CASCADE MULTI-CHANNEL NOISE REDUCTION AND ACOUSTIC FEEDBACK CANCELLATION

Santiago Ruiz, Toon van Waterschoot and Marc Moonen

KU Leuven
Dept. of Electrical Engineering, ESAT-STADIUS
Leuven, Belgium.

ABSTRACT

Acoustic feedback and noise are common problems that corrupt microphone signals and affect the performance of speech and audio signal processing applications and devices. In this paper, a cascade noise reduction (NR) and acoustic feedback cancellation (AFC) algorithm is presented for speech applications where a multi-channel Wiener filter (MWF) based NR is applied first followed by a single-channel prediction-error method (PEM) based adaptive feedback cancellation stage. It is shown that by using a rank-2 estimate of the speech correlation matrix in the NR stage it is possible to obtain a good feedback path estimate for the reference microphone in the AFC stage. Closed-loop simulations with M microphones and 1 loudspeaker are presented using both an M-channel rank-1 and an $(M+1)$-channel rank-2 MWF and it is shown that for the considered input signal-to-noise ratios the proposed algorithm increases the added stable gain (ASG) of the system.

***Index Terms*—** combined acoustic feedback cancellation and noise reduction, multichannel Wiener filter, prediction-error method based adaptive filtering with row operations.

1. 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 coupling between the microphone and loudspeaker may give rise to instabilities in the system, which translates into signal degradation and, in the worst case, acoustic howling. Four 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 decorrelation of the microphone and loudspeaker signals to obtain an unbiased feedback path estimate [1, 2]. Different approaches for decorrelation have been proposed such as probe-noise injection [3], time-varying or nonlinear processes in the forward path [4], and prewhitening [5]. The latter approach has been shown to provide limited perceptual distortion [6, 7]. In the literature, many different solutions for AFC can be found using different decorrelation procedures [1, 3, 4, 8, 9, 10]. Similarly for noise reduction (NR), a wide range of solutions can be found in the literature, where state-of-the-art algorithms such as the multi-channel Wiener filter (MWF) are often used [11, 12, 13], as well as more recently deep learning-based methods [14]. However, only few solutions to combined AFC and NR have been reported [15].

In this paper, a cascade NR and AFC algorithm is presented for speech applications, where a NR stage is applied first followed by a single-channel AFC stage. The novelty of the paper is to consider an $(M+1)$-channel data model (with M being the number of microphones) in the MWF formulation with two different signals, i.e., the speech component in the reference microphone and in the loudspeaker signals, both defined by the desired speech signal but not equal to each other. This allows to estimate a rank-2 $(M+1)$-MWF in the NR stage [13] which is then used to obtain estimates of the speech component in the reference microphone and in the loudspeaker signal. These estimates are later used in a single-channel prediction-error method (PEM) based AFC [5, 16].

The paper is organized as follows: Section 2 describes the signal model, where the underlying assumptions for the rank-2 estimate of the speech correlation matrix are detailed. Sections 3 and 4 describe the NR and AFC stages of the baseline and proposed algorithms, respectively. Sections 5 and 6 present the simulation results and conclusions of the paper, respectively.

2. SIGNAL MODEL

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

$$x_m(t) = H_m(q,t)s(t) + F_m(q,t)u(t) + n_{x,m}(t) \qquad (1)$$

where $H_m(q,t)$ and $F_m(q,t)$ are the transfer functions from the desired speech source and from the loudspeaker to the m^{th} microphone, respectively. The latter is also known as the feedback path. The noise signal in the m^{th} microphone is denoted by $n_{x,m}(t)$. The discrete time index is represented by t and q^{-1} is the delay operator, i.e., $q^{-k}u(t) = u(t-k)$. The loudspeaker signal can be expressed as

$$u(t) = \sum_{m=1}^{M} G_m(q,t)x_m(t) \qquad (2)$$

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", VLAIO O&O Project nr. HBC.2020.2197 'SPIC: Signal Processing and Integrated Circuits for Communications', Fonds de la Recherche Scientifique - FNRS and Fonds voor Wetenschappelijk Onderzoek - Vlaanderen 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.

where $G_m(q,t)$ is the forward path transfer function for the m^{th} microphone. 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 desired speech signal can be modeled as

$$s(t) = H_s(q,t)e(t) \tag{3}$$

where $H_s(q,t)$ is defined by an autoregressive (AR) process excited by the white noise signal $e(t)$. 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, i.e.,

$$d(t) = H_r(q,t)s(t) \tag{4}$$

where $H_r(q,t)$ is the transfer function from the desired speech source to the reference microphone.

