RESEARCH

Cascade algorithms for combined acoustic feedback cancellation and noise reduction

Santiago Ruiz*, Toon van Waterschoot and Marc Moonen

*Correspondence:
santiago.ruiz@esat.kuleuven.be
Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium
Full list of author information is available at the end of the article

Abstract

This paper presents three cascade algorithms for combined acoustic feedback cancellation (AFC) and noise reduction (NR) in speech applications. A prediction error method (PEM) based adaptive feedback cancellation (PEM-based AFC) algorithm is used for the AFC stage while a multichannel Wiener filter (MWF) is applied for the NR stage. A scenario with M microphones and 1 loudspeaker is considered, without loss of generality. The first algorithm is the baseline algorithm, namely the cascade M-channel rank-1 MWF and PEM-AFC, where a NR stage is performed first using a rank-1 MWF followed by a single-channel AFC stage using a PEM-based AFC algorithm. The second algorithm is the cascade $(M+1)$-channel rank-2 MWF and PEM-AFC, where again a NR stage is applied first followed by a single-channel AFC stage. The novelty of this algorithm is to consider an $(M+1)$-channel data model in the MWF formulation with two different desired signals, i.e., the speech component in the reference microphone signal and in the loudspeaker signal, both defined by the speech source signal but not equal to each other. The two desired signal estimates are later used in a single-channel PEM-based AFC stage. The third algorithm is the cascade M-channel PEM-AFC and rank-1 MWF where an M-channel AFC stage is performed first followed by an M-channel NR stage. Although in cascade algorithms where NR is performed first and then AFC the estimation of the feedback path is usually affected by the NR stage, it is shown here that by performing a rank-2 approximation of the speech correlation matrix this issue can be avoided and the feedback path can be correctly estimated. The performance of the algorithms is assessed by means of closed-loop simulations where it is shown that for the considered input signal-to-noise ratios (iSNRs) the cascade $(M+1)$-channel rank-2 MWF and PEM-AFC and the cascade M-channel PEM-AFC and rank-1 MWF algorithms outperform the cascade M-channel rank-1 MWF and PEM-AFC algorithm in terms of the added stable gain (ASG) and misadjustment (Mis) as well as in terms of perceptual metrics such as the short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ) and signal distortion (SD).

Keywords: Combined acoustic feedback cancellation and noise reduction; multichannel Wiener filter; prediction-error method based adaptive feedback cancellation

Introduction

Acoustic feedback and noise are common problems that corrupt microphone signals and affect the performance of speech and audio signal processing applications and devices, such as hearing aids, public address (PA) systems, in-car communication and teleconferencing systems. Acoustic feedback occurs whenever a signal is

captured by a microphone, amplified and played back by a loudspeaker within the same acoustic environment. This acoustic coupling between the microphone (array) and loudspeaker may give rise to instabilities in the system, which translates into signal degradation and, in the worst case, acoustic howling. Different approaches can be found to tackle this problem, with the two most popular being howling suppression and acoustic feedback cancellation (AFC) [1]. AFC solutions rely on a decorrelation of the microphone and loudspeaker signals to obtain an unbiased feedback path estimate [1, 2]. In the literature, many different solutions for AFC can be found using different decorrelation procedures such as probe-noise injection [3], time-varying or nonlinear processes in the forward path [4], null-steering (array)[5, 6] and prewhitening [7]. The latter approach has been shown to provide limited perceptual distortion [8, 9]. Similarly for multi-microphone noise reduction (NR), a wide range of solutions can be found in the literature, where one of the popular algorithms is the multi-channel Wiener filter (MWF) [10, 11, 12], and more recently deep learning-based methods have appeared [13].

Few solutions for combined multi-microphone AFC and NR have been reported in the literature [14, 15]. Similarly to combined acoustic echo cancellation (AEC) and NR, combined AFC and NR can be tackled with integrated and cascade approaches. An integrated approach combines the AFC and NR tasks in a single optimization criterion [14, 15]. A cascade approach consists of an AFC stage and a NR stage which can be combined in two ways, i.e., a multichannel AFC stage followed by a multichannel NR stage, or a single-channel AFC stage preceded by a multichannel NR stage. The order of the stages has performance implications on the combined system [14, 15].

Existing solutions to combined AFC and NR mainly cover single-microphone scenarios [16] and hearing aid applications [14, 5]. In [16] the prediction-error method (PEM)-based adaptive filtering with row operations (PEM-AFROW) algorithm [17] is used in combination with an NR stage based on a minimum mean squared error short-time log-spectral amplitude (MMSE-LSA) estimation, for a single-microphone scenario. In [14] and [15] multiple schemes are presented for combined AFC and NR using a generalised sidelobe canceller (GSC) for the NR stage and a PEM-based AFC stage. In [18] active feedback suppression for one microphone in a hearing aid is proposed using multiple loudspeakers, without considering the presence of noise in the microphone signal. A real-time implementation of a combined NR and feedback suppression method using spectral subtraction in a smartphone-based hearing aid is presented in [19].

In [20], two cascade algorithms are presented for combined multi-microphone AFC and NR for speech applications using a PEM-based AFC algorithm and MWF. The aim of these cascade algorithms is to estimate a desired speech signal without the feedback and noise components, as observed at a chosen reference microphone. A scenario with M microphones and one loudspeaker is considered, without loss of generality. The first algorithm in [20] is the baseline algorithm, namely the cascade M-channel rank-1 MWF and PEM-AFC, where a NR stage is performed first using a rank-1 MWF followed by a single-channel AFC stage using the PEM-based AFC algorithm. It is shown by means of simulations that this algorithm does not improve the added stable gain (ASG) in the closed-loop system. The second algorithm is the

cascade $(M+1)$-channel rank-2 MWF and PEM-AFC where again a NR stage is applied first followed by a single-channel AFC stage. The novelty of this algorithm is to consider an $(M+1)$-channel data model in the MWF formulation (i.e. by including the loudspeaker signal) with two different desired signals, i.e., the speech component in the reference microphone signal and in the loudspeaker signal, both defined by the speech source signal but not equal to each other [12]. The two desired signal estimates are later used in a single-channel PEM-based AFC stage [7, 21]. Although in cascade algorithms where NR is performed first and then AFC, the estimation of the feedback path is usually affected by the NR stage, it is shown in [20] that by performing a rank-2 approximation of the speech correlation matrix this issue can be avoided and the feedback path can be correctly estimated.

In this paper, a third cascade algorithm for AFC and NR using the PEM-based AFC algorithm and MWF is also presented, and then the three algorithms are further analysed and compared. The third algorithm is the cascade M-channel PEM-AFC and rank-1 MWF, where an M-channel AFC stage is performed first followed by an M-channel rank-1 NR stage. A comparison of the performance of the three algorithms is provided based on closed-loop simulations using three different scenarios under three signal-to-noise ratios (SNRs). It is shown that for the considered input SNRs (iSNRs) both the cascade $(M+1)$-channel rank-2 MWF and PEM-AFC and the cascade M-channel PEM-AFC and rank-1 MWF algorithms outperform the cascade M-channel rank-1 MWF and PEM-AFC algorithm in terms of ASG and misadjustment (Mis) as well as in terms of perceptual metrics such as the short-time objective intelligibility (STOI), perceptual evaluation of speech quality (PESQ) and signal distortion (SD). Additionally, the ASG definition is modified to account for the presence of the NR filters in the closed-loop system.

The paper is organized as follows. The signal model is presented in Section . The formulation of the cascade M-channel rank-1 MWF and PEM-AFC algorithm is provided in Section . The cascade $(M+1)$-channel rank-2 MWF and PEM-AFC algorithm is described in Section . The cascade M-channel PEM-AFC and rank-1 MWF is described in Section . Simulation results are given in Section , and finally Section concludes the paper.

Signal model

Consider a room with M microphones and L loudspeakers where the aim is to record a desired speech signal, amplify it and play it back in the loudspeakers. The case when $L=1$ will be considered, without loss of generality, with the speech source signal denoted by $s(t)$, the loudspeaker signal denoted by $u(t)$ and the m^{th} microphone signal, with $m=1,\ldots,M$, modeled as

$$x^{(m)}(t) = H^{(m)}(q,t)s(t) + F^{(m)}(q,t)u(t) + n^{(m)}(t) \tag{1}$$

where $H^{(m)}(q,t)$ and $F^{(m)}(q,t)$ are the transfer function from the speech source position and from the loudspeaker to the m^{th} microphone, respectively. The latter is also known as the feedback path transfer function. The direct noise signal in the m^{th} microphone is denoted by $n^{(m)}(t)$ [1]. The discrete time index is represented by

[1]It is noted that $u(t)$ may also add an additional noise component to $x^{(m)}(t)$, cfr. (3).

t and q^{-1} is the delay operator, i.e., $q^{-k}u(t) = u(t-k)$. The loudspeaker signal can be expressed as

$$u(t) = \sum_{m=1}^{M} G^{(m)}(q,t)x^{(m)}(t), \tag{2}$$

$$u(t) = u_s(t) + u_n(t) \tag{3}$$

where $G^{(m)}(q,t)$ is the forward path transfer function for the m^{th} microphone signal, $u_s(t)$ is the desired speech component and $u_n(t)$ is the noise component in the loudspeaker signal. The presence of the forward path creates a closed-loop system which introduces signal correlation between the loudspeaker and microphone signals. It is assumed that the speech source signal can be modeled as

$$s(t) = \frac{1}{A(q,t)}e(t) \tag{4}$$

where $\frac{1}{A(q,t)}$ is an autoregressive (AR) process excited by the white noise signal $e(t)$, which is a common assumption in PEM-based AFC [1, 7, 21] . A combined NR and AFC algorithm aims to estimate the desired speech signal without the feedback and noise components, as observed at a chosen reference microphone ($m = r$), i.e.,

$$d(t) = H^{(r)}(q,t)s(t) \tag{5}$$

where $H^{(r)}(q,t)$ is the transfer function from the speech source position to the reference microphone.

