper part showing the microphone signals and a bottom part showing the AEC performance. The upper part shows the echo (dark blue) and near-end (light green) signals, and the bottom part shows the 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. Results for bursting DT

Figures 4 and 5 show the AEC performance when a bursting DT (from 12.5 s during 12.5 s) occurs at two different SERs (15 dB and 5 dB) immersed in WGN at 30 dB SNR (Figure 4) and immersed in speech babble noise at 20 dB SNR (Figure 5). More specifically, in Figure 4a and 4b a DT burst is shown at 15 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 5 dB improvement in final MSD compared to the other two algorithms during single talk, i.e., -45 to -40 dB respectively. However, during DT, WISE-GRASS-FDAF-PEM-AFROW outperforms the other three algorithms: in Figure 4a by 5 dB in the case of PVSS and TD-PEM-AFROW, and by around 10 dB in the PC-VSS case; in



(a) (b)

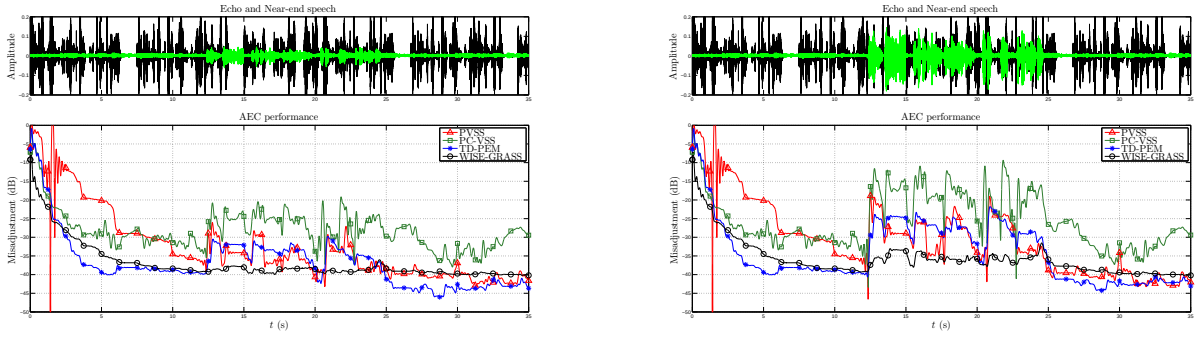
Figure 4: AEC performance in WGN at 30 dB SNR. Bursting DT occurs at different SER: (a) DT at 15 dB SER. (b) DT at 5 dB SER.

Figure 4b these differences are increased since WISE-GRASS-FDAF-PEM-AFROW outperforms TD-PEM-AFROW by $8-10$ dB, PVSS by $5-8$ dB and PC-VSS by $10-20$ dB.

Figure 5a and 5b show two scenarios where the near-end noise is speech babble noise at 20 dB SNR and two DT bursts at 15 and 5 dB SER occur after 12.5 s In these adverse scenarios, the improved performance of WISE-GRASS-FDAF-PEM-AFROW compared to the other three algorithms is demonstrated. The convergence of PVSS is seriously damaged 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. On the other hand, TD-PEM-AFROW and WISE-GRASS-FDAF-PEM-AFROW obtain a smoother and faster convergence. During DT, the MSD of WISE-GRASS-FDAF-PEM-AFROW remains at -40 dB, insensitive to DT. The other three algorithms perform as in the WGN case during DT, which evidences the great difference w.r.t. the WISE-GRASS-FDAF-PEM-AFROW performance. These differences are clearly visible in the case of DT at 5 dB SER. After DT, WISE-GRASS-FDAF-PEM-AFROW restores the low MSD value (-40 dB) although it is outperformed by TD-PEM-AFROW followed by PVSS. PC-VSS seems to have serious problems in speech babble noise since its MSD is almost 10 dB larger than for the other algorithms.