In the short-time Fourier transform (STFT) domain, an $N \times 1$ multi-channel signal vector ($N = M+1$), consisting of microphone and loudspeaker signals, can be expressed as

$$\mathbf{y}(\kappa,l) = \begin{bmatrix} 0 \\ \mathbf{h}_s(\kappa,l) \end{bmatrix} s(\kappa,l) + \begin{bmatrix} 1 \\ \mathbf{h}_f(\kappa,l) \end{bmatrix} u_s(\kappa,l) + \mathbf{n}(\kappa,l) \tag{5}$$

$$= \begin{bmatrix} u(\kappa,l) \\ \mathbf{x}(\kappa,l) \end{bmatrix} \tag{6}$$

where $s(\kappa,l)$, $u_s(\kappa,l)$, $u(\kappa,l)$ and $\mathbf{n}(\kappa,l)$ are the STFT representations of the speech signal, the desired speech component in the loudspeaker signal, the loudspeaker signal and the noise in the microphone and loudspeaker signals, respectively. It is noted that $\mathbf{n}(\kappa,l)$ includes the noise component in the loudspeaker signal, as well as, its coupling into the microphones, which is added to the direct noise components in the microphones. The steering vectors from the desired speech source and loudspeaker to the microphones are respectively denoted by $\mathbf{h}_s(\kappa,l)$ and $\mathbf{h}_f(\kappa,l)$. The time-frame and frequency-bin indices are l and κ, respectively (for brevity l and κ will be mostly omitted in the following). The speech correlation matrix is defined as follows

$$\bar{\mathbf{R}}_{\boldsymbol{ss}} = \begin{bmatrix} 1 & 0 \\ \mathbf{h}_f & \mathbf{h}_s \end{bmatrix} \begin{bmatrix} \Phi_{uu} & \Phi_{us} \\ \Phi_{su} & \Phi_{ss} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{h}_f^H \\ 0 & \mathbf{h}_s^H \end{bmatrix}. \tag{7}$$

where $\Phi_{ss} = E\{s^*s\}$,$\Phi_{su} = E\{s^*u_s\}$, $\Phi_{us} = E\{u_s^*s\}$, $\Phi_{uu} = E\{u_s^*u_s\}$, $E\{\cdot\}$ denotes statistical expectation, and $(\cdot)^*$ and $(\cdot)^H$ are the conjugate and conjugate transpose operators, respectively. Performing an LDL factorisation on the matrix with the Φ's in (7), $\bar{\mathbf{R}}_{\boldsymbol{ss}}$ can alternatively be expressed as

$$\bar{\mathbf{R}}_{\boldsymbol{ss}} = \begin{bmatrix} 1 & 0 \\ \mathbf{h}_f + \epsilon\mathbf{h}_s & \mathbf{h}_s \end{bmatrix} \begin{bmatrix} \Phi_{uu} & 0 \\ 0 & \Gamma \end{bmatrix} \begin{bmatrix} 1 & \mathbf{h}_f^H + \epsilon^*\mathbf{h}_s^H \\ 0 & \mathbf{h}_s^H \end{bmatrix} \tag{8}$$

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 $\bar{\mathbf{R}}_{\boldsymbol{ss}}$ in (8) alone, $\mathbf{h}_f$ and $\mathbf{h}_s$ cannot be uniquely defined whenever there is a non-zero correlation between s and u_s.

A cascade NR and AFC is presented here by performing a multi-channel NR stage first followed by a single-channel AFC stage. MWF-based NR is used to estimate the contribution of $s(\kappa,l)$ and $u_s(\kappa,l)$ in a reference microphone as well as in the loudspeaker, and then a single-channel AFC is performed on the resulting signals. Although in this type of cascade usually the estimation of the feedback path is 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 indeed be correctly estimated.

3. CASCADE M-CHANNEL RANK-1 MWF AND PEM-AFC

This section describes the baseline cascade algorithm. Here, an M-channel rank-1 MWF (i.e. using microphone signals only) is applied first, followed by a single-channel AFC stage.