The short-time Fourier transform (STFT) domain representation of the time-domain signals will be used here, which is obtained by means of an R samples long analysis window in a weighted overlap-add (WOLA) filterbank with 50% overlap [22]. Therefore, the STFT $x^{(m)}(\kappa,l)$ of the m^{th} microphone signal, $x^{(m)}(t)$, at frame l can be defined as

$$\begin{bmatrix} x^{(m)}(0,l) \\ \vdots \\ x^{(m)}(R-1,l) \end{bmatrix} = \boldsymbol{\mathcal{F}}_R \begin{bmatrix} x^{(m)}\left(l\frac{R}{2}\right) g_a(0) \\ \vdots \\ x^{(m)}\left(R-1+l\frac{R}{2}\right) g_a(R-1) \end{bmatrix} \tag{6}$$

with $\kappa \in \{0,1,\ldots,R-1\}$ the frequency bin index, $l \in \{0,1,\ldots,L_f-1\}$ with L_f being the number of frames, $\boldsymbol{\mathcal{F}}_R$ being the discrete Fourier transform (DFT) matrix of size R and $g_a(t)$ being an analysis window. Using the STFT representation of each microphone signal, the following $M \times 1$ STFT-domain microphone vector is defined

$$\mathbf{x}(\kappa,l) = \begin{bmatrix} x^{(1)}(\kappa,l) & \cdots & x^{(M)}(\kappa,l) \end{bmatrix}^T. \tag{7}$$

Furthermore, an $(M+1)\times 1$ signal vector, consisting of loudspeaker and microphone signals, can be expressed as

$$\mathbf{y}(\kappa,l) \triangleq \begin{bmatrix} u(\kappa,l) \\ \mathbf{x}(\kappa,l) \end{bmatrix} = \underbrace{\begin{bmatrix} u_s(\kappa,l) \\ \mathbf{x}_s(\kappa,l) \end{bmatrix}}_{\mathbf{y}_s(\kappa,l)} + \underbrace{\begin{bmatrix} u_n(\kappa,l) \\ \mathbf{x}_n(\kappa,l) \end{bmatrix}}_{\mathbf{y}_n(\kappa,l)} \tag{8}$$

$$= \underbrace{\begin{bmatrix} 0 \\ \mathbf{h}(\kappa,l) \end{bmatrix} s(\kappa,l) + \begin{bmatrix} 1 \\ \mathbf{f}(\kappa,l) \end{bmatrix} u_s(\kappa,l)}_{\mathbf{y}_s(\kappa,l)} + \mathbf{y}_n(\kappa,l) \tag{9}$$

where $s(\kappa,l)$, $u_s(\kappa,l)$, $u(\kappa,l)$ and $\mathbf{y}_n(\kappa,l)$ are the STFT-domain speech source signal, speech component in the loudspeaker signal, loudspeaker signal and noise component in the microphone and loudspeaker signals, respectively. It is noted that $\mathbf{y}_n(\kappa,l)$ includes the noise component in the loudspeaker signal (first vector component), as well as, its coupling into the microphones, added to the direct noise components in the microphones (all other vector components). The STFT-domain transfer functions from the speech source position to the microphones and from the loudspeaker to the microphones are respectively denoted by $\mathbf{h}(\kappa,l)$ and $\mathbf{f}(\kappa,l)$. The time-frame and frequency-bin indices l and κ will be mostly omitted in the following for brevity.

The speech correlation matrix is defined as

$$\bar{\mathbf{R}}_{\mathbf{yy|ss}} = E\{\mathbf{y}_s\mathbf{y}_s^H\} = \begin{bmatrix} 1 & 0 \\ \mathbf{f} & \mathbf{h} \end{bmatrix} \begin{bmatrix} \Phi_{uu} & \Phi_{us} \\ \Phi_{su} & \Phi_{ss} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{f}^H \\ 0 & \mathbf{h}^H \end{bmatrix} \tag{10}$$

where $\Phi_{ss} = E\{s^*s\}$,$\Phi_{su} = E\{s^*u_s\}$, $\Phi_{us} = E\{u_s^*s\} = \Phi_{us}^*$, $\Phi_{uu} = E\{u_s^*u_s\}$, $E\{\cdot\}$ denotes statistical expectation, and $(\cdot)^*$ and $(\cdot)^H$ are the conjugate and conjugate transpose operator, respectively. Performing an LDL factorisation on the matrix with the Φ's in (10), $\bar{\mathbf{R}}_{\mathbf{yy|ss}}$ can alternatively be expressed as

$$\bar{\mathbf{R}}_{\mathbf{yy|ss}} = \begin{bmatrix} 1 & 0 \\ \mathbf{f}+\epsilon\mathbf{h} & \mathbf{h} \end{bmatrix} \begin{bmatrix} \Phi_{uu} & 0 \\ 0 & \Gamma \end{bmatrix} \begin{bmatrix} 1 & \mathbf{f}^H+\epsilon^*\mathbf{h}^H \\ 0 & \mathbf{h}^H \end{bmatrix} \tag{11}$$

where $\epsilon = \dfrac{\Phi_{su}}{\Phi_{uu}}$ and $\Gamma = \Phi_{ss} - \dfrac{\Phi_{su}\Phi_{us}}{\Phi_{uu}}$. It is clear that from the knowledge of $\bar{\mathbf{R}}_{\mathbf{yy|ss}}$ in (11) alone, $\mathbf{f}$ and $\mathbf{h}$ cannot be uniquely defined whenever there is a non-zero correlation between s and u_s.

Three different cascade algorithms are presented in the following sections for AFC and NR. The first algorithm performs an M-channel rank-1 MWF-based NR to estimate the contribution of $s(\kappa,l)$ and $u_s(\kappa,l)$ in the reference microphone, and then a single-channel AFC is performed on the resulting signals. The second algorithm performs an $(M+1)$-channel rank-2 MWF-based NR stage first followed by a single-channel AFC stage, where the rank-2 MWF-based NR is used to estimate the contribution of $s(\kappa,l)$ and $u_s(\kappa,l)$ in the reference microphone as well as in the loudspeaker, and then a single-channel AFC is performed on the resulting signals.

The third algorithm performs an M-channel AFC stage first followed by an M-channel rank-1 MWF-based NR stage. In this case, after the M-channel AFC stage removes the feedback component in each microphone, a rank-1 MWF-based NR is used to estimate the contribution of $s(\kappa, l)$ in the reference microphone.

Cascade M-channel rank-1 MWF and PEM-AFC

NR stage

The objective of the NR stage is to provide an estimate of the speech component in the reference microphone signal. The feedback component will still be present in the output of the NR stage, hence a single-channel AFC stage is required to remove it.

In the STFT domain, the correlation matrix of the microphone signal vector $\mathbf{x}$ can be expressed as

$$\bar{\mathbf{R}}_{\mathbf{xx}} = E\{\mathbf{x}\mathbf{x}^H\} = \bar{\mathbf{R}}_{\mathbf{xx|ss}} + \bar{\mathbf{R}}_{\mathbf{xx|nn}} \tag{12}$$

where

$$\bar{\mathbf{R}}_{\mathbf{xx|ss}} = E\{\mathbf{x}_s\mathbf{x}_s^H\} = E\{(\mathbf{h}s + \mathbf{f}u_s)(\mathbf{h}s + \mathbf{f}u_s)^H\}, \tag{13}$$

$$\bar{\mathbf{R}}_{\mathbf{xx|nn}} = E\{\mathbf{x_n}\mathbf{x_n}^H\} \tag{14}$$

are the $M \times M$ microphone-only speech and noise correlation matrix, respectively. The expressions in (12)-(14) are obtained based on the assumption that s and $\mathbf{x}_n$ are uncorrelated. The minimization of the mean squared error (MSE) between the desired signal and the filtered microphone signals defines an optimal filter

$$\bar{\mathbf{w}} = \min_{\mathbf{w}} \; E\left\{\left\|d - \mathbf{w}^H\mathbf{x}\right\|^2\right\} \tag{15}$$

with $d = x_s^{(r)}$ representing the speech component (total contribution of s together with u_s) in the reference microphone signal. The desired signal estimate $\hat{d}$ is obtained as

$$\hat{d} = \bar{\mathbf{w}}^H\mathbf{x}. \tag{16}$$

The solution to (15) is the MWF [12, 10], given by

$$\bar{\mathbf{w}} = \bar{\mathbf{R}}_{\mathbf{xx}}^{-1}\bar{\mathbf{R}}_{\mathbf{xx|ss}}\mathbf{e}_r \tag{17}$$

where $\mathbf{e}_r$ selects the r^{th} column of $\bar{\mathbf{R}}_{\mathbf{xx}}^{-1}\bar{\mathbf{R}}_{\mathbf{xx|ss}}$.

In practice, by using a voice activity detector (VAD), $\bar{\mathbf{R}}_{\mathbf{xx}}$ and $\bar{\mathbf{R}}_{\mathbf{xx|nn}}$ are first estimated during *speech-plus-noise* periods where the speech source signal and noise are active, and *noise-only* periods where only the noise is active, i.e.,

$$\begin{aligned}
&\text{if VAD}(\kappa, l) = 1 : \\
&\hat{\mathbf{R}}_{\mathbf{xx}}(\kappa, l) = \beta\hat{\mathbf{R}}_{\mathbf{xx}}(\kappa, l-1) + (1-\beta)\mathbf{x}(\kappa, l)\mathbf{x}^H(\kappa, l), \\
&\text{if VAD}(\kappa, l) = 0 : \\
&\hat{\mathbf{R}}_{\mathbf{xx|nn}}(\kappa, l) = \beta\hat{\mathbf{R}}_{\mathbf{xx|nn}}(\kappa, l-1) + (1-\beta)\mathbf{x}(\kappa, l)\mathbf{x}^H(\kappa, l),
\end{aligned} \tag{18}$$

where $\hat{\mathbf{R}}_{\mathbf{xx}}(\kappa,l)$ and $\hat{\mathbf{R}}_{\mathbf{xx|nn}}(\kappa,l)$ represent estimates of $\bar{\mathbf{R}}_{\mathbf{xx}}$ and $\bar{\mathbf{R}}_{\mathbf{xx|nn}}$ at frame l and frequency bin κ, respectively. The forgetting factor $0<\beta<1$ can be chosen depending on the variation of the statistics of the signals, i.e., if the statistics change slowly then β should be chosen close to 1 to obtain long-term estimates that mainly capture the spatial coherence between the microphone signals. The following criterion will then be used to estimate $\bar{\mathbf{R}}_{\mathbf{xx|ss}}$ [12],

$$\hat{\mathbf{R}}_{\mathbf{xx|ss}} = \min_{\mathbf{R}_{\mathbf{xx|ss}}} \left\| \hat{\mathbf{R}}_{\mathbf{xx|nn}}^{-1/2} \left(\hat{\mathbf{R}}_{\mathbf{xx}} - \hat{\mathbf{R}}_{\mathbf{xx|nn}} - \mathbf{R}_{\mathbf{xx|ss}} \right) \hat{\mathbf{R}}_{\mathbf{xx|nn}}^{-H/2} \right\|_F^2 \tag{19}$$

