KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Authors: Arnout Roebben, Jesper Jensen, Toon van Waterschoot, Jan Wouters, and Marc Moonen Title: Analysis of Cascaded Multi-Microphone Feedback Cancellation and Noise Reduction Design Approaches in Hearing Instruments
Archived version	■ Submitted version ☐ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher) ☐ Published ☐ Published (Open Access)
Acknowledgements	In this manuscript, the following funding is acknowledged: ☐ EU H22020 ☐ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ☐ Other



Highlights

Analysis of Cascaded Multi-Microphone Feedback Cancellation and Noise Reduction Design Approaches in Hearing Instruments

Arnout Roebben, Jesper Jensen, Toon van Waterschoot, Jan Wouters, Marc Moonen

- Experimentally analysing algorithms for joint feedback and noise reduction.
- Focusing on cascaded adaptive feedback cancellation and noise reduction filters.
- Analysing steady-state performance for different design approaches for each filter.
- Studying existing solutions, and proposing novel solutions.
- Proposing analysis framework, finetuning strategies, and adapting training strategy.

Analysis of Cascaded Multi-Microphone Feedback Cancellation and Noise Reduction Design Approaches in Hearing Instruments

Arnout Roebben[a,*], Jesper Jensen[b,c], Toon van Waterschoot[a], Jan Wouters[d] and Marc Moonen[a]

[a]KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, Leuven, 3001, Belgium

[b]Demant, Eriksholm Research Centre, Rørtangvej 20, Snekkersten, 3070, Denmark

[c]Aalborg University, Department of Electronic Systems, Fredrik Bajers Vej 7B, Aalborg, 9220, Denmark

[d]KU Leuven, Department of Neurosciences, Research Group ExpORL, Herestraat 49 bus 721, Leuven, 3000, Belgium

ARTICLE INFO

Keywords:
Adaptive Feedback Cancellation
Noise Reduction
Model-Based
Data-Driven
Hybrid
Hearing Instruments

ABSTRACT

In speech processing applications with microphone recording and loudspeaker playback in the same noisy environment, e.g., in hearing instruments, a closed loop is created. Consequently, the recorded desired speech is corrupted by noise and acoustic feedback. To suppress both, a conventional, cascaded system consists of adaptive feedback cancellation (AFC) filters, to partially cancel the feedback, followed by noise reduction (NR) filters, to reduce the noise and residual feedback. In this paper, the aim is to experimentally analyse the performance of different combinations of AFC-filters and NR-filters, constructed following different design approaches. A general multi-microphone scenario with a primary focus on the steady-state performance is considered. The AFC-filters are constructed following a model-based or hybrid design approach, and the NR-filters following a model-based, hybrid, or data-driven design approach. The aim is not to propose novel AFC- or NR-filters, but rather to employ one representative instance for each design approach. Nevertheless, through this comparison novel high-performing combinations are obtained. In particular, the hybrid and data-driven NR-filters, possibly preceded by hybrid AFC-filters, are shown to be effective. Given the multi-microphone system and the interaction of the AFC- and NR-filters through the acoustic coupling, multi-microphone training and finetuning strategies are shown to be effective.

1. Introduction

In speech processing applications with microphone recording and loudspeaker playback in the same acoustic environment, e.g., in hearing instruments, the loudspeaker signals may feed back into the microphones, corrupting the recorded desired speech, as illustrated in Fig. 1 [9, 20, 42]. Through the acoustic coupling ($\mathbf{F}(.)$ in Fig. 1), a closed loop is created, which can be rendered unstable [9, 20, 42]. This onset of instability therefore limits the achievable amplification, which is essential in hearing instruments for compensating hearing loss. Hence feedback cancellation is required to reduce this coupling, thereby allowing for larger amplifications before instability. Adaptive feedback cancellation (AFC) is a particular feedback cancellation strategy, where an estimate of the feedback path ($\hat{\mathbf{F}}(q,t)$ in Fig. 1) is computed and used to filter the loudspeaker signals ($l[t]$ in Fig. 1). This results in an estimate of the feedback signal, which is subtracted from the microphone signals ($\mathbf{m}[t]$ in Fig. 1) [24, 37, 42]. The desired speech may also be corrupted by noise ($\mathbf{n}[t]$ in Fig. 1), hence requiring noise reduction (NR) filters ($\hat{\mathbf{W}}(q,t)$ in Fig. 1) [5].

NR-filters ($\hat{\mathbf{W}}(q,t)$ in Fig. 1), designed for noise reduction, can be constructed following a model-based, data-driven, or hybrid design approach [14]. In a model-based design approach, a signal model is formulated, which is substituted into a cost function, together with the NR-filters that are subject to the optimisation. This leads to NR-filters that are optimal with respect to the formulated signal model and cost function [10, 34]. In a data-driven design

*This research was carried out at the ESAT Laboratory of KU Leuven and at Eriksholm Research Centre, in the frame of Research Council KU Leuven project C14-21-0075 'A holistic approach to the design of integrated and distributed digital signal processing algorithms for audio and speech communication devices', Aspirant Grant 11PDH24N (for A. Roebben), and Grant For A Long Stay Abroad V438025N (for A. Roebben) from the Research Foundation - Flanders (FWO). The scientific responsibility is assumed by its authors. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the FWO and the Flemish Government.

*Corresponding author

arnout.roebben@esat.kuleuven.be (A. Roebben); jesj@demant.com (J. Jensen); toon.vanwaterschoot@esat.kuleuven.be (T.v. Waterschoot); jan.wouters@med.kuleuven.be (J. Wouters); marc.moonen@esat.kuleuven.be (M. Moonen)

ORCID(s): 0000-0003-4665-669X (A. Roebben); 0000-0003-1478-622X (J. Jensen); 0000-0002-6323-7350 (T.v. Waterschoot); 0000-0002-0093-698X (J. Wouters); 0000-0003-4461-0073 (M. Moonen)

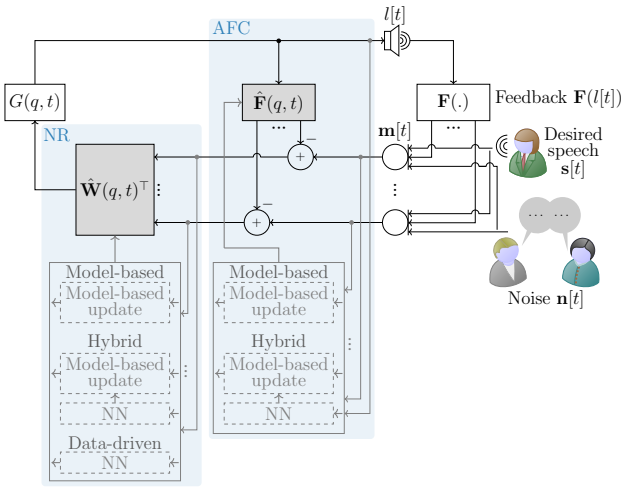

Figure 1: The cascaded filters $\hat{\mathbf{F}}(q,t)$ and $\hat{\mathbf{W}}(q,t)$, implementing adaptive feedback cancellation (AFC) followed by noise reduction (NR), can be designed following a model-based, hybrid, or data-driven design approach. In hybrid and data-driven designs a neural network (NN) is employed.

approach, no specific signal model is assumed. The microphone signals are passed to a neural network, which either directly returns an estimate of the desired speech [40], or a so-called mask to be applied to the microphone signals [6, 27]. This leads to NR-filters that are optimised for a training set and cost function. Data-driven NR-filters are not restricted by model assumptions. These filters, thus, typically, achieve a higher performance than model-based NR-filters, but require training data, lack the interpretability of the signal model in model-based designs, and may be less robust to deviations from prior assumptions [14]. In a hybrid design approach, model-based and data-driven designs are fused to combine the benefits of both [14]. Specific parameters of the model-based NR-filters are estimated using neural networks, unifying the performance of data-driven optimisation and the interpretable signal model of model-based designs [27, 39].