4.3.2. Results of RIR change during bursting DT

Figure 6 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 Figure 6a the AEC performance is shown where



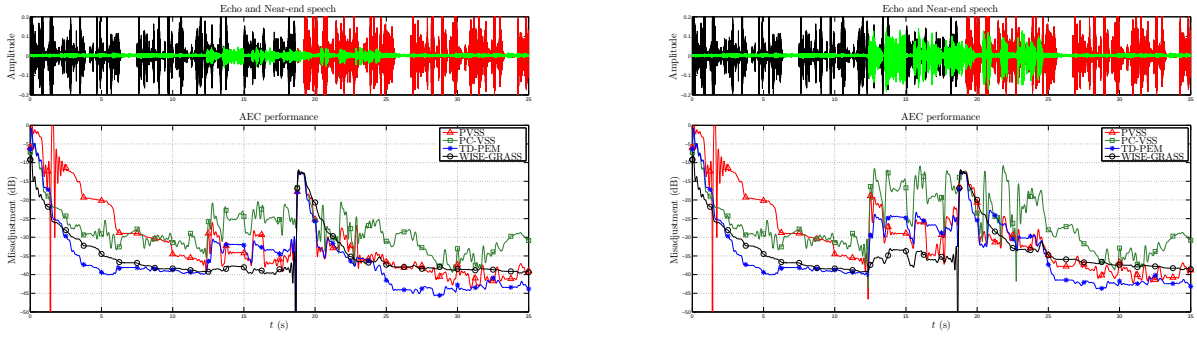
(a) (b)

Figure 5: AEC performance in speech babble noise at 20 dB SNR. Bursting DT occurs at different SER: (a) DT at 15 dB SER. (b) DT at 5 dB SER.

the change of RIR occurs during DT at 15 dB SER and in Figure 6b the SER is 5 dB. The poorest performance of PC-VSS becomes apparent now since the DT contaminates its convergence and hence the MSD is still 15 dB. On the other hand, PVSS and TD-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 Figure 6a and 6b, it is observed that the performance of WISE-GRASS-FDAF-PEM-AFROW is surprisingly almost constant. This fact highlights the robustness of WISE-GRASS-FDAF-PEM-AFROW in an adverse scenario. In fact, its performance is barely degraded which demonstrates the robustness of WISE-GRASS-FDAF-PEM-AFROW as compared to the other three algorithms.

4.3.3. WISE evolution

Figure 7 shows the WISE evolution, i.e., $W(\omega_m, k)$ for $m = 0, ..., 79$ and $k = 1, ..., \frac{L-N}{N}$, in two different scenarios: (1) in Figure 7a speech babble noise at 40 dB SNR and bursting DT at 5 dB SER is used, and (2) in Figure 7b both the loudspeaker signal and the near-end noise are WGN at 20 dB SNR, and an abrupt change of the RIR ($\mathbf{f}_2 = 0.5\mathbf{f}_1$) occurs at the 17.5 s mark. Figure 7a 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. In the bottom figure, the WISE 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 $0 - 10$ are not increased because the echo PSD is also low



(a) (b)

Figure 6: AEC performance in terms of MSD(t) in speech babble noise at 20 dB SNR and an abrupt change of RIR after 17 s occurs during bursting DT (a) DT at 15 dB SER. (b) DT at 5 dB SER.

in the frequency bins and, thus, the Wiener filter gain should be low. The step sizes in frequency bins from 75 and on, are also smaller due to the fact that the echo PSD is lower. In Figure 7b it is shown how the step size follows the echo spectrum 'weighted' by the near-end signal PSD.