$$\text{s.t.} \quad \text{rank}(\mathbf{R}_{\mathbf{xx|ss}}) = 1, \quad \mathbf{R}_{\mathbf{xx|ss}} \succeq 0 \tag{20}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Spatial pre-whitening is applied by pre- and post-multiplying by $\hat{\mathbf{R}}_{\mathbf{xx|nn}}^{-1/2}$ and $\hat{\mathbf{R}}_{\mathbf{xx|nn}}^{-H/2}$, respectively. The solution to (19)-(20) is based on a generalized eigenvalue decomposition (GEVD) of the $(M\times M)$ matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{xx}}, \hat{\mathbf{R}}_{\mathbf{xx|nn}}\}$ [12, 23]

$$\hat{\mathbf{R}}_{\mathbf{xx}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{xx}}\hat{\mathbf{Q}}^H \tag{21}$$

$$\hat{\mathbf{R}}_{\mathbf{xx|nn}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{xx|nn}}\hat{\mathbf{Q}}^H \tag{22}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{xx|nn}}$ are diagonal matrices and $\hat{\mathbf{Q}}$ is an invertible matrix. The rank-1 speech correlation matrix estimate $\hat{\mathbf{R}}_{\mathbf{xx|ss}}$ is then [12]

$$\hat{\mathbf{R}}_{\mathbf{xx|ss}} = \hat{\mathbf{Q}}\text{diag}\{\hat{\sigma}_{xx,1} - \hat{\sigma}_{xx|nn,1}, 0, \ldots, 0\}\hat{\mathbf{Q}}^H \tag{23}$$

where $\hat{\sigma}_{xx,i}$ and $\hat{\sigma}_{xx|nn,i}$ are the ith diagonal element of $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{xx|nn}}$, respectively, corresponding to the ith largest ratio $\hat{\sigma}_{xx,i}/\hat{\sigma}_{xx|nn,i}$. Using (23) and $\hat{\mathbf{R}}_{\mathbf{xx}}$ (cfr. (21)) in (17), the rank-1 MWF estimate $\hat{\mathbf{w}}$ can be expressed as

$$\hat{\mathbf{w}} = \hat{\mathbf{Q}}^{-H}\text{diag}\left\{1 - \frac{\hat{\sigma}_{xx|nn,1}}{\hat{\sigma}_{xx,1}}, 0, \ldots, 0\right\}\hat{\mathbf{Q}}^H\mathbf{e}_r. \tag{24}$$

The estimate, $\hat{x}_s^{(r)}$, is obtained as in (16) with $\hat{\mathbf{w}}$ replacing $\bar{\mathbf{w}}$

$$\hat{x}_s^{(r)} = \hat{\mathbf{w}}^H\mathbf{x}. \tag{25}$$

The corresponding time-domain signals are obtained by adding the L_f overlapping windowed frames as

$$\hat{\mathbf{x}}_{s,seg}^{(r)}(l) = \boldsymbol{\mathcal{F}}_R^{-1}\left[\hat{x}_s^{(r)}(0,l) \quad \cdots \quad \hat{x}_s^{(r)}(R-1,l)\right]^T, \tag{26}$$

$$\hat{\mathbf{x}}_{s,seg}^{(r)}(l) = \left[\hat{x}_{s,seg}^{(r)}\left(l\frac{R}{2}\right), \ldots, \hat{x}_{s,seg}^{(r)}\left(R-1+l\frac{R}{2}\right)\right]^T, \tag{27}$$

$$\hat{x}_s^{(r)}(t-d_{\text{NR}}) = \sum_{l=0}^{L_f-1} \hat{x}_{s,seg}^{(r)}\left(t-l\frac{R}{2}\right) g_s\left(t-l\frac{R}{2}\right) \tag{28}$$

where g_s is a synthesis window with nonzero values in the interval $0 \leq t \leq R-1$ and d_{NR} is the delay from the NR stage.

AFC stage

The NR stage provides an estimate for

$$x_s^{(r)}(t) = H^{(r)}(q,t)s(t) + F^{(r)}(q,t)u_s(t), \tag{29}$$

from which the AFC stage will now estimate $H^{(r)}(q,t)s(t)$. A single-channel PEM-based AFC algorithm is used. This kind of algorithms were initially developed in [7, 17] and they provide estimates of both the feedback path and the speech source signal model. The PEM-based AFC algorithm used here is the frequency-domain version presented in [21] (the reader is referred to [21] for a detailed explanation of the AFC algorithm). The algorithm uses an overlap-save (OLS) filterbank to compute convolutions in the frequency domain, which requires a rectangular window. The input signals to the AFC algorithm are the (noisy) loudspeaker signal u and the estimate in (28). A short description of the single-channel PEM-based AFC algorithm is provided in Algorithm 1.

A complete description of the cascade M-channel rank-1 MWF and PEM-AFC algorithm is provided in Algorithm 2, with a block diagram provided in Figure 1(a).

Cascade $(M+1)$-channel rank-2 MWF and PEM-AFC

NR stage

The objective of the NR stage is to provide an estimate of the speech component in the reference microphone signal and in the loudspeaker signal. The feedback component will still be present in the former, hence a single-channel AFC stage is required to remove it.

In the STFT domain, the correlation matrix of the signal vector $\mathbf{y}$ in (8) can be expressed as

$$\bar{\mathbf{R}}_{\mathbf{yy}} = E\{\mathbf{y}\mathbf{y}^H\} = \bar{\mathbf{R}}_{\mathbf{yy|ss}} + \bar{\mathbf{R}}_{\mathbf{yy|nn}} \tag{30}$$

with $\bar{\mathbf{R}}_{\mathbf{yy|nn}} = E\{\mathbf{y}_n\mathbf{y}_n^H\}$ the $(M+1)\times(M+1)$ noise correlation matrix. The final expression in (30) is obtained based on the assumption that s and $\mathbf{n}$ are uncorrelated. The minimization of the mean squared error (MSE) between the desired signals and the filtered microphone and loudspeaker signals defines an optimal filter

$$\bar{\mathbf{W}} = \min_{\mathbf{W}} \; E\left\{\left\|\mathbf{d} - \mathbf{W}^H\mathbf{y}\right\|^2\right\}. \tag{31}$$

with $\mathbf{d} = \begin{bmatrix} u_s & x_s^{(r)} \end{bmatrix}^T$. The desired signal estimates $\hat{u}_s$ and $\hat{x}_s^{(r)}$ are obtained as

$$\hat{u}_s = \mathbf{e}_1^H\bar{\mathbf{W}}^H\mathbf{y}, \tag{32}$$

$$\hat{x}_s^{(r)} = \mathbf{e}_2^H\bar{\mathbf{W}}^H\mathbf{y}. \tag{33}$$

The solution to (31) is the MWF [12, 10], given by

$$\bar{\mathbf{W}} = \bar{\mathbf{R}}_{\mathbf{yy}}^{-1}\bar{\mathbf{R}}_{\mathbf{yy|ss}}[\mathbf{e}_1|\mathbf{e}_{r+1}] \tag{34}$$

where $\mathbf{e}_{r+1}$ selects the $(r+1)^{\text{st}}$ column of a matrix.

In practice, by using a voice activity detector (VAD), $\bar{\mathbf{R}}_{\mathbf{yy}}$ and $\bar{\mathbf{R}}_{\mathbf{yy|nn}}$ are first estimated during *speech-plus-noise* periods where the desired speech signal and noise are active, and *noise-only* periods where only the noise is active, i.e.,

$$\begin{aligned}
&\text{if VAD}(\kappa,l) = 1 : \\
&\hat{\mathbf{R}}_{\mathbf{yy}}(\kappa,l) = \beta\hat{\mathbf{R}}_{\mathbf{yy}}(\kappa,l-1) + (1-\beta)\mathbf{y}(\kappa,l)\mathbf{y}^H(\kappa,l), \\
&\text{if VAD}(\kappa,l) = 0 : \\
&\hat{\mathbf{R}}_{\mathbf{yy|nn}}(\kappa,l) = \beta\hat{\mathbf{R}}_{\mathbf{yy|nn}}(\kappa,l-1) + (1-\beta)\mathbf{y}(\kappa,l)\mathbf{y}^H(\kappa,l)
\end{aligned} \tag{35}$$

where $\hat{\mathbf{R}}_{\mathbf{yy}}(\kappa,l)$ and $\hat{\mathbf{R}}_{\mathbf{yy|nn}}(\kappa,l)$ represent estimates of $\bar{\mathbf{R}}_{\mathbf{yy}}$ and $\bar{\mathbf{R}}_{\mathbf{yy|nn}}$ at frame l and frequency bin κ, respectively. The following criterion will then be used to estimate $\bar{\mathbf{R}}_{\mathbf{yy|ss}}$ [12],

$$\hat{\mathbf{R}}_{\mathbf{yy|ss}} = \min_{\mathbf{R}_{\mathbf{yy|ss}}} \left\| \hat{\mathbf{R}}_{\mathbf{yy|nn}}^{-1/2} \left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{yy|nn}} - \mathbf{R}_{\mathbf{yy|ss}} \right) \hat{\mathbf{R}}_{\mathbf{yy|nn}}^{-H/2} \right\|_F^2 \tag{36}$$

$$\text{s.t.} \quad \text{rank}(\mathbf{R}_{\mathbf{yy|ss}}) = 2, \quad \mathbf{R}_{\mathbf{yy|ss}} \succeq 0. \tag{37}$$