3.1. 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, 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

$$\bar{\mathbf{R}}_{\mathbf{xx}} = E\{\mathbf{x}\mathbf{x}^H\} = \bar{\mathbf{R}}_{\boldsymbol{ss}} + \bar{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}} \tag{9}$$

with $\bar{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}} = E\{\mathbf{n_x}\mathbf{n_x}^H\}$ the $M \times M$ microphone-only noise correlation matrix. The minimization of the mean squared error (MSE) between the desired signal and the filtered microphone signals defines an optimal filter

$$\bar{\mathbf{w}} = \arg\min_{\mathbf{w}} \; E\left\{\left|d - \mathbf{w}^H\mathbf{y}\right|^2\right\}. \tag{10}$$

with $d = x_{rs}$ representing the total contribution of s together with u_s in the reference microphone signal. The estimate $\hat{x}_{rs}$ is obtained as

$$\hat{x}_{rs} = \bar{\mathbf{w}}^H\mathbf{x}. \tag{11}$$

The solution to (10) is the MWF [13, 11], given by

$$\bar{\mathbf{w}} = \bar{\mathbf{R}}_{\mathbf{xx}}^{-1}\bar{\mathbf{R}}_{\boldsymbol{ss}}\mathbf{e}_1 \tag{12}$$

where $\mathbf{e}_1$ selects the first column of $\bar{\mathbf{R}}_{\mathbf{xx}}^{-1}\bar{\mathbf{R}}_{\boldsymbol{ss}}$ which corresponds to the reference microphone. The final expression in (12) is obtained based on the assumption that s and $\mathbf{n}_x$ are uncorrelated.

In practice, by using a voice activity detector (VAD), $\bar{\mathbf{R}}_{\mathbf{xx}}$ and $\bar{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}$ 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.,

$$\begin{aligned} &\text{if VAD}(l) = 1 : \hat{\mathbf{R}}_{\mathbf{xx}}(l) = \beta\hat{\mathbf{R}}_{\mathbf{xx}}(l-1) + (1-\beta)\mathbf{x}(l)\mathbf{x}^H(l) \\ &\text{if VAD}(l) = 0 : \hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}(l) = \beta\hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}(l-1) + (1-\beta)\mathbf{x}(l)\mathbf{x}^H(l) \end{aligned} \tag{13}$$

where $\hat{\mathbf{R}}_{\mathbf{xx}}(l)$ and $\hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}(l)$ represent estimates of $\bar{\mathbf{R}}_{\mathbf{xx}}$ and $\bar{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}$ at frame l, 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 $\bar{\mathbf{R}}_{\boldsymbol{ss}}$ [13],

$$\hat{\mathbf{R}}_{\boldsymbol{ss}} = \underset{\substack{\text{rank}(\bar{\mathbf{R}}_{\boldsymbol{ss}})=1 \\ \bar{\mathbf{R}}_{\boldsymbol{ss}} \succeq 0}}{\arg\min} \left|\left|\hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}^{-1/2}\left(\hat{\mathbf{R}}_{\mathbf{xx}} - \hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}} - \mathbf{R}_{\boldsymbol{ss}}\right)\hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}^{-H/2}\right|\right|_F^2 \tag{14}$$

where $||\cdot||_F$ denotes the Frobenius norm. Spatial pre-whitening is applied by pre- and post-multiplying by $\hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}^{-1/2}$ and $\hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}^{-H/2}$, respectively. The solution to (14) is based on a generalized eigenvalue decomposition (GEVD) of the $(M \times M)$ matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{xx}}, \hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}}\}$ [13, 17]

$$\hat{\mathbf{R}}_{\mathbf{xx}} = \hat{\mathbf{Q}}\hat{\boldsymbol{\Sigma}}_{\mathbf{xx}}\hat{\mathbf{Q}}^H \tag{15}$$

$$\hat{\mathbf{R}}_{\mathbf{n_x}\mathbf{n_x}} = \hat{\mathbf{Q}}\hat{\boldsymbol{\Sigma}}_{\mathbf{n_x}\mathbf{n_x}}\hat{\mathbf{Q}}^H \tag{16}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{n_x n_x}}$ are diagonal matrices and $\hat{\mathbf{Q}}$ is an invertible matrix. The rank-1 speech correlation matrix estimate $\hat{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}}$ is then [13]