AFC-filters ($\hat{\mathbf{F}}(q,t)$ in Fig. 1), designed as finite impulse response (FIR) filters to approximate amd remove the acoustic coupling, can be constructed following a model-based design approach [13, 17, 30, 37, 42], or a hybrid design approach [24, 35, 49]. As for the NR-filters, the model-based AFC-filters are optimal with respect to the signal model and cost function, while the hybrid AFC-filters reuse this model-based structure but estimate the parameters using neural networks. As the AFC-filters constitute FIR approximations of the feedback path, the AFC-filters will always be interpretable, and hence a data-driven design does not fit this definition.

For joint feedback cancellation and noise reduction, employing cascaded AFC-filters and NR-filters, as visualised in Fig. 1, constitutes an attractive approach [20]. This approach has been widely adopted in the related field of joint acoustic echo cancellation and noise reduction [15], where this approach has also been shown to be optimal [28]. Additionally, this approach allows for modular optimisation of both filters involved, and for a multi-stage approach, where the AFC-filters aim at (partially) cancelling the feedback, such that the NR-filters then aim at reducing the noise and residual feedback[1].

The approach of employing cascaded AFC-filters and NR-filters, has, nevertheless, only been partially explored. The combination of model-based AFC-filters with model-based NR-filters for joint feedback cancellation and noise reduction has been explored in single-microphone scenarios in [3, 21] and in multi-microphone scenarios in [23, 31]. Model-based stand-alone NR-filters with no prior AFC-filters for joint feedback cancellation and noise reduction have been explored in multi-microphone scenarios in [29]. The combination of model-based AFC-filters with data-driven NR-filters for joint feedback cancellation and noise reduction has been explored in single-microphone scenarios in [46, 50] and in multi-microphone scenarios in [22], and data-driven NR-filters with no prior AFC-filters for joint

[1]Due to their origin from the noise reduction functionality, the NR-filters are referred to as such. Yet, for joint feedback cancellation and noise reduction, the NR-filters aim to suppress both noise and (residual) feedback. Furthermore, when there is no noise, the NR-filters only aim at suppressing feedback.

feedback cancellation and noise reduction have been explored in single-microphone scenarios in [47, 50] and in multi-microphone scenarios in [22]. For the case when there is only feedback and no noise, i.e., for feedback cancellation, the combination of model-based AFC-filters and data-driven NR-filters has been explored in single-microphone scenarios in [48], data-driven NR-filters with no prior AFC-filters have been explored for feedback cancellation in single-microphone scenarios in [16, 48], the combination of data-driven NR-filters with hybrid AFC-filters has been explored for feedback cancellation in single-microphone scenarios in [45], and model-based NR-filters with no prior AFC-filters have been explored for feedback cancellation in multi-microphone scenarios in [30]. Alternative approaches exist, such as the hybrid Kalman filter of [49], the cascade of model-based NR-filters and data-driven NR-filters [43], and the model-based AFC-filters integrated into the model-based NR-filters [38].

In this paper, the aim is to systematically, and experimentally, analyse the performance of different cascaded combinations of AFC-filters, designed following a model-based or hybrid design approach, with NR-filters, designed following a model-based, hybrid, or data-driven design approach, as schematically shown in Fig. 1. A general multi-microphone scenario for joint feedback cancellation and noise reduction is considered with a primary focus on the steady-state performance. The aim is not to offer a complete overview of all the methods available, but to focus on the differences between the design approaches by using one representative instance for each design approach. This, nonetheless, enables the study of previously unexplored combinations, including the combination of hybrid AFC-filters with hybrid NR-filters, and hybrid NR-filters with no prior AFC-filters. The deep marginal feedback cancellation (DeepMFC) training framework for data-driven and hybrid single-microphone algorithms employed in a closed loop system is used and adapted for multi-microphone scenarios [50]. An analysis framework is presented, as well as finetuning strategies. Indeed, the combination of hybrid AFC-filters with data-driven or hybrid NR-filters allows for a joint optimisation of the neural networks associated with both filters. For consistency across designs, time-domain AFC-filters are used, as performance-wise time-domain identification is preferred over short-time Fourier transform (STFT) domain identification [1], while for the NR, STFT-domain NR-filters are used as these are generally preferred performance-wise over time-domain NR-filters [36].

The problem of joint feedback cancellation and noise reduction in hearing instruments is described in Section 2. The different design approaches are reviewed in Section 3. In Section 4, the multi-microphone training and finetuning strategies are described, and the experimental setup is detailed in Section 5. Section 6 presents the experimental analysis, and Section 7 the conclusions. Python code is available in [26].

2. Problem Statement

Considering an M-microphone/single-loudspeaker setup[2], as represented in Fig. 1, the microphone signals $m_i[t]$, with $i \in \{1, \cdots, M\}$, at discrete time $t \in \mathbb{N}$, can be stacked in a microphone signal vector $\mathbf{m}[t] \in \mathbb{R}^{M\times 1}$ as

$$\mathbf{m}[t] = \begin{bmatrix} m_1[t] & \cdots & m_M[t] \end{bmatrix}^\top. \tag{1}$$

This microphone signal vector $\mathbf{m}[t]$ can be decomposed into a desired speech signal vector $\mathbf{s}[t] \in \mathbb{R}^{M\times 1}$, a noise signal vector $\mathbf{n}[t] \in \mathbb{R}^{M\times 1}$, and a feedback signal vector $\mathbf{F}(l[t]) \in \mathbb{R}^{M\times 1}$ as

$$\mathbf{m}[t] = \mathbf{F}(l[t]) + \mathbf{s}[t] + \mathbf{n}[t], \tag{2}$$

where $l[t]$ represents the loudspeaker signal, and $\mathbf{F}(.)$ represents a map that models the propagation from the loudspeaker to the microphones. This loudspeaker signal $l[t]$ is obtained by processing the microphone signals and prior loudspeaker signals. To this end, the q-notation is used to represent (possibly time-varying) FIR filters, where q^{-1} represents the one-sampling-period delay operator [37]. For example, the AFC-filter between the loudspeaker and microphone i, $\hat{F}_i(q,t)$, represents the following filter

$$\hat{F}_i(q,t) = \hat{f}_{i,0}[t] + \hat{f}_{i,1}[t]q^{-1} + \cdots + \hat{f}_{i,L_{\hat{F}_i}-1}[t]q^{-L_{\hat{F}_i}+1} \tag{3}$$

The filtering of the $l[t]$ with $\hat{F}_i(q,t)$, i.e., $\hat{F}_i(q,t)l[t]$, can then be alternatively represented as $\hat{\mathbf{f}}_i[t]^\top \mathbf{l}_i[t]$ with

$$\hat{\mathbf{f}}_i[t] = \begin{bmatrix} \hat{f}_{i,0}[t] & \cdots & \hat{f}_{i,L_{\hat{F}_i}-1}[t] \end{bmatrix}^\top \in \mathbb{R}^{L_{\hat{F}_i}\times 1}, \tag{4a}$$

[2]The approaches can be extended to a multi-loudspeaker setup. This is avoided here for clarity, and as the analysed algorithms were originally designed for a single-loudspeaker setup.