Spatial pre-whitening is applied by pre- and post-multiplying by $\hat{\mathbf{R}}_{\mathbf{yy|nn}}^{-1/2}$ and $\hat{\mathbf{R}}_{\mathbf{yy|nn}}^{-H/2}$, respectively. The solution to (36)-(37) is based on a GEVD of the $(M+1)\times(M+1)$ matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{yy|nn}}\}$ [12, 23]

$$\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{yy}}\hat{\mathbf{Q}}^H \tag{38}$$

$$\hat{\mathbf{R}}_{\mathbf{yy|nn}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{yy|nn}}\hat{\mathbf{Q}}^H \tag{39}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{yy}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{yy|nn}}$ are diagonal matrices and $\hat{\mathbf{Q}}$ is an invertible matrix. The rank-2 speech correlation matrix estimate $\hat{\mathbf{R}}_{\mathbf{yy|ss}}$ is then [12]

$$\hat{\mathbf{R}}_{\mathbf{yy|ss}} = \hat{\mathbf{Q}}\text{diag}\{\hat{\sigma}_{yy,1} - \hat{\sigma}_{yy|nn,1}, \hat{\sigma}_{yy,2} - \hat{\sigma}_{yy|nn,2}, 0, \ldots, 0\}\hat{\mathbf{Q}}^H \tag{40}$$

where $\hat{\sigma}_{yy,i}$ and $\hat{\sigma}_{yy|nn,i}$ are the ith diagonal element of $\hat{\mathbf{\Sigma}}_{\mathbf{yy}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{yy|nn}}$, respectively, corresponding to the ith largest ratio $\hat{\sigma}_{yy,i}/\hat{\sigma}_{yy|nn,i}$. Using (40) and $\hat{\mathbf{R}}_{\mathbf{yy}}$ (cfr. (38)) in (34), the rank-2 MWF estimate $\hat{\mathbf{W}}$ can be expressed as

$$\hat{\mathbf{W}} = \hat{\mathbf{Q}}^{-H}\text{diag}\left\{1 - \frac{\hat{\sigma}_{yy|nn,1}}{\hat{\sigma}_{yy,1}}, 1 - \frac{\hat{\sigma}_{yy|nn,2}}{\hat{\sigma}_{yy,2}}, 0, \ldots, 0\right\}\hat{\mathbf{Q}}^H[\mathbf{e}_1|\mathbf{e}_{r+1}]. \tag{41}$$

The estimates $\hat{u}_s$ and $\hat{x}_s^{(r)}$, are now obtained as in (32)-(33) with $\hat{\mathbf{W}}$ replacing $\bar{\mathbf{W}}$

$$\hat{u}_s = \mathbf{e}_1^H \hat{\mathbf{W}}^H \mathbf{y}, \tag{42}$$
$$\hat{x}_s^{(r)} = \mathbf{e}_2^H \hat{\mathbf{W}}^H \mathbf{y}. \tag{43}$$

The corresponding time-domain signals are obtained by adding the L_f overlapping windowed frames as

$$\hat{\mathbf{x}}_{s,seg}^{(r)}(l) = \boldsymbol{\mathcal{F}}_R^{-1} \left[\hat{x}_s^{(r)}(0,l) \quad \cdots \quad \hat{x}_s^{(r)}(R-1,l) \right]^T, \tag{44}$$
$$\hat{\mathbf{u}}_{s,seg}(l) = \boldsymbol{\mathcal{F}}_R^{-1} \left[\hat{u}_s(0,l) \quad \cdots \quad \hat{u}_s(R-1,l) \right]^T, \tag{45}$$
$$\hat{\mathbf{x}}_{s,seg}^{(r)}(l) = \left[\hat{x}_{s,seg}^{(r)}\left(l\frac{R}{2}\right), \ldots, \hat{x}_{s,seg}^{(r)}\left(R-1+l\frac{R}{2}\right) \right]^T, \tag{46}$$
$$\hat{\mathbf{u}}_{s,seg}(l) = \left[\hat{u}_{s,seg}\left(l\frac{R}{2}\right), \ldots, \hat{u}_{s,seg}\left(R-1+l\frac{R}{2}\right) \right]^T, \tag{47}$$
$$\hat{x}_s^{(r)}(t-d_{\mathrm{NR}}) = \sum_{l=0}^{L_f-1} \hat{x}_{s,seg}^{(r)}\left(t-l\frac{R}{2}\right) g_s\left(t-l\frac{R}{2}\right), \tag{48}$$
$$\hat{u}_s(t-d_{\mathrm{NR}}) = \sum_{l=0}^{L_f-1} \hat{u}_{s,seg}\left(t-l\frac{R}{2}\right) g_s\left(t-l\frac{R}{2}\right). \tag{49}$$

AFC stage

In the AFC stage a single-channel PEM-based AFC algorithm is used. The PEM-based AFC algorithm used here is the frequency-domain version presented in [21]. The input signals to the AFC algorithm are $\hat{u}_s$ and $\hat{x}_s^{(r)}$. A short description of the PEM-based AFC algorithm is provided in Algorithm 1.

A complete description of the cascade $(M+1)$-channel rank-2 MWF and PEM-AFC algorithm is provided in Algorithm 3, with block diagram provided in Figure 1(b).

Cascade M-channel PEM-AFC and rank-1 MWF

Assuming an exact speech signal model $A^{-1}(q,t)$ is available (see (4)), a prefilter $A(q,t)$ can be applied, such that the time-domain prefiltered loudspeaker and m^{th} microphone signal can be expressed as

$$\tilde{u}(t) = A(q,t)u(t), \tag{50}$$
$$\tilde{x}^{(m)}(t) = A(q,t)x^{(m)}(t). \tag{51}$$

Similarly, the prefiltered version of the signal vector $\mathbf{y}$ in (8) can be expressed as

$$\tilde{\mathbf{y}}(\kappa,l) = \begin{bmatrix} \tilde{u}(\kappa,l) \\ \tilde{\mathbf{x}}(\kappa,l) \end{bmatrix} \tag{52}$$
$$= \begin{bmatrix} 0 \\ \mathbf{h}(\kappa,l) \end{bmatrix} e(\kappa,l) + \begin{bmatrix} 1 \\ \mathbf{f}(\kappa,l) \end{bmatrix} \tilde{u}_s(\kappa,l) + \tilde{\mathbf{y}}_n(\kappa,l) \tag{53}$$

where $\tilde{u}(\kappa,l)$ and $\tilde{\mathbf{x}}(\kappa,l)$ represent the STFT-domain prefiltered loudspeaker and microphone signals. Similarly, $\tilde{u}_s(\kappa,l)$ is the STFT-domain prefiltered desired speech component in the loudspeaker signal and $\tilde{\mathbf{y}}_n(\kappa,l)$ is the STFT-domain prefiltered noise component in the loudspeaker and microphone signals. The speech correlation matrix can be rewritten as

$$\bar{\mathbf{R}}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}|\mathbf{ss}} = \begin{bmatrix} 1 & 0 \\ \mathbf{f} & \mathbf{h} \end{bmatrix} \begin{bmatrix} \Phi_{\tilde{u}\tilde{u}} & 0 \\ 0 & \Phi_{ee} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{f}^H \\ 0 & \mathbf{h}^H \end{bmatrix} \tag{54}$$

$$= \begin{bmatrix} 1 \\ \mathbf{f} \end{bmatrix} \Phi_{\tilde{u}\tilde{u}} \begin{bmatrix} 1 & \mathbf{f}^H \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{h} \end{bmatrix} \Phi_{ee} \begin{bmatrix} 0 & \mathbf{h}^H \end{bmatrix} \tag{55}$$

where $\Phi_{\tilde{u}\tilde{u}} = E\{\tilde{u}^*\tilde{u}\}$, $\Phi_{ee} = E\{e^H e\}$, $\Phi_{e\tilde{u}} = E\{e^*\tilde{u}\} = 0$ and $\Phi_{\tilde{u}e} = E\{\tilde{u}^* e\} = 0$, with $\tilde{u} \approx \tilde{u}_s$. Since (54) is computed in the STFT domain, the cross-correlation terms would only be zero if there is a delay of at least one STFT-frame in the forward path. It can be observed that, after prefiltering, $\mathbf{h}$ and $\mathbf{f}$ can be readily computed from $\bar{\mathbf{R}}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}|\mathbf{ss}}$. In this case, the order of the AFC and NR stages can be inverted so that an M-channel AFC stage is performed first, which will estimate the speech component together with the noise component, and then a multichannel NR stage can follow.

AFC stage

In the AFC stage a single-channel PEM-based AFC algorithm is used for each microphone, i.e., M times. The AR model is estimated for each single-channel PEM-based AFC algorithm. The same step-size tunning is used for all adaptive algorithms. The PEM-based AFC algorithm used here is the frequency-domain version presented in [21]. The input signals to the AFC algorithm are u and $x^{(m)}, \forall m$. A short description of the PEM-based AFC algorithm is provided in Algorithm 1.

NR stage

A rank-1 MWF is used for the NR stage which operates on the microphone signals after the AFC stage.

The STFT domain representation of the time-domain signals will be used here, which is obtained by means of an R samples long analysis window in a WOLA filterbank with 50% overlap [22]. Therefore, the STFT $x_f^{(m)}(\kappa,l)$ of the m^{th} microphone signal after the AFC stage, $x_f^{(m)}(t)$, at frame l can be defined as

$$\begin{bmatrix} x_f^{(m)}(0,l) \\ \vdots \\ x_f^{(m)}(R-1,l) \end{bmatrix} = \boldsymbol{\mathcal{F}}_R \begin{bmatrix} x_f^{(m)}\left(l\frac{R}{2}\right) g_a(0) \\ \vdots \\ x_f^{(m)}\left(R-1+l\frac{R}{2}\right) g_a(R-1) \end{bmatrix}. \tag{56}$$

The STFT-domain multi-channel microphone signal after the AFC stage, assuming perfect feedback cancellation, is modeled as

$$\mathbf{x}_f(\kappa,l) = \mathbf{h}(\kappa,l)s(\kappa,l) + \mathbf{x}_{f|n}(\kappa,l) \tag{57}$$

where $\mathbf{x}_{f|n}(\kappa,l)$ is the STFT-domain noise component in the microphone signal after feedback cancellation. The minimization of the mean squared error (MSE) between the desired signal and the filtered feedback-compensated microphone signals, $\mathbf{x}_f$, defines an optimal filter

$$\bar{\mathbf{w}} = \min_{\mathbf{w}} \; E\left\{\left\|d - \mathbf{w}^H\mathbf{x}_f\right\|^2\right\} \tag{58}$$

with $d = x_{f|s}^{(r)}$. The desired signal estimate is then obtained as $\hat{d} = \bar{\mathbf{w}}^H\mathbf{x}_f$. The solution to (58) is the well-known MWF [12, 10], given by

$$\bar{\mathbf{w}} = \bar{\mathbf{R}}_{\mathbf{x_f x_f}}^{-1}\bar{\mathbf{R}}_{\mathbf{x_f x_f|ss}}\mathbf{e}_r \tag{59}$$

where $\bar{\mathbf{R}}_{\mathbf{x_f x_f}} = E\{\mathbf{x}_f\mathbf{x}_f^H\}$, $\bar{\mathbf{R}}_{\mathbf{x_f x_f|ss}} = E\{\mathbf{h}ss^H\mathbf{h}^H\}$, and, similarly, $\bar{\mathbf{R}}_{\mathbf{x_f x_f|nn}} = E\{\mathbf{x}_{f|n}\mathbf{x}_{f|n}^H\}$. The final expression in (59) is obtained based on the assumption that s and $\mathbf{x}_{f|n}$ are uncorrelated.