$$\hat{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}} = \hat{\mathbf{Q}}\text{diag}\{\hat{\sigma}_{x_1} - \hat{\sigma}_{n_{x,1}}, 0, \ldots, 0\}\hat{\mathbf{Q}}^H \tag{17}$$

where $\hat{\sigma}_{x_i}$ and $\hat{\sigma}_{n_{x,i}}$ are the ith diagonal element of $\hat{\mathbf{\Sigma}}_{\mathbf{xx}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{n_x n_x}}$, respectively, corresponding to the ith largest ratio $\hat{\sigma}_{x_i}/\hat{\sigma}_{n_{x,i}}$. Using (17) and $\hat{\mathbf{R}}_{\mathbf{xx}}$ (cfr. (15)) in (12), the MWF estimate $\hat{\mathbf{w}}$ can be expressed as

$$\hat{\mathbf{w}} = \hat{\mathbf{Q}}^{-H}\text{diag}\left\{1 - \frac{\hat{\sigma}_{n_{x,1}}}{\hat{\sigma}_{x_1}}, 0, \ldots, 0\right\}\hat{\mathbf{Q}}^H\mathbf{e}_1. \tag{18}$$

The estimate of the total contribution of s and u_s in the reference microphone signal, $\hat{x}_{rs}$, is obtained as in (11) with $\hat{\mathbf{w}}$ replacing $\bar{\mathbf{w}}$

$$\hat{x}_{rs} = \hat{\mathbf{w}}^H\mathbf{x}. \tag{19}$$

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

$$\hat{\mathbf{x}}_{rs,seg}(l) = \boldsymbol{\mathcal{F}}_R^{-1}\hat{\mathbf{x}}_{rs}^{\kappa}(l) \tag{20}$$

$$x_{rs}(t - d_{\text{NR}}) = \sum_{l=0}^{L_f-1}\hat{x}_{rs,seg}\left(t - l\frac{R}{2}\right)g_s\left(t - l\frac{R}{2}\right) \tag{21}$$

where $\boldsymbol{\mathcal{F}}_R$ is the discrete Fourier transform (DFT) matrix of size R, g_s is a synthesis window, d_{NR} is the delay from the NR stage and the superscript $(\cdot)^{\kappa}$ is used to denote a frequency-domain vector at time frame l as

$$\mathbf{x}_{rs}^{\kappa}(l) = \begin{bmatrix} x_{rs}(0,l) & \cdots & x_{rs}(R,l)\end{bmatrix}^T. \tag{22}$$

3.2. AFC stage

In the AFC stage a single-channel so-called PEM-based adaptive filtering with row operations (PEM-AFROW) algorithm is used. This algorithm was initially developed in [5] and it provides estimates of both the feedback path and the desired speech signal model in (1) and (3), respectively. The implemented algorithm is the frequency-domain version presented in [16] (the reader is referred to [16] for a detailed explanation of the AFC algorithm). The input signals of the AFC algorithm are the noisy loudspeaker signal and the estimate of the total contribution of s and u_s in the reference microphone, u and x_{rs}, respectively.

4. CASCADE (M+1)-CHANNEL RANK-2 MWF AND PEM-AFC

This section describes the proposed cascade algorithm. Here a (M+1)-channel rank-2 MWF is applied first, followed by a single-channel AFC stage.

4.1. 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 (6) can be expressed as

$$\bar{\mathbf{R}}_{\mathbf{yy}} = E\{\mathbf{y}\mathbf{y}^H\} = \bar{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}} + \bar{\mathbf{R}}_{\mathbf{nn}} \tag{23}$$

with $\bar{\mathbf{R}}_{\mathbf{nn}} = E\{\mathbf{n}\mathbf{n}^H\}$ the $N \times N$ noise correlation matrix. The minimization of the mean squared error (MSE) between the desired signals and the filtered microphone and loudspeaker signals defines an optimal filter

$$\bar{\mathbf{W}} = \arg\min_{\mathbf{W}}\ E\left\{\left|\mathbf{d} - \mathbf{W}^H\mathbf{y}\right|^2\right\}. \tag{24}$$