$$\mathbf{l}_i[t] = \begin{bmatrix} l[t] & \cdots & l[t - L_{\hat{F}_i} + 1] \end{bmatrix}^\top \qquad \in \mathbb{R}^{L_{\hat{F}_i} \times 1} \tag{4b}$$

the vector containing the coefficients of the filter and the vector containing time-lagged loudspeaker samples, respectively. Thus, with the $M \times 1$-dimensional $\hat{\mathbf{F}}(q,t)$ containing the AFC-filters $\hat{F}_i(q,t)$ as

$$\hat{\mathbf{F}}(q,t) = \begin{bmatrix} \hat{F}_1(q,t) & \cdots & \hat{F}_M(q,t) \end{bmatrix}^\top, \tag{5}$$

with the $M \times 1$-dimensional $\hat{\mathbf{W}}(q,t)$ containing the NR-filters, and with $G(q,t)$ the successive so-called forward path processing (cfr. Fig. 1), the loudspeaker signal $l[t]$ is obtained as

$$l[t] = G(q,t)\hat{\mathbf{W}}(q,t)^\top(\mathbf{m}[t] - \hat{\mathbf{F}}(q,t)l[t]) \tag{6a}$$

$$\triangleq G(q,t)\hat{\mathbf{W}}(q,t)^\top \mathbf{y}[t]. \tag{6b}$$

This forward path processing $G(q,t)$, e.g., includes the amplification in hearing instruments. The NR-filters $\hat{\mathbf{W}}(q,t)$ are applied in the STFT domain with F frequency bins, and employing a half-frame overlap. For discrete-time index t, the frame index k is defined accordingly

$$k = \left\lfloor \frac{t}{R} \right\rfloor, \tag{7}$$

with R the hop size in samples between frames, i.e., $R = \frac{F}{2}$ for a half-frame overlap. Let $\mathbf{w}_s \in \mathbb{R}^{F \times 1}$ be defined as the STFT-synthesis window, $\mathcal{F}_F \in \mathbb{C}^{F \times F}$ as the discrete Fourier transform (DFT) matrix, diag(.) as the diagonal matrix operator, and $\hat{s}_r[f,k]$ as the STFT-domain estimate of the desired speech in reference microphone $r \in \{1, \cdots, M\}$ for frequency bin f and frame k, i.e., the estimate of the STFT-transformed $s_r[t]$, $s_r[f,k]$. The filtering $\hat{\mathbf{W}}(q,t)^\top \mathbf{y}[t]$, $\forall t \in [\frac{kF}{2}, \frac{kF}{2} + \frac{F}{2} - 1]$ corresponds to

$$\begin{bmatrix} \hat{\mathbf{W}}(q, \frac{kF}{2})^\top \mathbf{y}[\frac{kF}{2}] \\ \vdots \\ \hat{\mathbf{W}}(q, \frac{kF}{2} + \frac{F}{2} - 1)^\top \mathbf{y}[\frac{kF}{2} + \frac{F}{2} - 1] \end{bmatrix} = \underbrace{\begin{bmatrix} 0_{\frac{F}{2} \times \frac{F}{2}} & \mathbb{I}_{\frac{F}{2}} \\ \mathbb{I}_{\frac{F}{2}} & 0_{\frac{F}{2} \times \frac{F}{2}} \end{bmatrix}}_{\text{Overlap-add}} \underbrace{\begin{bmatrix} \text{diag}(\mathbf{w}_s)\mathcal{F}_F^{-1} & 0_{F \times F} \\ 0_{F \times F} & \text{diag}(\mathbf{w}_s)\mathcal{F}_F^{-1} \end{bmatrix}}_{\text{Synthesis}} \begin{bmatrix} \hat{s}_r[0, k-1] \\ \vdots \\ \hat{s}_r[F-1, k-1] \\ \hline \hat{s}_r[0,k] \\ \vdots \\ \hat{s}_r[F-1,k] \end{bmatrix}, \tag{8}$$

wherein $0_{F \times F}$ and $\mathbb{I}_{\frac{F}{2}}$ represent $\frac{F}{2} \times \frac{F}{2}$-dimensional zero- and identify-matrices, respectively, and where $\hat{s}_r[f,k] = \hat{\mathbf{w}}[f,k]^H \mathbf{y}[f,k]$ is obtained by applying the STFT-domain NR-filters $\hat{\mathbf{w}}[f,k] \in \mathbb{C}^{M \times 1}$ to the STFT-transformed $\mathbf{y}[t]$, i.e., to $\mathbf{y}[f,k]$. To this end, $\mathbf{y}[f,k]$ is given as

$$\begin{bmatrix} y_i[0,k] & \cdots & y_i[F-1,k] \end{bmatrix}^\top = \mathcal{F}_F \text{diag}(\mathbf{w}_a) \begin{bmatrix} y_i[kR] & \cdots & y_i[kR+F-1] \end{bmatrix}^\top, \tag{9a}$$

$$\mathbf{y}[f,k] = \begin{bmatrix} y_1[f,k] & \cdots & y_M[f,k] \end{bmatrix}^\top, \tag{9b}$$

with $\mathbf{w}_a \in \mathbb{R}^{F \times 1}$ the STFT-analysis window.

Referring to (6), the goal is to design the AFC-filters $\hat{\mathbf{F}}(q,t)$ to (partially) remove the feedback $\mathbf{F}(l[t])$, and to subsequently employ the NR-filters $\hat{\mathbf{W}}(q,t)$ to remove the noise $\mathbf{n}[t]$ and residual feedback $\mathbf{F}(l[t]) - \hat{\mathbf{F}}(q,t)l[t]$, such that an estimate of the desired speech in reference microphone r is obtained. The design of $\hat{\mathbf{W}}(q,t)$ and $\hat{\mathbf{F}}(q,t)$ can be realised following different design approaches as detailed next.

3. Design Approaches

In Section 3.1, representative examples of the model-based design (Section 3.1.1) and the hybrid design (Section 3.1.2) approaches for constructing AFC-filters are reviewed. Similarly, in Section 3.2, representative examples of the model-based design (Section 3.2.1), the hybrid design (section 3.2.2), and the data-driven design (Section 3.2.3) approaches for constructing NR-filters are reviewed.

3.1. Adaptive Feedback Cancellation (AFC)

3.1.1. Model-Based Design

The model-based design for AFC is exemplified by the prediction error method-based adaptive filtering with row operations (PEM-AFROW) algorithm [30]. The PEM-AFROW algorithm employs adaptive filters to update $\hat{F}_i(q,t)$. Yet, as the loudspeaker signal is correlated with the desired speech signal and noise components in the microphone signals, regular adaptive filter updates would cause a bias on the estimate of the feedback path [17, 30]. Consequently, the PEM-AFROW algorithm employs an additional prewhitening filter $\hat{A}_i(q,t)$ of $L_{\hat{A}_i}$ coefficients to reduce this correlation [30].

For the adaptive filter updates, in this paper, a block-NLMS update is used with frames as defined in (7) with $R = F$. With $\hat{\mathbf{f}}_{i,k-1} \in \mathbb{R}^{L_{\hat{F}_i} \times 1}$ the feedback path estimate between the loudspeaker and microphone i, for frame $k-1$, the block-NLMS update is defined as

$$\boxed{\hat{\mathbf{f}}_{i,k} = \hat{\mathbf{f}}_{i,k-1} + \frac{\mu}{\alpha + \left\| L_{i,k}^{(w)} \right\|_F^2} L_{i,k}^{(w)} \left(\mathbf{m}_{i,k}^{(w)} - L_{i,k}^{(w)^\top} \hat{\mathbf{f}}_{i,k-1} \right)}, \tag{10}$$

with step size μ, with regularisation α, with

$$\mathbf{m}_{i,k}^{(w)} = \begin{bmatrix} m_{i,k}^{(w)}[kR] & \cdots & m_{i,k}^{(w)}[kR+F-1] \end{bmatrix}^\top \tag{11a}$$

$$L_{i,k}^{(w)} = \begin{bmatrix} l_{i,k}^{(w)}[kR] & \cdots & l_{i,k}^{(w)}[kR+F-1] \\ \vdots & \ddots & \vdots \\ l_{i,k}^{(w)}[kR-L_{\hat{F}_i}+1] & \cdots & l_{i,k}^{(w)}[kR+F-L_{\hat{F}_i}] \end{bmatrix}, \tag{11b}$$

and with $.^{(w)}$ denoting prewhitening using prewhitening filter $\hat{\mathbf{a}}_{i,k} \in \mathbb{R}^{L_{\hat{A}_i} \times 1}$. To this end, the prewhitened loudspeaker and microphone signals are defined as

$$l_{i,k}^{(w)}[t] = \hat{\mathbf{a}}_{i,k}^\top \begin{bmatrix} l[t] & \cdots & l[t-L_{\hat{A}_i}+1] \end{bmatrix}^\top, \tag{12a}$$

$$m_{i,k}^{(w)}[t] = \hat{\mathbf{a}}_{i,k}^\top \begin{bmatrix} m_i[t] & \cdots & m_i[t-L_{\hat{A}_i}+1] \end{bmatrix}^\top. \tag{12b}$$

This prewhitening filter $\hat{\mathbf{a}}_{i,k}$ is obtained by applying the Levinson-Durbin (LD(.)) algorithm on the feedback-cancelled microphone signals of frame k using the feedback canceller of frame $k-1$ [11], i.e.,

$$\hat{\mathbf{a}}_{i,k} = \mathrm{LD}\left(\begin{bmatrix} m_i[kR] \\ \vdots \\ m_i[kR+F-1] \end{bmatrix} - L_{i,k}^\top \hat{\mathbf{f}}_{i,k-1} \right), \tag{13}$$

with $L_{i,k}$ similar to (11b), yet without prior prewhitening.