In practice, by using a voice activity detector (VAD), $\bar{\mathbf{R}}_{\mathbf{x_f x_f}}$ and $\bar{\mathbf{R}}_{\mathbf{x_f x_f|nn}}$ are first estimated during *speech-plus-noise* periods where the desired speech signal and background noise are active, and *noise-only* periods where only the noise is active [24], i.e.,

$$\text{if VAD}(\kappa,l) = 1: \tag{60}$$
$$\hat{\mathbf{R}}_{\mathbf{x_f x_f}}(\kappa,l) = \beta\hat{\mathbf{R}}_{\mathbf{x_f x_f}}(\kappa,l-1) + (1-\beta)\mathbf{x}_f(\kappa,l)\mathbf{x}_f^H(\kappa,l),$$
$$\text{if VAD}(\kappa,l) = 0: \tag{61}$$
$$\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}(\kappa,l) = \beta\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}(\kappa,l-1) + (1-\beta)\mathbf{x}_f(\kappa,l)\mathbf{x}_f^H(\kappa,l) \tag{62}$$

where $\hat{\mathbf{R}}_{\mathbf{x_f x_f}}(\kappa,l)$ and $\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}(\kappa,l)$ represent estimates of $\bar{\mathbf{R}}_{\mathbf{x_f x_f}}$ and $\bar{\mathbf{R}}_{\mathbf{x_f x_f|nn}}$ at frame l and frequency bin index κ, respectively. The following criterion will then be used to estimate $\bar{\mathbf{R}}_{\mathbf{x_f x_f|ss}}$ [12],

$$\hat{\mathbf{R}}_{\mathbf{x_f x_f|ss}} = \min_{\mathbf{R}_{\mathbf{x_f x_f|ss}}} \left\|\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}^{-1/2}\left(\hat{\mathbf{R}}_{\mathbf{x_f x_f}} - \hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}} - \mathbf{R}_{\mathbf{x_f x_f|ss}}\right)\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}^{-H/2}\right\|_F^2 \tag{63}$$

$$\text{s.t.} \quad \text{rank}(\mathbf{R}_{\mathbf{x_f x_f|ss}}) = 1, \quad \mathbf{R}_{\mathbf{x_f x_f|ss}} \succeq 0. \tag{64}$$

Spatial pre-whitening is applied by pre- and post-multiplying by $\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}^{-1/2}$ and $\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}^{-H/2}$, respectively. The solution to (63)-(64) is based on a GEVD of the $(M\times M)$ matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{x_f x_f}}, \hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}\}$ [12, 23]

$$\hat{\mathbf{R}}_{\mathbf{x_f x_f}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{x_f x_f}}\hat{\mathbf{Q}}^H, \tag{65}$$
$$\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{x_f x_f|nn}}\hat{\mathbf{Q}}^H \tag{66}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{x_f x_f}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{x_f x_f|nn}}$ are diagonal matrices and $\hat{\mathbf{Q}}$ is an invertible matrix. The speech correlation matrix estimate $\hat{\mathbf{R}}_{\mathbf{x_f x_f|ss}}$ is then [12]

$$\hat{\mathbf{R}}_{\mathbf{x_f x_f|ss}} = \hat{\mathbf{Q}}\text{diag}\{\hat{\sigma}_{x_f x_f,1} - \hat{\sigma}_{x_f x_f|nn,1}, 0, \ldots, 0\}\hat{\mathbf{Q}}^H \tag{67}$$

where $\hat{\sigma}_{x_f x_f,1}$ and $\hat{\sigma}_{x_f x_f|nn,1}$ are the first diagonal element of $\hat{\mathbf{\Sigma}}_{\mathbf{x_f x_f}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{x_f x_f|nn}}$, respectively, corresponding to the largest ratio $\hat{\sigma}_{x_f x_f,i}/\hat{\sigma}_{x_f x_f|nn,i}$. Using (67) and $\hat{\mathbf{R}}_{\mathbf{x_f x_f}}$ (cfr. (65)) in (59), the rank-1 MWF estimate $\hat{\mathbf{w}}$ can be expressed as

$$\hat{\mathbf{w}} = \hat{\mathbf{Q}}^{-H}\text{diag}\left\{1 - \frac{\hat{\sigma}_{x_f x_f|nn,1}}{\hat{\sigma}_{x_f x_f,1}}, 0, \ldots, 0\right\}\hat{\mathbf{Q}}^H\mathbf{e}_r. \tag{68}$$

The desired signal estimate is then obtained as $\hat{d} = \hat{\mathbf{w}}^H\mathbf{x}_f$. The time-domain desired signal is obtained by adding the L_f overlapping windowed frames as

$$\hat{\mathbf{d}}_{seg}(l) = \boldsymbol{\mathcal{F}}_R^{-1}\left[\hat{d}\,(0,l) \quad \cdots \quad \hat{d}\,(R-1,l)\right]^T, \tag{69}$$

$$\hat{\mathbf{d}}_{seg}(l) = \left[\hat{d}_{seg}\left(l\frac{R}{2}\right), \ldots, \hat{d}_{seg}\left(R-1+l\frac{R}{2}\right)\right]^T, \tag{70}$$

$$\hat{d}(t-d_t) = \sum_{l=0}^{L_f-1}\hat{d}_{seg}\left(t-l\frac{R}{2}\right)g_s\left(t-l\frac{R}{2}\right) \tag{71}$$

where $g_s(t)$ is a synthesis window, $d_t = d_{\text{AFC}} + d_{\text{NR}}$ is the total delay from both stages and d_{AFC} is the delay from the AFC stage. A complete description of the cascade M-channel PEM-AFC and rank-1 MWF algorithm is provided in Algorithm 4 and a block diagram is provided in Figure 1(c).

Simulation results

Scenario description

In order to assess the performance of the presented cascade algorithms, closed-loop simulations were performed using the following three scenarios.

- **Scenario 1** consists of a 4-microphone linear array and a loudspeaker in front of it which reproduces an amplified version of the desired speech source signal. The desired source is also in front of the microphone array but closer than the loudspeaker. Artificial impulse responses from the loudspeaker and the desired source to the microphones were generated using the randomized image method in [25], and the speech source signal was generated using a cascade of AR models. Results for different SNRs are shown.
- **Scenario 2** has the same set-up as Scenario 1, however the source signal is replaced by a speech signal [26] and the reverberation time is set to 0.14 s. Results for different SNRs are shown.
- **Scenario 3** consists of a 4-microphone array and a loudspeaker located diagonally from it, which reproduces an amplified version of the desired signal. The desired source is in front of the microphone array. Measured impulse responses [27] the from loudspeaker and the desired source to the microphones were used and the source signal was a speech signal [26]. Results for different

SNRs are shown. Although the reverberation time of these impulse responses is 0.5 s, they were truncated to 0.31 s which keeps most of the reverberant tail. The loudspeaker signal in all scenarios was obtained by using the desired signal estimate $\hat{d}(t)$, multiplied and delayed by the forward path gain and delay respectively. The window and impulse response length for each scenario are shown in Table 1. The forward path gain profile used for Scenario 1 is shown in Figure 2 with K_{MSG} defined in Section . Similar forward path gain profiles were used for Scenario 2 and Scenario 3, however the duration of the signals is different. The gain profile was chosen such that the *noise-only* and *speech-plus-noise* correlation matrices in the three algorithms could be updated while the system is stable, then the gain is gradually increased to test the proposed algorithms. The forward path delay in the simulations depends on the window size used for both the WOLA and OLS procedures. In all simulations, the forward path delay was set to $\frac{3R}{2}$. An R-samples long root-squared Hann window was used in the WOLA filterbank for the NR stage and an R-samples long rectangular window was used in the OLS filterbank for the AFC stage.

Feedback cancellation performance measures

Misadjustment (Mis)

The Mis measure is defined as the normalised distance in dB between the true and estimated feedback path in the time domain. Alternatively, due to Parseval's energy theorem, the Mis can be expressed in the frequency domain as [9]

$$\text{Mis}(l) = 20\log_{10}\left[\frac{\frac{1}{R}\sum_{\kappa=0}^{R-1}\left(f^{(r)}(\kappa) - \hat{f}^{(r)}(\kappa,l)\right)^2}{\frac{1}{R}\sum_{\kappa=0}^{R-1}\left(f^{(r)}(\kappa)\right)^2}\right] \quad \text{dB}. \tag{72}$$

Added stable gain (ASG)

The ASG measure is based on the so-called maximum stable gain (MSG) which is the maximum gain achievable in the system without it becoming unstable. In a single-channel scenario with a spectrally flat forward path, the MSG is given by [1]

$$\text{MSG}(l) = -20\log_{10}\left[\max_{\kappa\in\mathcal{P}^{(r)}(l)}\left|f^{(r)}(\kappa) - \hat{f}^{(r)}(\kappa,l)\right|\right] \quad \text{dB} \tag{73}$$

where $\mathcal{P}^{(r)}(l)$ is the set of frequencies that satisfy the phase condition of the Nyquist stability criterion [1] at the reference microphone. The ASG is then obtained as

$$\text{ASG}(l) = \text{MSG}(l) - K_{\text{MSG}} \quad \text{dB} \tag{74}$$

where K_{MSG} is the MSG of the system when no feedback canceller is included, i.e., $\hat{f}^{(m)}(\kappa,l) = 0\ \forall\kappa,l$, in (73).

When a NR stage is included in the closed-loop system the expression in (73) can be modified to account for the NR filters. For this, the MSG is defined at a reference microphone as

$$\text{MSG}(l) = -20\log_{10}\left[\max_{\kappa\in\mathcal{P}^{(r)}(l)}\left|f^{\star(r)}(\kappa,l) - \hat{f}^{(r)}(\kappa,l)\right|\right] \quad \text{dB} \tag{75}$$

where for an M-channel NR stage $f^{\star(r)}(\kappa, l)$ is defined as

$$f^{\star(r)}(\kappa, l) = \mathbf{e}_r^H \hat{\mathbf{W}}^H(\kappa, l)\mathbf{f}(\kappa, l), \tag{76}$$

and for an $M+1$-channel NR stage $f^{\star(r)}(\kappa, l)$ is

$$f^{\star(r)}(\kappa, l) = \mathbf{e}_{r+1}^H \hat{\mathbf{W}}^H(\kappa, l) \begin{bmatrix} 0 \\ \mathbf{f}(\kappa, l) \end{bmatrix}. \tag{77}$$

Then, the ASG can be computed as in (74), noting that K_{MSG} should be computed similarly to (75) with the initial value of $\hat{\mathbf{W}}$.