4.3.4. Results in highly adverse scenarios

The performance of the algorithms is challenged in Figure 8 where two highly adverse scenarios are proposed: In this DT situation WISE-GRASS-FDAF-PEM-AFROW significantly outperforms PC-VSS which appears to be strongly affected by DT, and moreover shows a significant improvement w.r.t. PVSS which itself outperforms TD-PEM-AFROW. In Figure 8a a change of the RIR occurs which decreases the PVSS performance, while WISE-GRASS-FDAF-PEM-AFROW performance is only decreased w.r.t. Figure 6b. WISE-GRASS-FDAF-PEM-AFROW shows a robust convergence after the RIR change even in a highly adverse DT level obtaining more than 8 dB improvement compared to PVSS after the RIR change. On the other hand Figure 8b shows another highly adverse situation where a continuous DT occurs. The simulation is done using four equal-length segments of speech babble noise at four different SERs $(20, 10, 5, 30$ dB). It shows that, in these noise levels, the convergence of PVSS decreases considerably and PC-VSS diverges as the SER increases. After convergence, WISE-GRASS-FDAF-PEM-AFROW shows the lowest MSD, followed by that of TD-PEM-AFROW ($3-5$ dB above) and then PVSS ($6-7$ dB above). Finally, Figure 8b shows that, after a change of the RIR in continuous DT, the only two algorithms that are able to converge are TD-PEM-AFROW and WISE-GRASS-FDAF-PEM-AFROW.

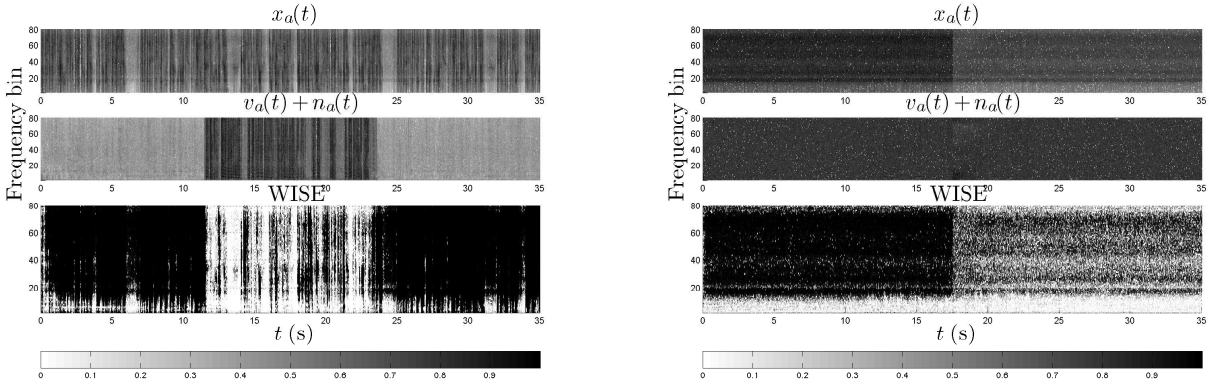


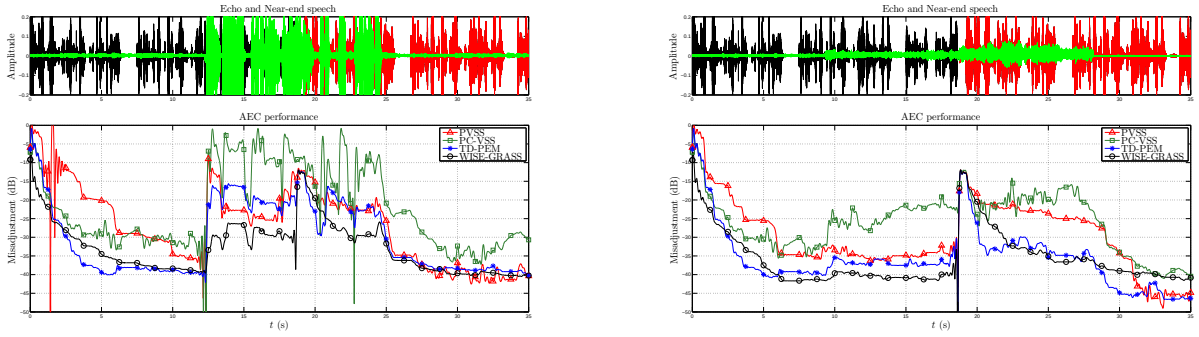
Figure 7: WISE evolution in different scenarios together with the spectrogram of the echo signal and the near-end signals: (a) speech babble noise at 40 dB SNR and bursting DT at 5 dB SER. (b) Both the the near-end noise and the loudspeaker signal are WGN at 20 dB SNR and abrupt change of RIR ($\mathbf{f}_2 = 1/2\mathbf{f}_1$) occurs at the 17.5 s mark.