As the block-NLMS update adopts frame size F, the in-the-loop estimate of the feedback path is updated at this rate,

$$\boxed{\hat{\mathbf{f}}_i[t] = \hat{\mathbf{f}}_{i,k-1}, \forall t \in [kR, kR+F-1]}. \tag{14}$$

To guarantee identifiability of the estimate of the feedback path, however, the following conditions are assumed to be satisfied [30, 38]

$$\mathbf{F}(l[t]) = \mathbf{F}(q,t) l[t], \tag{15a}$$

$$s_i[t] + n_i[t] = \frac{1}{D_i(q,t)} w[t], \tag{15b}$$

$$\text{Delay of } G(q,t)(\mathbf{F}(q,t) - \hat{\mathbf{F}}(q,t)) \geq L_{\hat{A}_i} \tag{15c}$$

i.e., the feedback path of order $L_{F_i} < L_{\hat{F}_i}$ is linear, both the desired speech and the noise in microphone i are generated by passing white noise $w[t]$ through an autoregressive model $\frac{1}{D_i(q,t)}$ with $L_{D_i} \leq L_{\hat{A}_i}$ denominator coefficients, and the system delay is larger than $L_{\hat{A}_i}$. If this is not the case, a residual bias on the estimate persists [37, 38].

3.1.2. Hybrid Design

In a hybrid design for AFC, specific parameters of a model-based algorithm, e.g., the PEM-AFROW algorithm, are estimated using a neural network. This design is exemplified by the deep feedback cancellation (DFC) algorithm [24], where the estimate of the feedback path is directly supplied by a neural network on a frame-by-frame basis, and which has originally been proposed for single-microphone scenarios. Therefore, with $\mathbf{m}_{i,k} \in \mathbb{C}^{L_K F \times 1}$, $i \in \{1, \cdots, M\}$, and $\mathbf{l}_k \in \mathbb{C}^{L_K F \times 1}$ containing the most recent L_K frames of the STFT-transformed microphone and loudspeaker signals [24],

$$\mathbf{m}_{i,k} = [m_i[0, k - L_K + 1] \ \cdots \ m_i[F-1, k - L_K + 1] \ \cdots \ m_i[0, k] \ \cdots \ m_i[F-1, k]]^\top \tag{16a}$$

$$\mathbf{l}_k = [l[0, k - L_K + 1] \ \cdots \ l[F-1, k - L_K + 1] \ \cdots \ l[0, k] \ \cdots \ l[F-1, k]]^\top \tag{16b}$$

the time-domain estimate for frame k, i.e., $\hat{\mathbf{f}}_{i,k} \in \mathbb{R}^{L_{\hat{F}_i} \times 1}$, is supplied by the neural network $\mathrm{NN}^{\{\hat{\mathbf{f}}\}}(.)$ as

$$\boxed{\hat{\mathbf{f}}_{i,k} = \mathrm{NN}^{\{\hat{\mathbf{f}}\}}\left(\begin{bmatrix}\mathbf{m}_{i,k}^\top & \mathbf{l}_k^\top\end{bmatrix}^\top\right)}. \tag{17}$$

The parameters of the neural network are then optimised on a training set to minimise the normalised Euclidean system distance (NESD) cost function [24]

$$\hat{\mathbf{f}}_{i,k} = \underset{\tilde{\mathbf{f}}_{i,k}}{\operatorname{argmin}} \, 10\log_{10}\left(\frac{\left\|\mathbf{f}_{i,k} - \tilde{\mathbf{f}}_{i,k}\right\|_2^2}{\left\|\mathbf{f}_{i,k}\right\|_2^2}\right), \tag{18}$$

with $\mathbf{f}_{i,k} \in \mathbb{R}^{L_{F_i} \times 1}$ the ground-truth feedback path for frame k. Assumption (15a) is, thus, still adopted, while assumptions (15b) and (15c) are omitted. The in-the-loop estimate of the feedback path is updated at the following rate, with R as per (7),

$$\boxed{\hat{\mathbf{f}}_i[t] = \hat{\mathbf{f}}_{i,k-1}, \forall t \in [kR, kR + F - 1]}. \tag{19}$$

3.2. Noise Reduction (NR)

3.2.1. Model-Based Design

The model-based design for NR is exemplified by the multi-channel Wiener filter (MWF) $\hat{\mathbf{w}}[f,k] \in \mathbb{C}^{M \times 1}$ [4, 10, 34], obtained by minimising the MSE cost function,

$$\hat{\mathbf{w}}[f,k] = \underset{\tilde{\mathbf{w}}[f,k]}{\operatorname{argmin}} \, \mathbb{E}\{|s_r[f,k] - \tilde{\mathbf{w}}[f,k]^H \mathbf{y}[f,k]|^2\}. \tag{20}$$

This results in the MWF $\hat{\mathbf{w}}[f,k]$ [4, 10, 34],

$$\hat{\mathbf{w}}[f,k] = R_{yy}[f,k]^{-1} R_{ys}[f,k]\mathbf{t}_r, \tag{21}$$

with $R_{yy}[f,k] = \mathbb{E}\{\mathbf{y}[f,k]\mathbf{y}[f,k]^H\}$ and $R_{ys}[f,k] = \mathbb{E}\{\mathbf{y}[f,k]\mathbf{s}[f,k]^H\}$, and with $\mathbf{t}_r \in \mathbb{R}^{M \times 1}$ a unit vector with all 0-elements, except for a 1-element at row r. To estimate these correlation matrices, the MWF assumes that the interfering signals are uncorrelated to the desired speech, such that, with $R_{ss}[f,k] = \mathbb{E}\{\mathbf{s}[f,k]\mathbf{s}[f,k]^H\}$, (21) becomes equivalent to

$$\hat{\mathbf{w}}[f,k] = R_{yy}[f,k]^{-1} R_{ss}[f,k]\mathbf{t}_r. \tag{22}$$

Additionally, the availability of a voice activity detector (VAD), VAD$[k]$, is assumed, which allows to differentiate between periods of desired speech activity (VAD$[k]=1$) and inactivity (VAD$[k]=0$). Finally, all interfering signals are assumed to be present when VAD$[k]=0$. Let $\hat{R}_{yy|0}[f,k]$, estimated when VAD$[k]=0$, and $\hat{R}_{yy|1}[f,k]$, estimated when VAD$[k]=1$, be defined as

$$\begin{cases} \mathrm{VAD}[k]=1: \ \hat{R}_{yy|1}[f,k] = \lambda \hat{R}_{yy|1}[f,k-1] + (1-\lambda)(\mathbf{y}[f,k]\mathbf{y}[f,k]^H) \\ \mathrm{VAD}[k]=0: \ \hat{R}_{yy|0}[f,k] = \lambda \hat{R}_{yy|0}[f,k-1] + (1-\lambda)(\mathbf{y}[f,k]\mathbf{y}[f,k]^H), \end{cases} \tag{23}$$

with $\lambda \in [0, 1]$ a smoothing factor, then the estimate of $\boldsymbol{R}_{ss}[f, k]$, i.e., $\hat{\boldsymbol{R}}_{ss}[f, k]$, is assumed to be obtainable as

$$\hat{\boldsymbol{R}}_{ss}[f, k] = \hat{\boldsymbol{R}}_{yy|1}[f, k] - \hat{\boldsymbol{R}}_{yy|0}[f, k], \tag{24}$$