Signal distortion (SD)
The SD gives an indication of the distortion of the processed signal. Unweighted and weighted SD measures have been used in the literature [8, 28, 9, 29] for different speech enhancement algorithms. The frequency-weighted SD is defined as in [8]

$$\text{SD}(l) = \left(\int_{f_l}^{f_h} w_{\text{ERB}}(f) \left(10 \log_{10} \frac{\Phi_e(f, l)}{\Phi_r(f, l)} \right)^2 df. \right)^{1/2} \tag{78}$$

where $\Phi_e(f, l)$ is the PSD of the estimated signal, $\Phi_r(f, l)$ is the PSD of the reference signal, f is the frequency index in Hz and $w_{\text{ERB}}(f)$ is a weighting function which gives equal weight to each auditory critical band between $f_l = 300\,\text{Hz}$ and $f_h = 6400\,\text{Hz}$. For this metric, the estimated signal is $\hat{d}(t)$ and the reference signal is $H^{(r)}(q, t)s(t)$ (cfr. (5)). The measure is computed only during *"speech-plus-noise"* periods and the average over all frames is presented.

Perceptual Performance measures

For the perceptual assessment of the cascade algorithms presented in this paper, two metrics have been selected, namely, the PESQ and the STOI [30, 31, 9]. The PESQ measure is part of an International Telecommunications Union (ITU) Standard and widely used to objectively assess the perceptual quality of a speech signal. The STOI measure is a correlation-based speech intelligibility measure that works on the temporal envelopes of short speech frames. We used a MATLAB implementation of the STOI measure from [31]. These metrics were chosen based on the results presented in [9] where objective metrics were compared to subjective evaluation results for AFC algorithms.

Closed-loop simulations

Closed-loop simulation results are presented in this section. The algorithms are abbreviated as follows in the legends and tables descriptors. The cascade M-channel rank-1 MWF and PEM-AFC algorithm is abbreviated as **Rank-1 NR-AFC**, the cascade $(M+1)$-channel rank-2 MWF and PEM-AFC algorithm as **Rank-2 NR-AFC** and the cascade M-channel PEM-AFC and rank-1 MWF as **AFC-NR**. The three algorithms and the data for Scenario 1 are available in [32].

Figure 3 shows the ASG and Mis for three iSNRs for all algorithms using Scenario 1. In addition the STOI and SD scores for each algorithm are shown in Table 2. It is observed that both the Rank-2 NR-AFC and AFC-NR increase the ASG and the Mis is reduced. Furthermore, the STOI and SD scores outperform those of the Rank-1 NR-AFC for all iSNRs.

Figure 4 shows the ASG and Mis for all algorithms using Scenario 2. The STOI, PESQ-MOS and SD scores are shown in Table 3. It can be seen that both the Rank-2 NR-AFC and the AFC-NR outperform the Rank-1 NR-AFC in terms of ASG and Mis. Similarly to the results using Scenario 1, the STOI, PESQ-MOS and SD scores of both the Rank-2 NR-AFC and AFC-NR algorithms outperform those of the Rank-1 NR-AFC algorithm for all iSNRs.

Figure 5 shows the ASG and Mis for all algorithms using Scenario 3. It can be seen that the ASG is slightly increased for the Rank-2 NR-AFC and the AFC-NR algorithms when the iSNR is 5 dB and 0 dB. It is also observed that the Rank-2 NR-AFC and AFC-NR decrease the Mis, however not as much as in Scenario 1 and 2. The STOI and SD scores are presented in Table 4. Both the Rank-2 NR-AFC and AFC-NR outperform the Rank-1 NR-AFC algorithm for all iSNRs.

The observed high ASG values for the Rank-2 NR-AFC and AFC-NR algorithms in Scenario 1 and 2 can be explained by the inclusion of the NR filters in the ASG computation (cfr. (76)-(77)) which means that the MWF also influences the stability of the system. The fluctuating ASG values for the Rank-1 NR-AFC algorithm mean that the system stability is not guaranteed. This has been confirmed both by the perceptual performance measures scores in Table 2 and 3, and by the presence of howling in the resulting audio signals. Additionally, it should be noted that the SD scores in Table 2, 3 and 4 for all algorithms are considerably higher than those reported in the literature [9]. The reason for this is the sensitivity of this metric to the presence of noise in the microphone signals, which distorts the signal. In the literature, most of the considered SNRs are around 30 dB, which is considerably higher than the ones in this paper. In Scenario 3, the feedback path estimate is being under-modeled (cfr. Table 1) which explains the low ASG values for all the algorithms. The estimated feedback path has a smoother frequency response than the true feedback path which can cause a magnitude difference in the ASG computation, resulting in a slowing increasing ASG. Similarly to Scenario 1 and 2, the system is not stable when using the Rank-1 NR-AFC algorithm.

Conclusions

Three cascade multi-channel NR and AFC algorithms have been presented. Three different scenarios have been used to compare the performance of these algorithms in simulations. It is shown that both the cascade $(M+1)$-channel rank-2 MWF and PEM-AFC and the cascade M-channel PEM-AFC and rank-1 MWF algorithms outperform the cascade M-channel rank-1 MWF and PEM-AFC in terms of ASG and Mis. It is then shown in Section that both the cascade $(M+1)$-channel rank-2 MWF and PEM-AFC and the cascade M-channel PEM-AFC and rank-1 MWF are suitable to solve the combined AFC and NR problem in speech applications. It is also shown that by performing a rank-2 approximation of the speech correlation matrix the feedback path can be correctly estimated when an NR stage precedes the AFC stage.

Declarations

Acknowledgements
Not applicable.

Funding
This research work was carried out at the ESAT Laboratory of KU Leuven, in the frame of Research Council KU Leuven Project C3-19-00221 "Cooperative Signal Processing Solutions for IoT-based Multi-User Speech Communication Systems", Fonds de la Recherche Scientifique - FNRS and the Fonds Wetenschappelijk Onderzoek - Vlaanderen under EOS Project no 30452698 '(MUSE-WINET) MUlti-SErvice WIreless NETwork' and the European Research Council under the European Union's Horizon 2020 Research and Innovation Program / ERC Consolidator Grant: SONORA (no. 773268). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. The scientific responsibility is assumed by its authors.

Availability of data and materials
The algorithms are publicly available at [32].

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SR, TvW and MM jointly developed the idea of using a $(M+1)$-channel data model in the multichannel Wiener filter formulation for combined acoustic feedback cancellation and noise reduction. SR, TvW and MM jointly developed the research methodology to turn this concept into a usable and effective algorithm. SR, TvW and MM jointly designed and interpreted the computer simulations. SR implemented the computer simulations. All authors contributed in writing the manuscript and further read and approved the final manuscript.

Author details
Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium.

References

1. van Waterschoot, T., Moonen, M.: Fifty years of acoustic feedback control: State of the art and future challenges. Proc. IEEE **99**(2), 288–327 (2011)
2. Guo, M., Jensen, S.H., Jensen, J.: Evaluation of state-of-the-art acoustic feedback cancellation systems for hearing aids. J. Audio Eng. Soc. **61**(3), 125–137 (2013)
3. Guo, M., Jensen, S.H., Jensen, J.: Novel acoustic feedback cancellation approaches in hearing aid applications using probe noise and probe noise enhancement. IEEE Trans. Audio Speech Lang. Process. **20**(9), 2549–2563 (2012). doi:10.1109/TASL.2012.2206025
4. Guo, M., Jensen, S.H., Jensen, J., Grant, S.L.: On the use of a phase modulation method for decorrelation in acoustic feedback cancellation. In: Proc. 20th European Signal Process. Conf. (EUSIPCO '12) (2012)
5. Schepker, H., Nordholm, S.E., Tran, L.T.T., Doclo, S.: Null-steering beamformer-based feedback cancellation for multi-microphone hearing aids with incoming signal preservation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(4), 679–691 (2019). doi:10.1109/TASLP.2019.2892234
6. Strasser, F., Puder, H.: Adaptive feedback cancellation for realistic hearing aid applications. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2322–2333 (2015). doi:10.1109/TASLP.2015.2479038
7. Spriet, A., Moonen, M., Proudler, I.: Feedback cancellation in hearing aids: an unbiased modelling approach. In: Proc. 11th European Signal Process. Conf. (EUSIPCO '02), pp. 1–4 (2002)
8. Spriet, A., Moonen, M., Wouters, J.: Evaluation of feedback reduction techniques in hearing aids based on physical performance measures. J. Acoust. Soc. Amer. **128**(3), 1245–1261 (2010)
9. Bernardi, G., van Waterschoot, T., Wouters, J., Moonen, M.: Subjective and objective sound-quality evaluation of adaptive feedback cancellation algorithms. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(5), 1010–1024 (2018)
10. Benesty, J., Chen, J., Huang, Y.A., Doclo, S.: Study of the wiener filter for noise reduction. In: Speech Enhancement, pp. 9–41. Springer, Berlin Heidelberg (2005)
11. Benesty, J., Jensen, J.R., Christensen, M.G., Chen, J.: Speech Enhancement: A Signal Subspace Perspective. Elsevier, Oxford (2014)
12. Serizel, R., Moonen, M., Van Dijk, B., Wouters, J.: Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(4), 785–799 (2014)
13. Wang, D., Chen, J.: Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018). doi:10.1109/TASLP.2018.2842159
14. Spriet, A., Rombouts, G., Moonen, M., Wouters, J.: Combined feedback and noise suppression in hearing aids. IEEE Trans. Audio Speech Lang. Process. **15**(6), 1777–1790 (2007). doi:10.1109/TASL.2007.896670
15. Rombouts, G., Spriet, A., Moonen, M.: Generalized sidelobe canceller based combined acoustic feedback- and noise cancellation. Signal Processing **88**(3), 571–581 (2008). doi:10.1016/j.sigpro.2007.08.018
16. Bastari, A., Squartini, S., Piazza, F.: Joint acoustic feedback cancellation and noise reduction within the prediction error method framework. In: 2008 Hands-Free Speech Communication and Microphone Arrays, pp. 228–231 (2008). doi:10.1109/HSCMA.2008.4538728
17. Rombouts, G., van Waterschoot, T., Struyve, K., Moonen, M.: Acoustic feedback cancellation for long acoustic paths using a nonstationary source model. IEEE Trans. Signal Process. **54**(9), 3426–3434 (2006)