4.3.5. Results of WISE-only, GRASS-only and WISE-GRASS in FDAF-PEM-AFROW

One figure that can shed some light on why WISE-GRASS-FDAF-PEM-AFROW is so robust and yet performs so well is depicted in Figure 9. A simulation of FDAF-PEM-AFROW using WISE-only, GRASS-only and the proposed combination WISE-GRASS is shown in two adverse scenarios: speech babble noise at 10 dB and 20 dB SNR both with DT at -5 dB SER and with an abrupt change of the RIR in Figure 9a and 9b respectively. In FDAF and stochastic gradient algorithms in general, the excess MSE depends on the step size and noise variance [1], [41]. GRASS-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 WISE-GRASS-FDAF-PEM-AFROW always obtains the best result compared to WISE-only and GRASS-only. It is interesting to note how WISE-GRASS-FDAF-PEM-AFROW takes over GRASS-only FDAF-PEM-AFROW to WISE-only FDAF-PEM-AFROW at around 3 s in Figure 9a and around 5 s in Figure 9b.

4.4. Simulation results for AFC

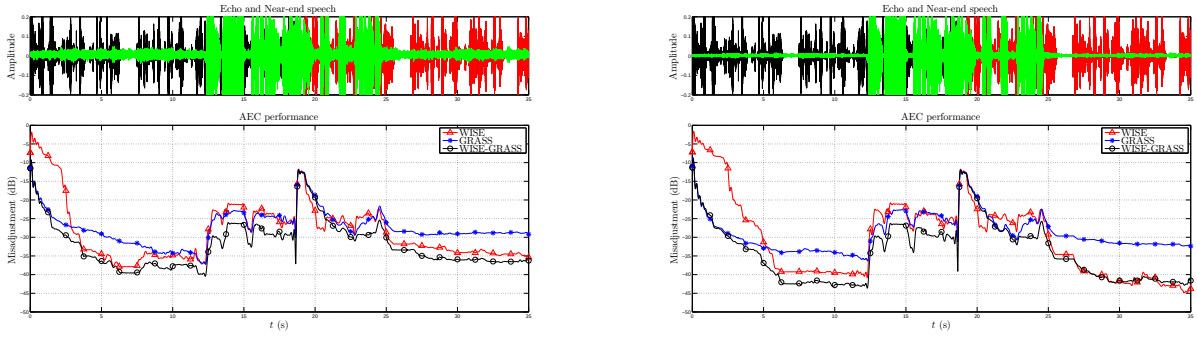
Three AFC scenarios are shown to compare the performance of PEM-AFROW, TD-PEM-AFROW and WISE-GRASS-FDAF-PEM-AFROW. The performance is given in terms of the MSG. The value of $W(\omega_m, k) = 1$ for $k = 1, ..., 20$ because at start-up the estimated $\hat{P}_{X_a}(\omega_m, k)$, and therefore $W(\omega_m, k)$, are very low. WISE-GRASS-FDAF-PEM-AFROW needs some initial iter-



(a) (b)

Figure 8: AEC performance comparison in highly adverse scenarios: (a) speech babble noise at 20 dB SNR and a change of the RIR after 17 s during bursting DT at -5 dB SER. (b) speech babble noise (continuous DT) at $20, 10, 5, 30$ dB SNR and change of RIR after 17 s.