This subtraction can be guided by a generalised eigenvalue decomposition (GEVD) to ensure positive semi-definiteness of $\hat{\boldsymbol{R}}_{ss}[f, k]$, and to enforce a rank constraint on $\hat{\boldsymbol{R}}_{ss}[f, k]$ [34]. For example, when there is only one desired speaker, the desired speech signal vector can be represented as $\mathbf{s}[f, k] = \mathbf{h}[f, k]s_0[f, k]$ with $\mathbf{h}[f, k] \in \mathbb{C}^{M\times 1}$ containing the acoustic transfer functions (ATFs) from the desired speech source to the microphones, and with $s_0[f, k]$ the desired speech source signal [10, 34]. Assuming one desired speaker, the desired speech correlation matrix is then a rank-1 matrix as $\boldsymbol{R}_{ss}[f, k] = \mathbb{E}\{s_0[f, k]s_0[f, k]^*\}\mathbf{h}[f, k]\mathbf{h}[f, k]^H$. Both constraints are enforced by computing the GEVD of the matrix pencil $\{\hat{\boldsymbol{R}}_{yy|1}[f, k], \hat{\boldsymbol{R}}_{yy|0}[f, k]\}$ [2, 34]

$$\hat{\boldsymbol{R}}_{yy|1}[f, k] = \hat{\boldsymbol{Q}}[f, k]\hat{\boldsymbol{\Sigma}}_{yy|1}[f, k]\hat{\boldsymbol{Q}}[f, k]^H \tag{25a}$$

$$= \hat{\boldsymbol{Q}}[f, k]\text{diag}(\hat{\sigma}^{(1)}_{y|1}[f, k], \cdots, \hat{\sigma}^{(M)}_{y|1}[f, k])\hat{\boldsymbol{Q}}[f, k]^H, \tag{25b}$$

$$\hat{\boldsymbol{R}}_{yy|0}[f, k] = \hat{\boldsymbol{Q}}[f, k]\hat{\boldsymbol{\Sigma}}_{yy|0}[f, k]\hat{\boldsymbol{Q}}[f, k]^H \tag{25c}$$

$$= \hat{\boldsymbol{Q}}[f, k]\text{diag}(\hat{\sigma}^{(1)}_{y|0}[f, k], \cdots, \hat{\sigma}^{(M)}_{y|0}[f, k])\hat{\boldsymbol{Q}}[f, k]^H, \tag{25d}$$

with $\hat{\boldsymbol{Q}}[f, k] \in \mathbb{C}^{M\times M}$ containing the generalised eigenvectors, and $\hat{\boldsymbol{\Sigma}}_{yy|1}[f, k] \in \mathbb{C}^{M\times M}$ and $\hat{\boldsymbol{\Sigma}}_{yy|0}[f, k] \in \mathbb{C}^{M\times M}$ containing diagonal elements $\left\{\hat{\sigma}^{(1)}_{y|1}[f, k], \cdots, \hat{\sigma}^{(M)}_{y|1}[f, k]\right\}$ and $\left\{\hat{\sigma}^{(1)}_{y|0}[f, k], \cdots, \hat{\sigma}^{(M)}_{y|0}[f, k]\right\}$ that define the generalised eigenvalues $\left\{\hat{\sigma}^{(1)}_{y|1}[f, k]/\hat{\sigma}^{(1)}_{y|0}[f, k], \cdots, \hat{\sigma}^{(M)}_{y|1}[f, k]/\hat{\sigma}^{(M)}_{y|0}[f, k]\right\}$. Herein, it is assumed that, without loss of generality, the generalised eigenvalues are sorted according to a decreasing ratio $\hat{\sigma}^{(i)}_{y|1}/\hat{\sigma}^{(i)}_{y|0}$. When subtracting $\hat{\boldsymbol{R}}_{yy|0}[f, k]$ from $\hat{\boldsymbol{R}}_{yy|1}[f, k]$ as in (24), only the first generalised eigenvalue is retained assuming one desired speaker, thereby leading to the GEVD-based MWF (GEVD-MWF) [34],

$$\boxed{\hat{\mathbf{w}}[f, k] = \hat{\boldsymbol{Q}}[f, k]^{-H}\text{diag}\left(\frac{\hat{\sigma}^{(1)}_{y|1}[f, k] - \hat{\sigma}^{(1)}_{y|0}[f, k]}{\hat{\sigma}^{(1)}_{y|1}[f, k]}, 0, \cdots, 0\right)\hat{\boldsymbol{Q}}[f, k]^H\mathbf{t}_r}. \tag{26}$$

While the assumptions used in the derivation of this GEVD-MWF are satisfied when the desired speech is corrupted by additive, uncorrelated noise, these assumptions are not met for the residual feedback under imperfect AFC. Due to the closed loop, the residual feedback originating from the desired speech is indeed correlated with the desired speech. Further, during the ideal VAD-regimes, the following signals are observed

$$\begin{cases} \text{VAD}[k] = 1: \ (\mathbf{F}(l[t]) - \hat{\mathbf{F}}(q, t)l[t]) + \mathbf{s}[t] + \mathbf{n}[t] \\ \text{VAD}[k] = 0: \ (\mathbf{F}(l^{(n)}[t]) - \hat{\mathbf{F}}(q, t)l^{(n)}[t]) + \mathbf{n}[t], \end{cases} \tag{27}$$

with $l^{(n)}[t]$ the loudspeaker signal when only the noise source is active. As the feedback component originating from the desired speech is available only when the desired speech is available, unless the AFC removes all feedback, the GEVD-MWF will effectively aim at reconstructing both the desired speech and all correlated feedback components.

3.2.2. Hybrid Design

The hybrid design for NR is exemplified by a hybrid design for the GEVD-MWF, as originally proposed for a plain NR task in [27]. The structure of the GEVD-MWF is retained, but the correlation matrices $\hat{\boldsymbol{R}}_{yy|1}[f, k]$ and $\hat{\boldsymbol{R}}_{yy|0}[f, k]$ are directly estimated using a neural network. Specifically, the upper-triangular Cholesky factors $\hat{\boldsymbol{\Gamma}}_{yy|1}[f, k] \in \mathbb{C}^{M\times M}$ and $\hat{\boldsymbol{\Gamma}}_{yy|0}[f, k] \in \mathbb{C}^{M\times M}$ are obtained as the output of the neural network, and used to construct the correlation matrices as

$$\hat{\boldsymbol{R}}_{yy|1}[f, k] = \hat{\boldsymbol{\Gamma}}_{yy|1}[f, k]\hat{\boldsymbol{\Gamma}}_{yy|1}[f, k]^H, \tag{28a}$$

$$\hat{\boldsymbol{R}}_{yy|0}[f, k] = \hat{\boldsymbol{\Gamma}}_{yy|0}[f, k]\hat{\boldsymbol{\Gamma}}_{yy|0}[f, k]^H. \tag{28b}$$

	#Microphones	In-the-loop processing
Single-microphone DeepMFC [50]	1	$G(q,t)$
Multi-microphone DeepMFC	M	$G(q,t)\mathbf{t}_r^\top$
FIN_1 Finetuning	M	$G(q,t)\mathbf{t}_r^\top$
FIN_2 Finetuning	M	$G(q,t)\hat{\mathbf{W}}(q,t)^\top$

Table 1
Summary of the training and finetuning frameworks.