18. Schepker, H., Doclo, S.: Active feedback suppression for hearing devices exploiting multiple loudspeakers. In: Proc. 2019 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '19), pp. 60–64 (2019). doi:10.1109/WASPAA.2019.8937187
19. Vashkevich, M., Azarov, E., Petrovsky, N., Likhachov, D., Petrovsky, A.: Real-time implementation of hearing aid with combined noise and acoustic feedback reduction based on smartphone. In: Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '17), pp. 6570–6571 (2017). doi:10.1109/ICASSP.2017.8005301
20. Ruiz, S., van Waterschoot, T., Moonen, M.: Cascade multi-channel noise reduction and acoustic feedback cancellation. In: Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '22), pp. 676–680 (2022). doi:10.1109/ICASSP43922.2022.9747291
21. Bernardi, G., van Waterschoot, T., Wouters, J., Moonen, M.: An all-frequency-domain adaptive filter with PEM-based decorrelation for acoustic feedback control. In: Proc. 2015 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '15), pp. 1–5 (2015)
22. Crochiere, R.: A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE Trans. Acoust., Speech, Signal Process. **28**(1), 99–102 (1980)
23. Jabloun, F., Champagne, B.: Signal subspace techniques for speech enhancement. In: Speech Enhancement, pp. 135–159. Springer, Berlin Heidelberg (2005)
24. Bertrand, A., Moonen, M.: Robust distributed noise reduction in hearing aids with external acoustic sensor nodes. EURASIP J. Adv. Signal Process **2009**, 1–14 (2009)
25. De Sena, E., Antonello, N., Moonen, M., van Waterschoot, T.: On the modeling of rectangular geometries in room acoustic simulations. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(4), 774–786 (2015)
26. Bang, Olufsen: Music for Archimedes. Compact Disc B&O (1992)
27. Dietzen, T., Ali, R., Taseska, M., van Waterschoot, T.: MYRiAD: A Multi-Array Room Acoustic Database. ESAT-STADIUS Tech. Rep. TR 22-118, KU Leuven, Belgium (submitted for publication) (2022)
28. Gannot, S., Cohen, I.: Speech enhancement based on the general transfer function GSC and postfiltering. IEEE Trans. Speech Audio Process. **12**(6), 561–571 (2004)
29. Aichner, R.: Acoustic blind source separation in reverberant and noisy environments. PhD thesis, Friedrich-Alexander-Universitat Erlangen-Nurnberg (2007)
30. P.862, I.-T.R.: Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union, Geneva, Switzerland (2001)
31. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '10), pp. 4214–4217 (2010)
32. Ruiz, S.: AFC-NR. `https://github.com/rogaits/AFC-NR`

Algorithm 1: Single-channel PEM-based AFC [21]

Initialize estimates of feedback path, desired signal, etc...;

for $l = 0, 1, 2, \ldots$ **do**

Define loudspeaker signal frame $\mathbf{u}(l) = \left[u\left(l\frac{R}{2} - R + 1\right), \ldots, u\left(l\frac{R}{2}\right)\right]$;

Compute $\mathbf{U}(l) = \operatorname{diag}\{\boldsymbol{\mathcal{F}}_R \mathbf{u}(l)\}$; % where $\boldsymbol{\mathcal{F}}_R$ represents the DFT matrix.

Estimate the feedback component

$$\hat{\mathbf{x}}_r(l) = \mathbf{C}^T \boldsymbol{\mathcal{F}}_R^{-1} \mathbf{U}(l) \left[\hat{f}^{(r)}(0,l) \quad \cdots \quad \hat{f}^{(r)}(R-1,l)\right]^T; \text{ \% With}$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{0}_{\frac{R}{2}\times\frac{R}{2}} & \mathbf{I}_{\frac{R}{2}\times\frac{R}{2}} \end{bmatrix}^T$$

Estimate the feedback-compensated signal $\hat{\mathbf{x}}_f^{(r)}(l) = \mathbf{x}^{(r)}(l) - \hat{\mathbf{x}}^{(r)}(l)$;

Estimate AR model coefficients $\hat{\mathbf{a}}(l) = \text{Levinson} - \text{Durbin}\left(\left[\hat{\mathbf{x}}_f^{(r)T}(l) \quad \hat{\mathbf{x}}_f^{(r)T}(l-1)\right]^T\right)$;

Compute the DFT of the AR model $\hat{\mathbf{A}}(l) = \operatorname{diag}\left\{\boldsymbol{\mathcal{F}}_R \left[\hat{\mathbf{a}}^T(l) \quad \mathbf{0}_{1\times(R-n_A-1)}\right]^T\right\}$; % where n_A is the AR model order.

Compute the time-domain prediction error

$$\boldsymbol{\varepsilon}^{(r)}(l) = \mathbf{C}_e^T \boldsymbol{\mathcal{F}}_R^{-1} \hat{\mathbf{A}}(l) \left(\begin{bmatrix} \hat{x}^{(r)}(0,l) \\ \vdots \\ \hat{x}^{(r)}(R-1,l) \end{bmatrix} - \mathbf{U}(l) \begin{bmatrix} \hat{f}^{(r)}(0,l) \\ \vdots \\ \hat{f}^{(r)}(R-1,l) \end{bmatrix} \right); \text{ \% with}$$

$$\mathbf{C}_e = \begin{bmatrix} \mathbf{0}_{(\frac{R}{2}-n_A)\times(\frac{R}{2}+n_A)} & \mathbf{I}_{(\frac{R}{2}-n_A)} \end{bmatrix}^T.$$

Compute the frequency-domain prediction error

$$\left[\varepsilon^{(r)}(0,l) \quad \cdots \quad \varepsilon^{(r)}(R-1,l)\right]^T = \boldsymbol{\mathcal{F}}_R \mathbf{C}_e \boldsymbol{\varepsilon}^{(r)}(l);$$

Compute the frequency-domain step size $\mu(\kappa,l) = \mu_{\text{fix}}\left[\alpha + |\hat{a}(\kappa,l)u(\kappa,l)|^2\right]^{-1}$; % where α is a regularization term and μ_{fix} is a real, constant frequency-independent value.

$$\begin{bmatrix} \Delta\hat{f}^{(r)}(0,l) \\ \vdots \\ \Delta\hat{f}^{(r)}(R-1,l) \end{bmatrix} = \operatorname{diag}\left\{ \begin{bmatrix} \mu(0,l) \\ \vdots \\ \mu(R-1,l) \end{bmatrix} \right\} \mathbf{U}^H(l) \hat{\mathbf{A}}^H(l) \begin{bmatrix} \varepsilon^{(r)}(0,l) \\ \vdots \\ \varepsilon^{(r)}(R-1,l) \end{bmatrix};$$

Update feedback path coefficients $\begin{bmatrix} \hat{f}^{(r)}(0,l+1) \\ \vdots \\ \hat{f}^{(r)}(R-1,l+1) \end{bmatrix} =$

$$\begin{bmatrix} \hat{f}^{(r)}(0,l) \\ \vdots \\ \hat{f}^{(r)}(R-1,l) \end{bmatrix} + \boldsymbol{\mathcal{F}}_R \left[\mathbf{I}_{R\times R} - \mathbf{C}\mathbf{C}^T\right] \boldsymbol{\mathcal{F}}_R^{-1} \begin{bmatrix} \Delta\hat{f}^{(r)}(0,l) \\ \vdots \\ \Delta\hat{f}^{(r)}(R-1,l) \end{bmatrix};$$

Algorithm 2: Cascade M-channel rank-1 MWF and PEM-AFC

Initialize estimates of feedback path, desired signal, β, etc...;

Transform $u(t)$ and $x^{(m)}(t)$ $\forall m$ to STFT domain;

for $l = 0, 1, 2, \ldots$ **do**

for $\kappa = 0, 1, \ldots, R-1$ **do**

Update $\hat{\mathbf{R}}_{\mathbf{xx}}(\kappa,l)$ or $\hat{\mathbf{R}}_{\mathbf{xx}|\mathbf{nn}}(\kappa,l)$ similar to (18) based on VAD;

Update $\hat{\mathbf{w}}(\kappa,l)$ using (24) by means of the GEVD of $\{\hat{\mathbf{R}}_{\mathbf{xx}}(\kappa,l), \hat{\mathbf{R}}_{\mathbf{xx}|\mathbf{nn}}(\kappa,l)\}$;

Compute $\hat{x}_s^{(r)}(\kappa,l)$ based on (25);

Obtain the time-domain signal $\hat{x}_s^{(r)}(t)$ based on (26)-(28);

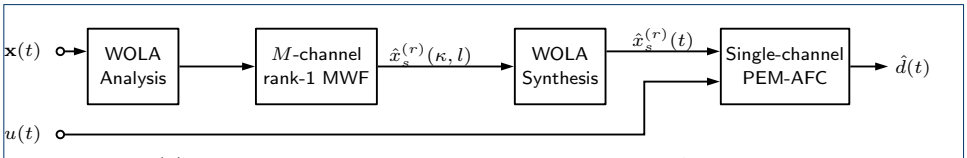
Perform Algorithm 1 using $u(t)$ and $\hat{x}_s^{(r)}(t)$ as input signals.