ations until a significant feedback signal PSD is estimated. In Figure 10 the circles represent points of PEM-AFROW, stars represent points of TD-PEM-AFROW and squares represent points of WISE-GRASS-FDAF-PEM-AFROW. More specifically, in Figure 10a 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 the PEM-AFROW achieves $5 - 7.5$ dB MSG improvement and TD-PEM-AFROW achieves $3 - 5$ dB MSG improvement w.r.t to PEM-AFROW. On the other hand, WISE-GRASS-FDAF-PEM-AFROW outperforms the other two algorithms by $8 - 10$ and $5 - 3$ dB respectively. The superior performance of WISE-GRASS-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 shown in Figure 10b. It can be seen that PEM-AFROW goes into the instability region and that TD-PEM-AFROW outperforms PEM-AFROW by $6 - 7$ dB when using low n_A orders. WISE-GRASS-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 Figure 10c, a $n_A = 55$ AR model order is used but, in this case, a speech babble noise at 20 dB SNR is chosen. In this case both PEM-AFROW and TD-PEM-AFROW remain at around 5 dB above the instability region. On the other hand, WISE-GRASS-FDAF-PEM-AFROW maintains its superior performance of $6 - 9$ dB w.r.t. the other two algorithms. In this last scenario, Figure 10d shows the WISE evolution in terms of its instantaneous values (in each frequency bin) as time evolves. It is clearly seen that



(a) (b)

Figure 9: Comparison of WISE-only, GRASS-only and the combination of the two WISE-GRASS. Bursting DT at -5 dB SER and RIR change with continuous speech babble noise at various SNRs: (a) speech babble noise at 10 dB SNR. (b) speech babble noise at 20 dB SNR.

some frequency bins usually have a smaller step size (e.g., frequency bin $5-35$ and $85-100$) than others (e.g., $35-60$ and from $70-85$). It is also interesting to notice that, during the first 7 s, frequency bins $5-30$ have smaller step sizes than after 7 s.

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 (a) the WIener variable Step sizE (WISE) and (b) the GRAdient Spectral variance Smoothing (GRASS) to be performed in FDAF-PEM-AFROW, leading to the WISE-GRASS-FDAF-PEM-AFROW algorithm. The WISE 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 GRASS aims at reducing the variance in the noisy gradient estimates based on time-recursive averaging of gradient estimates. WISE-GRASS-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. Simulations show that WISE-GRASS-FDAF-PEM-AFROW outperforms other competing algorithms in adverse scenarios both in AEC and AFC applications when the near-end signals (i.e., speech and/or background