To this end, $\mathbf{y}_k \in \mathbb{C}^{ML_KF\times 1}$, which is defined as

$$\mathbf{y}_k = [\mathbf{y}[0,k-L_K+1]^\top \ \cdots \ \mathbf{y}[F-1,k-L_K+1]^\top \ \cdots \ \mathbf{y}[0,k]^\top \ \cdots \ \mathbf{y}[F-1,k]^\top]^\top, \tag{29}$$

is supplied as input to the neural network $\text{NN}^{\{\hat{R}\}}(.)$, yielding

$$\boxed{\left[\hat{\Gamma}_{yy|1}[f,k]^\top \quad \hat{\Gamma}_{yy|0}[f,k]^\top\right]^\top = g\left(\text{NN}^{\{\hat{R}\}}\left(\mathbf{y}_k\right)\right)}, \tag{30}$$

with $g(.)$ the element-wise activation function of the final layer of the neural network $\text{NN}^{\{\hat{R}\}}$. As per [27], $g(.)$ is selected such that $\hat{\boldsymbol{R}}_{yy|1}[f,k]$ and $\hat{\boldsymbol{R}}_{yy|0}[f,k]$ are valid correlation matrices, i.e., Hermitian and positive semi-definite. The GEVD of these correlation matrices is then computed according to (25), after which the GEVD-MWF is obtained according to (26). The parameters of the neural network are optimised by minimising the MSE cost (20) on a training set.

By estimating the correlation matrices and subtracting them through GEVD operations, the model-structure of (26) is reused, such that the interfering signals are still assumed to be uncorrelated with the desired speech, as detailed in Section 3.2.1. Thus, a mismatch still persists between the model-structure and (21), although the hybrid GEVD-MWF does not require a VAD opposed to the model-based GEVD-MWF.

3.2.3. Data-Driven Design

In a data-driven design for NR, no specific signal model is assumed, such that the neural network either provides a direct estimate of the desired speech [40, 47, 50], or provides a so-called mask (filter) to be applied to the microphone signals [6, 27]. The mask estimation is expressed by means of $\mathbf{y}_k$, as defined in (29), that is supplied to the neural network $\text{NN}^{\{\hat{\mathbf{w}}\}}(.)$, resulting in

$$\boxed{\hat{\mathbf{w}}[f,k] = \text{NN}^{\{\hat{\mathbf{w}}\}}\left(\mathbf{y}_k\right)}, \tag{31}$$

which corresponds to the multi-microphone mask as defined in [6, 27]. The parameters of the neural network are optimised on a training set with respect to the MSE cost (20).

4. Training And Finetuning

Table 1 summarises the following training strategies for the optimisation of the neural network parameters in the hybrid and data-driven designs as used in a closed loop system.

4.1. Single-Microphone DeepMFC Training Framework [50]

Direct closed-loop optimisation of neural networks related to the AFC- and NR-filters is challenging, since the closed loop system can easily be rendered unstable during the optimisation [47, 50]. Therefore, the deep marginal feedback cancellation (DeepMFC) framework has been proposed for training data-driven and hybrid single-microphone ($M = 1$) algorithms employed in a closed loop system, either in the presence or absence of noise [50]. This DeepMFC training framework has been successfully applied when either $\hat{\mathbf{W}}(q,t)$ or $\hat{\mathbf{F}}(q,t)$ is estimated by a neural network, e.g., to optimise the neural network parameters associated with $\hat{\mathbf{W}}(q,t)$ with $\hat{\mathbf{F}}(q,t) = \mathbf{0}$ in [16, 50], or to optimise the neural network parameters associated with $\hat{\mathbf{F}}(q,t)$ with $\hat{\mathbf{W}}(q,t) = 1$ in [24]. In a first step, closed-loop

training signals are generated with $\hat{\mathbf{W}}(q,t) = 1$ and $\hat{\mathbf{F}}(q,t) = \mathbf{0}$ [50]. In a second step, the neural network is optimised in an off-line fashion similar to the supervised optimisation process for NR [6, 40]. Indeed, the generated microphone - loudspeaker signals can be interpreted as desired speech corrupted by interfering signals, from which $\hat{\mathbf{W}}(q,t)$ or $\hat{\mathbf{F}}(q,t)$ should remove the noise and/or feedback. As the microphone - loudspeaker signals are generated without the influence of $\hat{\mathbf{W}}(q,t)$ and $\hat{\mathbf{F}}(q,t)$, there is a mismatch between the training signals used to optimise the neural network and the closed loop system at inference when $\hat{\mathbf{W}}(q,t)$ or $\hat{\mathbf{F}}(q,t)$ are present. Despite this mismatch, the DeepMFC strategy has been applied successfully [16, 24, 50].

4.2. Multi-Microphone DeepMFC Training Framework

In this paper, the DeepMFC framework is adapted for multi-microphone scenarios. Similar to the single-microphone DeepMFC framework, $\hat{\mathbf{W}}(q,t)$ is set to $\mathbf{t}_r$ and $\hat{\mathbf{F}}(q,t)$ is set to $\mathbf{0}$ in the closed loop system, as visualised in Fig. 2. The loudspeaker signals are then generated in closed loop by applying $G(q,t)$ to the signals in reference microphone r, while the feedback, desired speech and noise are simultaneously captured by all microphones. Consequently, multiple microphone - loudspeaker signals are generated, which can be used for off-line optimisation of the neural networks similar to supervised optimisation for NR. As the loudspeaker signals, generated in this way, are not influenced by the other microphones, the mismatch between training and inference in this multi-microphone setting is increased with respect to the single-microphone framework. Still, in Section 6, it will be shown that this extension is effective.

In principle, after this initial training, closed-loop training does become feasible as the neural network weights are well-initialised due to the prior DeepMFC training [16, 48]. Yet, such additional training step is omitted here as solely training using the DeepMFC strategy has been shown to be effective [24, 50], and as the primary focus is a comparison of the design approaches rather than maximising performance.

4.3. Finetuning

The structure of cascaded AFC- and NR-filters is inherently modular as both filters can be designed separately and inserted in the loop in a plug-and-play fashion. To maintain this modular structure of the AFC- and NR-filters, the multi-microphone DeepMFC training strategy is initially applied separately for $\hat{\mathbf{F}}(q,t)$ (with $\hat{\mathbf{W}}(q,t) = \mathbf{t}_r$) and $\hat{\mathbf{W}}(q,t)$ (with $\hat{\mathbf{F}}(q,t) = \mathbf{0}$), i.e., both filters are optimised in isolation. Nevertheless, as a consequence of adhering to the modular nature of the AFC- and NR-filters, the mutual influence of $\hat{\mathbf{F}}(q,t)$ and $\hat{\mathbf{W}}(q,t)$ at inference is neglected during training. To address this limitation, two finetuning strategies are proposed, which can be applied after the prior training of $\hat{\mathbf{F}}(q,t)$ and $\hat{\mathbf{W}}(q,t)$ in isolation. In a first finetuning strategy (FIN_1), the training signals, as generated according to Fig. 2, are passed through both $\hat{\mathbf{F}}(q,t)$ and $\hat{\mathbf{W}}(q,t)$ during training, as the input signals to the neural network associated with $\hat{\mathbf{W}}(q,t)$ are then affected by $\hat{\mathbf{F}}(q,t)$, and as the cost can subsequently flow back all the neural networks involved. In a second finetuning strategy (FIN_2), new training signals are generated by including $\hat{\mathbf{W}}(q,t)$, as previously trained in isolation, in the closed loop system, as visualised in Fig. 3. This allows to create training signals, where the loudspeaker signals are affected by all microphone signals and by $\hat{\mathbf{W}}(q,t)$. The multi-microphone DeepMFC framework can then be applied to finetune the neural network associated with $\hat{\mathbf{F}}(q,t)$. This FIN_2 finetuning strategy is similar to the single-microphone finetuning strategy of [45] where $\hat{\mathbf{F}}(q,t)$ was included in the loop to finetune $\hat{\mathbf{W}}(q,t)$. However, as multi-microphone systems are considered, the microphone signals should be combined into one loudspeaker signal, such that the strategy of [45] cannot be readily applied. To reduce the computational overhead, as the generation of new training signals is required, one fourth of the original training data is used to generate these examples, and this will be shown effective in Section 6.