with $\mathbf{d} = \begin{bmatrix} u_s^* & y_{rs}^*\end{bmatrix}^H$, where y_{rs} represents the total contribution of s together with u_s in the reference microphone signal. The estimates $\hat{u}_s$ and $\hat{y}_{rs}$ are obtained as

$$\hat{u}_s = \left(\bar{\mathbf{W}}\mathbf{e}_1\right)^H\mathbf{y} = \mathbf{e}_1^H\bar{\mathbf{W}}^H\mathbf{y} \tag{25}$$

$$\hat{y}_{rs} = \left(\bar{\mathbf{W}}\mathbf{e}_2\right)^H\mathbf{y} = \mathbf{e}_2^H\bar{\mathbf{W}}^H\mathbf{y} \tag{26}$$

where $\mathbf{e}_2$ selects the second column of a matrix. The solution to (24) is the MWF [13, 11], given by

$$\bar{\mathbf{W}} = \bar{\mathbf{R}}_{\mathbf{yy}}^{-1}\bar{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}}\mathbf{E} \tag{27}$$

where $\mathbf{E}$ selects the columns of $\bar{\mathbf{R}}_{\mathbf{yy}}^{-1}\bar{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}}$ that correspond to the loudspeaker and reference microphone signal. The final expression in (27) is obtained based on the assumption that s and $\mathbf{n}$ are uncorrelated. In practice, by using a voice activity detector (VAD), $\bar{\mathbf{R}}_{\mathbf{yy}}$ and $\bar{\mathbf{R}}_{\mathbf{nn}}$ 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.,

$$\begin{aligned}&\text{if VAD}(l) = 1 : \hat{\mathbf{R}}_{\mathbf{yy}}(l) = \beta\hat{\mathbf{R}}_{\mathbf{yy}}(l-1) + (1-\beta)\mathbf{y}(l)\mathbf{y}^H(l)\\ &\text{if VAD}(l) = 0 : \hat{\mathbf{R}}_{\mathbf{nn}}(l) = \beta\hat{\mathbf{R}}_{\mathbf{nn}}(l-1) + (1-\beta)\mathbf{y}(l)\mathbf{y}^H(l)\end{aligned} \tag{28}$$

where $\hat{\mathbf{R}}_{\mathbf{yy}}(l)$ and $\hat{\mathbf{R}}_{\mathbf{nn}}(l)$ represent estimates of $\bar{\mathbf{R}}_{\mathbf{yy}}$ and $\bar{\mathbf{R}}_{\mathbf{nn}}$ at frame l, respectively. The following criterion will then be used to estimate $\bar{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}}$ [13],

$$\hat{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}} = \underset{\substack{\text{rank}(\bar{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}})=2\\ \bar{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}}\succeq 0}}{\arg\min}\ \left|\left|\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}\left(\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{nn}} - \mathbf{R}_{\boldsymbol{\mathcal{SS}}}\right)\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}\right|\right|_F^2 \tag{29}$$

where $||\cdot||_F$ denotes the Frobenius norm. Spatial pre-whitening is applied by pre- and post-multiplying by $\hat{\mathbf{R}}_{\mathbf{nn}}^{-1/2}$ and $\hat{\mathbf{R}}_{\mathbf{nn}}^{-H/2}$, respectively. The solution to (29) is based on a GEVD of the $(N \times N)$ matrix pencil $\{\hat{\mathbf{R}}_{\mathbf{yy}}, \hat{\mathbf{R}}_{\mathbf{nn}}\}$ [13, 17]

$$\hat{\mathbf{R}}_{\mathbf{yy}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{yy}}\hat{\mathbf{Q}}^H \tag{30}$$

$$\hat{\mathbf{R}}_{\mathbf{nn}} = \hat{\mathbf{Q}}\hat{\mathbf{\Sigma}}_{\mathbf{nn}}\hat{\mathbf{Q}}^H \tag{31}$$

where $\hat{\mathbf{\Sigma}}_{\mathbf{yy}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{nn}}$ are diagonal matrices and $\hat{\mathbf{Q}}$ is an invertible matrix. The rank-2 speech correlation matrix estimate $\hat{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}}$ is then [13]