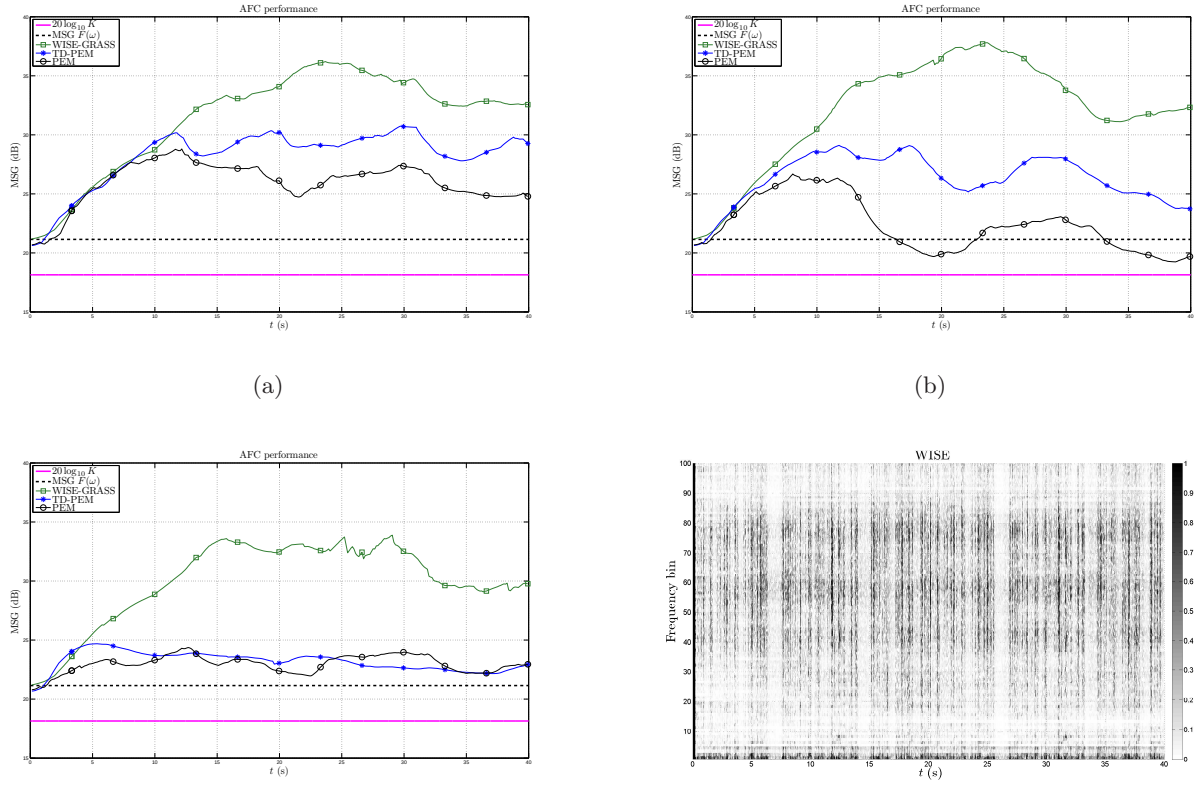


Figure 10: AFC performance in terms of MSG(t) of the three algorithms: PEM-AFROW, TD-PEM-AFROW and WISE-GRASS-FDAF-PEM-AFROW: (a) $n_A = 55$ and WGN at 40 dB SNR. (b) $n_A = 12$ and WGN at 40 dB SNR. (c) $n_A = 55$ and speech babble noise at 20 dB SNR. (d) WISE evolution in the same scenario as (c).

noise) strongly affect the microphone signal. WISE-GRASS-FDAF-PEM-AFROW combines the best characteristics of robust adaptation both in continuous DT and bursting DT without the need for a DTD. WISE-GRASS-FDAF-PEM-AFROW consequently gathers all the characteristic we are seeking for namely, decorrelation properties (PEM, FDAF), minimum variance (GRASS, FDAF, PEM), variable step size (WISE), and computational efficiency (FDAF).

References

[1] S. Haykin, Adaptive Filter Theory, Upper Saddle River, New Jersey: Prentice Hall, 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, no. 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: Proc. 18th European Signal Process. Conf. (EUSIPCO'10), Aalborg, Denmark, 2010.

[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, no. 3 (2007) 846–858.

[5] D. L. Duttweiler, A twelve-channel digital echo canceler, IEEE Trans. Commun. 26, no. 5 (1978) 647–653.

[6] H. K. Jung, N. S. Kim, T. Kim, A new double-talk detector using echo path estimation, Speech Commun. 45, no. 1 (2005) 41–48.

[7] G. Enzner, P. Vary, Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones, Signal Process. 86, no. 6 (2006) 1140–1156.

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

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

[10] 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: Proc. 2008 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'08), Las Vegas, USA, Mar. 2008, pp. 253–256.

[11] T. S. Wada, B.-H. Juang, Enhancement of residual echo cancellation for improved acoustic echo cancellation, in: Proc. 15th European Signal Process. Int. Conf. (EUSIPCO'07), Poznan, Poland, Sep. 2007, pp. 1620–1624.

[12] V. J. Mathews, Z. Xie, A stochastic gradient adaptive filter with gradient adaptive step size, IEEE Trans. Signal Process. 41, no. 6 (1993) 2075–2087.

[13] 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: Proc. 2007 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'07), Vol. 3, Honolulu, Hawaii, USA, Apr. 2007, pp. 1349–1352.

[14] W.-P. Ang, B. Farhang-Boroujeny, A new class of gradient adaptive step-size LMS algorithms, IEEE Trans. Signal Process. 49, no. 4 (2001) 805–810.

[15] H.-C. Shin, A. H. Sayed, W.-J. Song, Variable step-size NLMS and affine projection algorithms, IEEE Signal Process. Lett. 11, no. 2 (2004) 132–135.