5. Experiment Design

5.1. Dataset

5.1.1. Acoustic Scenarios

To represent the ATFs from the sources to the hearing instrument microphones, measured impulse responses are required from external loudspeakers to the hearing instrument microphones. These external microphones then model desired speech and noise sources. Additionally, impulse responses from the hearing instrument loudspeaker to the hearing instrument microphones are required in order to represent the feedback paths. In Table 2, the datasets used to create the training, validation and test sets are summarised. The test set is constructed using the Hearpiece dataset of [7], which considers in-the-ear (ITE) hearing instruments and does provide impulse responses from external loudspeakers

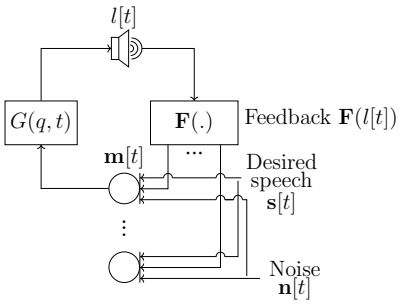

Figure 2: Generation of training signals for the multi-microphone adaptation of the DeepMFC training strategy, wherein $\hat{\mathbf{W}}(q,t)$ is set to $\mathbf{t}_r$ and $\hat{\mathbf{F}}(q,t)$ is set to $\mathbf{0}$ in the closed loop system. These examples are also used for the FIN_1 finetuning.

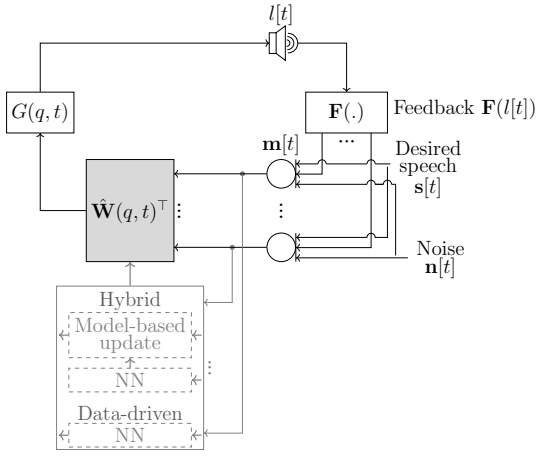

Figure 3: Generation of multi-microphone DeepMFC training signals for the FIN_2 finetuning of $\hat{\mathbf{F}}(q,t)$ to the effect of $\hat{\mathbf{W}}(q,t)$ by including the latter filter in the closed loop system.

to the hearing instrument microphones, as well as impulse responses from the hearing instrument loudspeaker to the microphones. Although this dataset has been recorded in an anechoic room, it is, to the best of our knowledge, the only large-scale, publicly available, dataset for hearing instruments that contains the ATFs and the corresponding feedback paths to the hearing instrument microphones. To this end, $M = 2$ microphones are considered together with one desired speech source and one noise source. As microphones, the in-ear - outer-vent microphone pair of the hearing instrument considered in [7] is used. The algorithms cannot exactly cancel all interfering sources as the total number of interfering sources, i.e., the feedback and noise source, exceeds the degrees of freedom as defined by the $M = 2$ microphones. The test set, containing 968 scenarios, is subsequently constructed from twenty subjects, where one of fifteen external loudspeaker positions is sampled with replacement as the desired speech source position and a distinct external loudspeaker position is sampled for the noise source position. Either the inner-driver or the outer-driver hearing instrument loudspeaker is selected from the hearing instrument considered in [7], as well as the hearing instrument microphones at the same side of this loudspeaker, thus considering monaural scenarios.

The training set is created by combining the source-to-microphone impulse responses from three distinct subjects and fifty distinct external loudspeaker positions, modelling desired speech and noise sources, of the Hearpiece dataset [7] with simulated feedback path impulse responses [50] and measured feedback path impulse responses from other feedback path datasets [18, 32]. The validation set is similarly created by combining the source-to-microphone impulse responses from three distinct subjects and fifteen distinct external loudspeaker positions, sampled from the Hearpiece dataset [7], with the simulated feedback paths [50] and feedback paths from other datasets [18, 32]. While this creates a mismatch between the ATFs and the feedback paths, it will be shown (Section 6) to be effective.

	Training	Validation	Testing
Source-to-microphone	Hearpiece dataset [7] 3 subjects 50 source positions	Hearpiece dataset [7] 3 unseen subjects 15 unseen source positions	Hearpiece dataset [7] 20 unseen subjects 15 unseen source positions
Feedback	Dataset [18] (16 203 paths) Dataset [32] (4 050 paths) Simulated [50] (20 256 paths)	Dataset [18] (1 800 paths) Dataset [32] (450 paths) Simulated [50] (2 251 paths)	Dataset [7] (968 paths)
	40 509 scenarios	4 501 scenarios	968 scenarios

Table 2
Summary of the acoustic scenarios of the training, validation and test set.

For the simulated feedback paths, 10 000 pairs are generated. The feedback path to the first microphone is generated according to the procedure of [50], modelling the feedback path as the product of a sine wave, a zero-mean Gaussian random process, and a decaying exponential. The feedback path to the second microphone is generated with the same parameters except for a random increase of the start-decay time parameter as considered in the framework of [50] with either zero or one sample, in order to create a pair of impulse responses with similar properties as the inter-microphone spacing in hearing instruments is small [9]. Indeed, typically the inter-microphone spacing is around 10 mm, corresponding to a delay of less than one sample for a sampling frequency of 16 kHz. These simulated feedback paths are used for 50 % of the training and validation set, as 20 256 and 2 251 feedback paths, respectively, are sampled with replacement. Additionally, the training set contains 16 203 feedback paths from 80 % of the available feedback paths from [18] and 4 050 feedback paths from [32], as sampled with replacement, thereby accounting for 40 % and 10 % of the feedback paths. Similarly, the validation set contains 1 800 feedback paths from [18] with distinct subjects as for the training set, and 450 feedback paths from the other 20 % of the available feedback paths from [32].

5.1.2. Source Signals

The desired speech signal is selected from the centre for speech technology voice cloning toolkit (CSTR VCTK) dataset [44]. 72%, 8%, and 20% of the speakers are assigned to training, validation, and test set. Segments of prominent silence are identified using a VAD applied to the clean speech, and removed. For training and validation, 2 s-segments are selected with possible overlap. The test set contains 3 s-segments. The noise signal is selected from the diverse environments multi-channel acoustic noise database (DEMAND) [41], with 76.5%, 8.5%, and 15% of the noise types assigned to training, validation, and test set. All signals are sampled at 16 kHz.

5.1.3. Dataset-Specific Considerations

As the inter-subject variance for the feedback paths is small in the Hearpiece dataset [7], the feedback path impulse response to the first microphone is rescaled such that the coefficient with the largest magnitude in absolute value lies randomly within the range $[0.1, 0.31]$ as per [24]. The feedback path impulse response to the second microphone is then rescaled, such that its largest magnitude coefficient in absolute value corresponds to a factor, which is randomly sampled within the range $[0.8, 1]$, times the largest magnitude coefficient of the feedback path impulse response to the first microphone. This range ensures that the magnitude of the second feedback path is close to the magnitude of the first feedback path in line with the small inter-microphone distance [9, 20]. The source-to-microphone impulse responses are analogously rescaled to have a maximum amplitude in absolute value randomly sampled from the range $[0.09, 0.1]$ to promote diversity in the amplitudes. This procedure is also applied to the other datasets.