$$\hat{\mathbf{R}}_{\boldsymbol{\mathcal{SS}}} = \hat{\mathbf{Q}}\text{diag}\{\hat{\sigma}_{y_1} - \hat{\sigma}_{n_1}, \hat{\sigma}_{y_2} - \hat{\sigma}_{n_2}, 0, \ldots, 0\}\hat{\mathbf{Q}}^H \tag{32}$$

where $\hat{\sigma}_{y_i}$ and $\hat{\sigma}_{n_i}$ are the ith diagonal element of $\hat{\mathbf{\Sigma}}_{\mathbf{yy}}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{nn}}$, respectively, corresponding to the ith largest ratio $\hat{\sigma}_{y_i}/\hat{\sigma}_{n_i}$. Using (32) and $\hat{\mathbf{R}}_{\mathbf{yy}}$ (cfr. (30)) in (27), the MWF estimate $\hat{\mathbf{W}}$ can be expressed as

$$\hat{\mathbf{W}} = \hat{\mathbf{Q}}^{-H}\text{diag}\left\{1 - \frac{\hat{\sigma}_{n_1}}{\hat{\sigma}_{y_1}}, 1 - \frac{\hat{\sigma}_{n_2}}{\hat{\sigma}_{y_2}}, 0, \ldots, 0\right\}\hat{\mathbf{Q}}^H\mathbf{E}. \tag{33}$$

The estimates of the total contribution of s and u_s in the loudspeaker and in the reference microphone signals, $\hat{u}_s$ and $\hat{y}_{rs}$, respectively, are now obtained as in (25)-(26) with $\hat{\mathbf{W}}$ replacing $\bar{\mathbf{W}}$

$$\hat{u}_s = \mathbf{e}_1^H\hat{\mathbf{W}}^H\mathbf{y} \tag{34}$$

$$\hat{y}_{rs} = \mathbf{e}_2^H\hat{\mathbf{W}}^H\mathbf{y}. \tag{35}$$

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

$$\hat{\mathbf{y}}_{rs,seg}(l) = \boldsymbol{\mathcal{F}}_R^{-1}\hat{\mathbf{y}}_{rs}^{\kappa}(l) \quad (36)$$

$$\hat{\mathbf{u}}_{s,seg}(l) = \boldsymbol{\mathcal{F}}_R^{-1}\hat{\mathbf{u}}_{s}^{\kappa}(l) \quad (37)$$

$$y_{rs}(t-d_{\mathrm{NR}}) = \sum_{l=0}^{L_f-1} \hat{y}_{rs,seg}\left(t-l\frac{R}{2}\right) g_s\left(t-l\frac{R}{2}\right) \quad (38)$$

$$u_{s}(t-d_{\mathrm{NR}}) = \sum_{l=0}^{L_f-1} \hat{u}_{s,seg}\left(t-l\frac{R}{2}\right) g_s\left(t-l\frac{R}{2}\right). \quad (39)$$

4.2. AFC stage

In the AFC stage a single-channel so-called PEM-AFROW algorithm is used. The implemented algorithm is the frequency-domain version presented in [16] (the reader is referred to [16] for a detailed explanation of the AFC algorithm). The input signals of the AFC algorithm are $\hat{u}_s$ and $\hat{y}_{rs}$.

5. RESULTS

Closed-loop simulations were generated to assess the performance of the proposed algorithm. The investigated scenario had a 4-microphone array with a loudspeaker in front of it reproducing an amplified version of the filtered microphone signals. Room impulse responses of 2048 samples were generated using the randomized image method in [18] at a sampling frequency of 16 kHz for a room of size 5 m × 5 m × 3 m. The STFT frame length was 2048 with a 50% overlap and a squared-root Hann window was used. Since we are mostly focusing on the direct path of component of the feedback path, a $T_{60} = 14$ ms was chosen. A forward path delay and gain of 64 ms and 6.7 dB were used, respectively. The maximum stable gain (MSG) of the closed-loop system was 9.7 dB. A speech signal was used as desired source, i.e. near-end signal, and white noise was added to the microphone signals using different input signal-to-noise ratio (iSNR). The order of the AR model in the PEM-AFROW algorithm was 12. Three metrics are used for the performance evaluation: the misadjustment (Mis), the ASG and the log-spectral distance (SD). The Mis is defined as the normalised distance in dB between the true and estimated feedback path at the reference, $\mathbf{f}_r$ and $\hat{\mathbf{f}}_r$ respectively, as [7]

$$\mathrm{Mis}(l) = 20\log_{10}\left|\frac{\mathbf{f}_r(l)-\hat{\mathbf{f}}_r(l)}{\mathbf{f}_r(l)}\right| \quad \mathrm{dB}. \quad (40)$$