Algorithm 3: Cascade $(M+1)$-channel rank-2 MWF and PEM-AFC

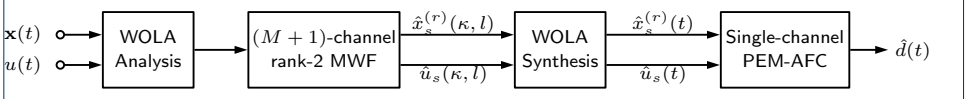
Initialize estimates of feedback path, desired signal, β, etc...;
Transform $u(t)$ and $x^{(m)}(t)\ \forall m$ to STFT domain;
for $l = 0, 1, 2, \ldots$ **do**
  for $\kappa = 0, 1, \ldots, R-1$ **do**
    Construct the signal vector $\mathbf{y}(\kappa, l)$ based on $\mathbf{x}(\kappa, l)$ and $u(\kappa, l)$;
    Update $\hat{\mathbf{R}}_{\mathbf{yy}}(\kappa, l)$ or $\hat{\mathbf{R}}_{\mathbf{yy|nn}}(\kappa, l)$ similar to (35) based on VAD;
    Update $\hat{\mathbf{W}}(\kappa, l)$ using (41) by means of the GEVD of $\{\hat{\mathbf{R}}_{\mathbf{yy}}(\kappa, l), \hat{\mathbf{R}}_{\mathbf{yy|nn}}(\kappa, l)\}$;
    Compute $\hat{u}_s(\kappa, l)$ and $\hat{x}_s^{(r)}(\kappa, l)$ based on (42) and (43);
  Obtain the time-domain signals $\hat{u}_s(t)$ and $\hat{x}_s^{(r)}(t)$ based on (44)-(49);
  Perform Algorithm 1 using $\hat{u}_s(t)$ and $\hat{x}_s^{(r)}(t)$ as input signals.

Algorithm 4: Cascade M-channel PEM-AFC and rank-1 MWF

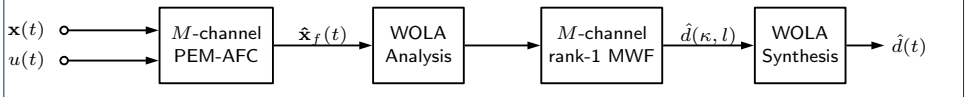
Initialize estimates of feedback path, desired signal, β, etc...;
for $l = 0, 1, 2, \ldots$ **do**
  for $m = 1, \ldots, M$ **do**
    Perform Algorithm 1 using $\hat{\mathbf{u}}(l)$ and $\hat{\mathbf{x}}^{(m)}(l)$ as input signals.
  Transform $u(t)$ and $\hat{x}_f^{(m)}(t)\ \forall m$ to STFT domain based on (56)
  for $\kappa = 0, \ldots, R-1$ **do**
    Construct the signal vector $\mathbf{x}_f(\kappa, l)$;
    Update $\hat{\mathbf{R}}_{\mathbf{x_f x_f}}(\kappa, l)$ or $\hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}(\kappa, l)$ similar to (60) based on VAD;
    Update $\hat{\mathbf{w}}(\kappa, l)$ using (68) by means of the GEVD of $\{\hat{\mathbf{R}}_{\mathbf{x_f x_f}}(\kappa, l), \hat{\mathbf{R}}_{\mathbf{x_f x_f|nn}}(\kappa, l)\}$;
    Compute desired signal $\hat{d}(\kappa, l) = \hat{\mathbf{w}}^H(\kappa, l)\mathbf{x}_f(\kappa, l)$;



(a) Cascade M-channel rank-1 MWF and PEM-AFC algorithm.



(b) Cascade $M+1$-channel rank-2 MWF and PEM-AFC algorithm.



(c) Cascade M-channel PEM-AFC and rank-1 MWF algorithm.

Figure 1: Block diagrams for cascade algorithms.

Table 1: Scenarios and simulation parameters

Parameter	Scenario 1	Scenario 2	Scenario 3
Impulse response length (samples)	1024	2048	5000
Window length (samples)	1024	2048	5000
Estimated Impulse response length (samples)	512	1024	2500
T_{60} (s)	0	0.14	0.5
Sampling rate (Hz)	16000	16000	16000
Source signal	Cascaded AR models	Speech	Speech

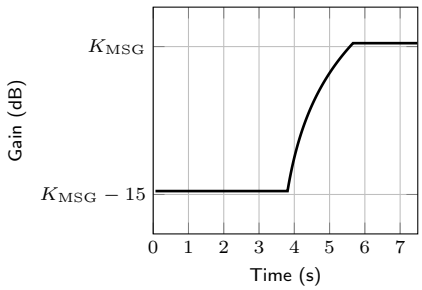


Figure 2: Forward path gain profile for Scenario 1.

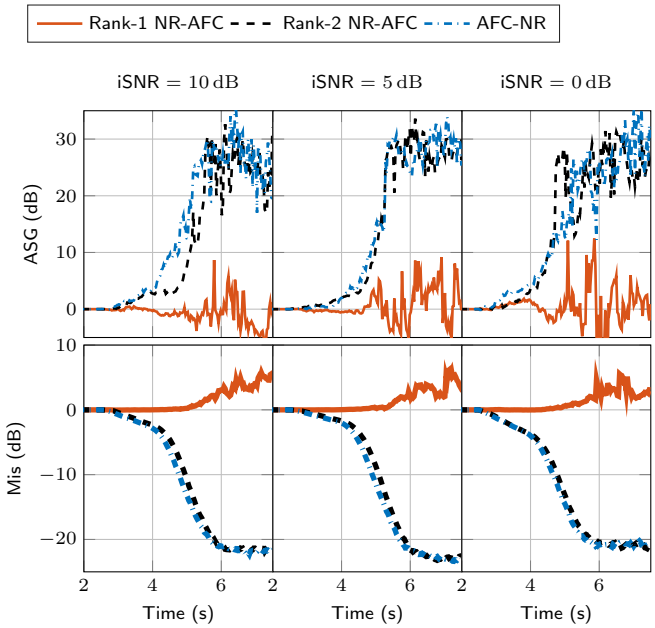


Figure 3: ASG and Mis for the three cascade algorithms in Scenario 1.

Table 2: STOI and SD for the three cascade algorithms using Scenario 1.

SNR	Algorithm	STOI	mean(SD)	max(SD)
10 dB	Rank-1 NR-AFC	0.21	122.60	162.81
	Rank-2 NR-AFC	0.71	4.63	6.69
	AFC-NR	0.70	4.73	10.36
5 dB	Rank-1 NR-AFC	0.17	115.48	153.43
	Rank-2 NR-AFC	0.53	6.12	8.28
	AFC-NR	0.55	6.11	8.17
0 dB	Rank-1 NR-AFC	0.11	130.88	173.73
	Rank-2 NR-AFC	0.40	8.19	9.93
	AFC-NR	0.41	8.14	10.16

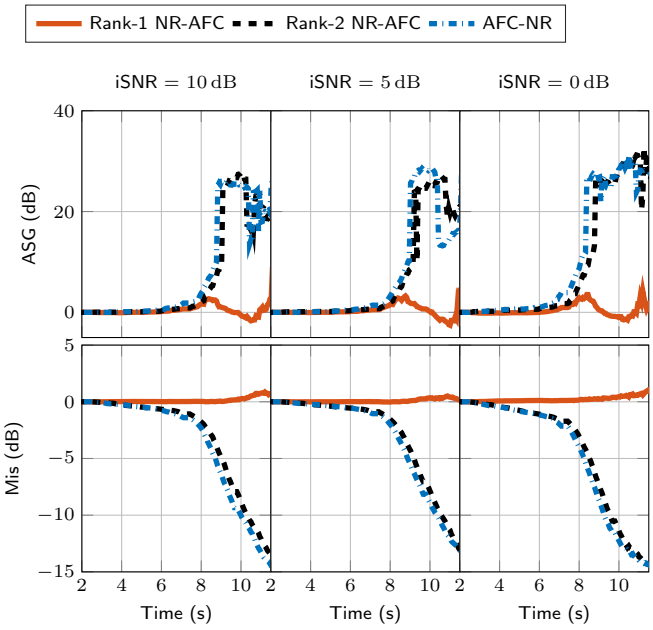


Figure 4: ASG and Mis for the three cascade algorithms in Scenario 2.

Table 3: STOI, SD and PESQ for the three cascade algorithms using Scenario 2.

SNR	Algorithm	STOI	mean(SD)	max(SD)	PESQ MOS
10 dB	Rank-1 NR-AFC	0.68	35.02	69.53	1.04
	Rank-2 NR-AFC	0.78	22.22	45.35	1.43
	AFC-NR	0.77	22.38	46.28	1.43
5 dB	Rank-1 NR-AFC	0.59	38.52	81.36	1.04
	Rank-2 NR-AFC	0.70	26.27	50.09	1.26
	AFC-NR	0.70	26.37	49.82	1.26
0 dB	Rank-1 NR-AFC	0.56	41.83	70.78	1.03
	Rank-2 NR-AFC	0.65	30.00	53.96	1.18
	AFC-NR	0.65	30.16	53.58	1.18

Table 4: STOI and SD for the three cascade algorithms using Scenario 3.

SNR	Algorithm	STOI	mean(SD)	max(SD)
10 dB	Rank-1 NR-AFC	0.56	24.59	56.40
	Rank-2 NR-AFC	0.63	22.58	52.01
	AFC-NR	0.59	22.29	51.20
5 dB	Rank-1 NR-AFC	0.48	28.35	60.03
	Rank-2 NR-AFC	0.55	26.50	55.54
	AFC-NR	0.53	26.09	54.60
0 dB	Rank-1 NR-AFC	0.41	32.64	64.10
	Rank-2 NR-AFC	0.49	30.89	60.76
	AFC-NR	0.47	30.39	61.29

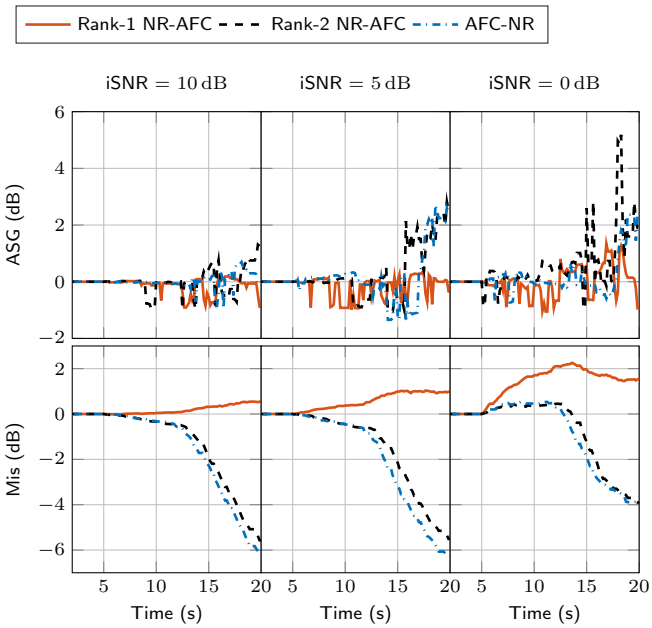


Figure 5: ASG and Mis for the three cascade algorithms in Scenario 3.