[16] Y. Zhang, N. Li, J. A. Chambers, Y. Hao, New gradient-based variable step size lms algorithms, EURASIP J. Advances Signal Process. 2008, no. 105.

[17] T. Creasy, T. Aboulnasr, A projection-correlation algorithm for acoustic echo cancellation in the presence of double talk, in: Proc. 2000 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'00), Vol. 1, Istanbul, Turkey, June 2008, pp. 436–439.

[18] K. Ozeki, T. Umeda, An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties, Electronics and Communication in Japan 67, no. 5 (1984) 19–27.

[19] J. Benesty, H. Rey, L. R. Vega, S. Tressens, A nonparametric VSS NLMS algorithm, IEEE Signal Process. Lett. 13, no. 10 (2006) 581–584.

[20] M. A. Iqbal, S. L. Grant, Novel variable step size NLMS algorithms for echo cancellation, in: Proc. 2008 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'08), Las Vegas, USA, Mar. 2008, pp. 241 – 244.

[21] C. Paleologu, S. Ciochină, J. Benesty, Double-talk robust VSS-NLMS algorithm for under-modeling acoustic echo cancellation, in: Proc. 2008 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'08), Las Vegas, USA, Mar. 2008.

[22] 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, no. 8 (2008) 1466 – 1478.

[23] S. M. Kay, Fundamentals of statistical signal processing: estimation theory, Upper Saddle River, New Jersey: Prentice Hall, 1993.

[24] T. van Waterschoot, M. Moonen, Double-talk robust acoustic echo cancellation with continuous near-end activity, in: Proc. 13th European Signal Process. Conf. (EUSIPCO'05), Antalya, Turkey, 2005.

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

[26] 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, no. 9 (2006) 3426–3434.

[27] T. van Waterschoot, M. Moonen, Fifty years of acoustic feedback control: state of the art and future challenges, Proc. IEEE 99, no. 2 (2011) 288–327.

[28] T. van Waterschoot, M. Moonen, Adaptive feedback cancellation for audio applications, Signal Process. 89, no. 11 (2009) 2185–2201.

[29] 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: Proc. 2010 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'10), Dallas, USA, 2010.

[30] K. Ngo, T. van Waterschoot, M. G. Christensen, M. Moonen, S. H. Jensen, Improved prediction error filters for adaptive feedback cancellation in hearing aids, Elsevier Signal Processing.

[31] 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: Proc. 20th European Signal Process. Int. Conf. (EUSIPCO'12), Bucharest, Rumania, Aug. 2012, pp. 2422–2426.

[32] J. M. Gil-Cacho, T. van Waterschoot, M. Moonen, S. H. Jensen, On the use of the frequency-domain prediction error method for double-talk-robust acoustic echo cancellation, in: Technical Report KULuven, ESAT-SCD, Dec. 2012, pp. 1–27 (online ftp://ftp.esat.kuleuven.be/pub/SISTA/pepegilcacho/reports/gilcacho13_13.pdf).

[33] T. Trump, A frequency domain adaptive algorithm for colored measurement noise environment, in: Proc. 1998 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP'98), Seattle, USA, Mar. 1998, pp. 1705–1708.

[34] J. J. Shynk, Frequency-domain and multirate adaptive filtering, IEEE Signal Process. Mag. 9, no. 1 (1992) 14–37.

[35] N. J. Bershad, P. L. Feintuch, Analysis of the frequency domain adaptive filter, Proc. IEEE 67, no. 12 (1979) 1658–1659.

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

[37] N. Krishnamurthy, J. H. L. Hansen, Babble noise: Modeling, analysis, and applications, IEEE Trans. Audio Speech Lang. Process. 17, no. 7 (2009) 1394–1407.

[38] C. Rohrs, R. Younce, Double talk detector for echo canceler and method (04 1990).
URL `http://www.patentlens.net/patentlens/patent/US4918727/`

[39] 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, no. 6 (2000) 656–663.

[40] 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, no. 5 (2007) 2096–2108.

[41] B. Farhang-Boroujeny, Adaptive filters: Theory and applications, Chichster, UK: Wiley, 1998.