The ASG is based on the so-called MSG which is the maximum gain achievable in the system without it becoming unstable. If the forward path is spectrally flat, the MSG is given by [1]

$$\mathrm{MSG}(l) = -20\log_{10}\left[\max_{\kappa\in\mathcal{P}(l)}\left|f_r(\kappa,l)-\hat{f}_r(\kappa,l)\right|\right] \quad \mathrm{dB} \quad (41)$$

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

$$\mathrm{AGS}(l) = \mathrm{MSG}(l) - K_{\mathrm{MSG}}(l) \quad \mathrm{dB} \quad (42)$$

where $K_{\mathrm{MSG}}(l)$ is the MSG of the system when no feedback canceller is included, i.e., $\hat{f}_r(\kappa,l) = 0\ \forall\kappa,l$, in (41). The SD gives an indication of the distortion of the processed signal. Unweighted and weighted SD measures have been used in the literature [6, 19, 7, 20] for different speech enhancement algorithms. The frequency-weighted SD is defined as in [6]

$$SD(l) = \left(\int_{f_l}^{f_h} w_{ERB}(f)\left(10\log_{10}\frac{\Phi_e(f,l)}{\Phi_r(f,l)}\right)^2 df.\right)^{1/2} \quad (43)$$

Table 1: STOI, SD and PESQ for the Rank-1 and Rank-2 NR-AFC algorithm using a speech as desired signal in closed-loop processing.

SNR	Algorithm	STOI	mean(SD)	max(SD)	PESQ MOS
10 dB	Rank-2 NR-AFC	0.77	19.13	35.89	1.54
	Rank-1 NR-AFC	0.65	46.00	67.52	1.27
0 dB	Rank-2 NR-AFC	0.60	26.48	44.21	1.19
	Rank-1 NR-AFC	0.51	46.74	70.49	1.13

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 $w_{ERB}(f)$ is a weighting function which gives equal weight to each auditory critical band between $f_l = 300$ Hz and $f_h = 6400$ Hz. The measure was computed only during *"speech-plus-noise"* periods and the average over each frame is presented. Additionally, the short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) metrics are also used to assess the performance of the algorithms [21, 22, 7]. The metrics were chosen based on the results presented in [7] where objective and subjective metrics have been assessed for AFC algorithms.

Figure 1 shows the ASG for the proposed algorithm in Section 4 (Rank-2 NR-AFC) and the baseline algorithm in Section 3 (Rank-1 NR-AFC). It is observed that the ASG increases for the proposed algorithm while it decreases for the baseline algorithm for the considered iSNRs. Figure 2 shows the Mis for both algorithms. For the Rank-2 NR-AFC algorithm the Mis slowly decreases whereas for the Rank-1 NR-AFC algorithm it diverges for the considered iSNRs. Table 1 shows the STOI, PESQ and SD metrics for the two iSNRs considered. It is observed that for both iSNRs the Rank-2 NR-AFC outperforms the Rank-1 NR-AFC in terms of STOI and PESQ and SD.

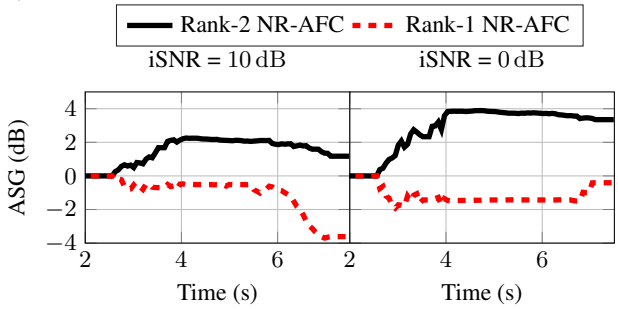

Fig. 1: ASG for the Rank-1 and Rank-2 NR-AFC algorithm using a speech signal as desired signal in closed-loop processing.

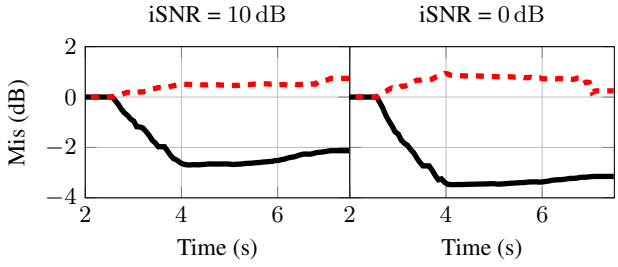

Fig. 2: Mis for the three cascade algorithms using a speech signal as desired signal in closed-loop processing.

6. CONCLUSIONS

A cascade multi-channel NR and AFC algorithm has been presented which uses a rank-2 estimate of the desired speech correlation matrix to compute the MWF in the NR stage. It is shown that for the considered iSNRs the proposed algorithm increases the ASG whilst a similar algorithm using a rank-1 estimate (which is the usual assumption [13] when there is one speaker) does not.

7. REFERENCES

[1] T. van Waterschoot and M. Moonen, "Fifty years of acoustic feedback control: State of the art and future challenges," *Proc. IEEE*, vol. 99, no. 2, pp. 288–327, 2011.

[2] M. Guo, S. H. Jensen, and J. Jensen, "Evaluation of state-of-the-art acoustic feedback cancellation systems for hearing aids," *J. Audio Eng. Soc.*, vol. 61, no. 3, pp. 125–137, 2013.

[3] M. Guo, S. H. Jensen, and J. Jensen, "Novel acoustic feedback cancellation approaches in hearing aid applications using probe noise and probe noise enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2549–2563, 2012.

[4] M. Guo, S. H. Jensen, J. Jensen, and S. L. Grant, "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] A. Spriet, M. Moonen, and I. Proudler, "Feedback cancellation in hearing aids: an unbiased modelling approach," in *Proc. 11th European Signal Process. Conf. (EUSIPCO '02)*. IEEE, 2002, pp. 1–4.

[6] A. Spriet, M. Moonen, and J. Wouters, "Evaluation of feedback reduction techniques in hearing aids based on physical performance measures," *J. Acoust. Soc. Amer.*, vol. 128, no. 3, pp. 1245–1261, 2010.

[7] G. Bernardi, T. van Waterschoot, J. Wouters, and M. Moonen, "Subjective and objective sound-quality evaluation of adaptive feedback cancellation algorithms," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 5, pp. 1010–1024, 2018.

[8] J. Franzen and T. Fingscheidt, "Improved measurement noise covariance estimation for N-channel feedback cancellation based on the frequency domain adaptive Kalman filter," in *Proc. 2019 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '19)*, 2019, pp. 965–969.

[9] H. Schepker, S. E. Nordholm, L. T. T. Tran, and S. Doclo, "Null-steering beamformer-based feedback cancellation for multi-microphone hearing aids with incoming signal preservation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 4, pp. 679–691, 2019.

[10] F. Strasser and H. Puder, "Adaptive feedback cancellation for realistic hearing aid applications," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 12, pp. 2322–2333, 2015.

[11] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the wiener filter for noise reduction," in *Speech Enhancement*, pp. 9–41. Springer, 2005.

[12] J. Benesty, J. R. Jensen, M. G. Christensen, and J. Chen, *Speech enhancement: A signal subspace perspective*, Elsevier, 2014.

[13] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 4, pp. 785–799, 2014.

[14] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[15] A. Spriet, G. Rombouts, M. Moonen, and J. Wouters, "Combined feedback and noise suppression in hearing aids," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 6, pp. 1777–1790, 2007.

[16] G. Bernardi, T. van Waterschoot, J. Wouters, and M. Moonen, "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)*. IEEE, 2015, pp. 1–5.

[17] F. Jabloun and B. Champagne, "Signal subspace techniques for speech enhancement," in *Speech Enhancement*, pp. 135–159. Springer, 2005.

[18] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 4, pp. 774–786, 2015.

[19] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 6, pp. 561–571, 2004.

[20] R. Aichner, *Acoustic blind source separation in reverberant and noisy environments*, Ph.D. thesis, Friedrich-Alexander-Universitat Erlangen-Nurnberg, 2007.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP '10)*. IEEE, 2010, pp. 4214–4217.

[22] ITU-T Rec. P.862, "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.