The root mean square (RMS) of the desired speech source signal is set to one, and the power of the noise source signal is adjusted to achieve a signal-to-noise ratio (SNR) uniformly sampled within the range $[-5, 5]$ dB for the training and validation set, and an SNR of 0 dB for the test set. The filter $G(q, t)$ consists of a delay and full-band gain to model postprocessing by the hearing instrument as per [24, 30, 37, 42]. The overall processing delay is set to 10 ms, which is typical for hearing instruments [9, 20]. The gain for the training and validation set is randomly sampled to be $[3, 0.1]$ dB below the maximum stable gain (MSG) as per [24, 42], for the scenario of Fig. 2. The MSG defines the gain for

which the system becomes unstable, and is defined as [42]

$$\text{MSG [dB]} = -20\log_{10}\left(\max_{\omega\in\mathcal{P}} |G(\omega)F_r(\omega)|\right), \tag{32}$$

with radial frequency ω and $\mathcal{P} = \{\omega|\angle G(\omega)F_r(\omega) = n2\pi\}$ the frequencies for which the system can become unstable [42]. The gain for the test set is fixed 1 dB below the MSG. All feedback path impulse responses are of length $L_{F_1} = L_{F_2} = 64$ samples. For the test set, when explicitly stated, mild non-linearities are inserted to simulate loudspeaker non-linearities as per [33]. The feedback signal $\mathbf{F}(l[t])$ is then generated as

$$\mathbf{F}(l[t]) = \mathbf{F}(q,t)\frac{\text{atan}(\alpha l[t])}{\alpha}, \tag{33}$$

with $\mathbf{F}(q,t)$ the feedback path impulse responses and $\alpha = 10^{-1}$.

5.2. Algorithm Settings

Table 3 summarises the employed algorithms used in the simulations.

5.2.1. Neural Network Architectures

The causal neural network architecture from [24] is adopted for the DFC (Section 3.1.2). This architecture with 857 235 parameters accepts the log-magnitude and phase of the ith microphone signal and the loudspeaker signal in the STFT domain as its input, and returns the estimate for the feedback path between the ith microphone and the loudspeaker according to (17). A discrete Fourier transform (DFT) of 8 ms and a hop size of 2 ms, i.e., $F = 128$ and $R = 32$, is adopted as per [24]. The causal convolutional recurrent network (CRN) architecture from [27], a pruned version of the architecture of [40], is adopted for the hybrid GEVD-MWF (Section 3.2.2) and its data-driven counterpart (Section 3.2.3). A discrete Fourier transform (DFT) of 8 ms and a hop size of 4 ms, i.e., $F = 128$ and $R = 64$, is adopted. This architecture has 109 337 and 108 940 parameters for the correlation estimation and mask estimation, and takes the real and imaginary parts of $\mathbf{y}_k$ as its input. When comparing the case where $\hat{\mathbf{F}}(q,t)$ is omitted to the case where both $\hat{\mathbf{W}}(q,t)$ and $\hat{\mathbf{F}}(q,t)$ are estimated by neural networks, the neural network architecture related to $\hat{\mathbf{W}}(q,t)$ is altered to have around 1M neural network parameters by setting the number of output channels to six in each encoder layer. This enables a fair comparison in terms of neural network parameters, as the case where both $\hat{\mathbf{W}}(q,t)$ and $\hat{\mathbf{F}}(q,t)$ are estimated by neural networks also employs around 1M neural network parameters in total.

5.2.2. Optimisation

The neural networks associated with the AFC- or NR- filters are initially optimised for their respective feedback cancellation and noise reduction task in isolation. To this end, the neural network associated with the NR-filters is initially optimised using the multi-microphone training set of Section 5.1 in the absence of any feedback signals, i.e., $G(q,t) = 0$. The algorithms resulting from the optimisation on this training set are denoted by subscript $._{\text{NR}}$. Similarly, the neural network associated with the AFC-filters is initially optimised using the training set of Section 5.1, in the absence of the noise sources and using only the reference microphone ($\hat{\mathbf{W}}(q,t) = \mathbf{t}_r$). The Adam optimiser with an initial learning rate of $2\cdot10^{-4}$, a decay factor of 0.98 every 2 epochs, early stopping with a 5-epoch patience, and a maximum of 50 epochs is used. The batch size is set to 8, and gradient norms exceeding 1 are clipped. As the NR-filters are typically also leveraged to cancel (residual) feedback, the neural network associated with the NR-filters is also optimised for the multi-microphone dataset of Section 5.1 in the presence of both noise and feedback, but in absence of the AFC-filters to preserve its modular nature, as detailed in Section 4.2. The neural network weights are initialised using the neural network trained on the noise reduction task, and the same optimisation parameters are employed. The resulting algorithms are denoted by subscript $._{\text{AFC+NR}}$. For the finetuning (Section 4), the initial learning rate is reduced to $2\cdot10^{-5}$ and the maximum number of epochs to 10.

5.2.3. Algorithm Specifics

For the model-based GEVD-MWF, either an oracle VAD is assumed available as denoted by a $._{\text{s}}$ subscript, or the WebRTC VAD, which operates on the microphone signals, is used as denoted by a $._{\text{m}}$ subscript [12]. As the model-based NR-filters require substantial convergence time, the model-based NR-filters are only inserted into the loop during

Algorithm	AFC-filters			NR-filters		
	Design	#Parameters NN	Interferer training NN	Design	#Parameters NN	Interferer training NN
CAF	Model-based	0				
PEM-AFROW	Model-based	0				
DFC	Hybrid	857 235	Feedback			
MWF_*				Model-based	0	
Corr_{NR} [100k]				Hybrid	109 337	Noise
$\text{Corr}_{\text{AFC+NR}}$ [100k]				Hybrid	109 337	Noise and feedback
Mask_{NR} [100k] / [1M]				Data-driven	108 940 / 1 007 190	Noise
$\text{Mask}_{\text{AFC+NR}}$ [100k] / [1M]				Data-driven	108 940 / 1 007 190	Noise and feedback
PEM-AFROW - MWF_*	Model-based	0		Model-based	0	
PEM-AFROW - $\text{Corr}_{\text{AFC+NR}}$ [100k]	Model-based	0		Hybrid	109 337	Noise and feedback
PEM-AFROW - $\text{Mask}_{\text{AFC+NR}}$ [100k]	Model-based	0		Data-driven	108 940	Noise and feedback
DFC - MWF_*	Hybrid	857 235	Feedback	Model-based	0	
DFC - $\text{Corr}_{\text{AFC+NR}}$ [100k]	Hybrid	857 235	Feedback	Hybrid	109 337	Noise and feedback
DFC - $\text{Mask}_{\text{AFC+NR}}$ [100k]	Hybrid	857 235	Feedback	Data-driven	108 940	Noise and feedback

Table 3
Summary of the algorithms. MWF_* can either correspond to MWF_s, representing usage of an oracle VAD, or to MWF_m, representing usage of the WebRTC VAD [12], which operates on the microphone signals. CAF represents regular adaptive filter updates, i.e., the PEM-AFROW algorithm without prewhitening filters.

the second half of the signal duration. The smoothing factor λ is set to 0.995 as per [27]. For the PEM-AFROW, the step size μ is set to 0.04, and the regularisation α to 10^{-6} as per [42]. To promote distinct, positive eigenvalues $\hat{\boldsymbol{R}}_{mm}[f,k]$ and $\hat{\boldsymbol{R}}_{nn}[f,k]$ are regularised with $\text{diag}\left(\begin{bmatrix}1.2 & 1.4\end{bmatrix}\right)\cdot 10^{-6}$ and $\text{diag}\left(\begin{bmatrix}1.1 & 0.9\end{bmatrix}\right)\cdot 10^{-6}$. The first microphone is always selected as reference microphone.

5.3. Performance Measures

Using impulse responses to model the feedback path, as described in Section 5.1, the system is governed by linear time-varying filters, such that (2) can be written as

$$\mathbf{m}[t] = \mathbf{F}(q,t)l[t] + \mathbf{s}[t] + \mathbf{n}[t] \tag{34a}$$
$$= \mathbf{F}(q,t)(l^{(s)}[t] + l^{(n)}[t]) + \mathbf{s}[t] + \mathbf{n}[t] \tag{34b}$$
$$= \mathbf{e}^{(s)}[t] + \mathbf{e}^{(n)}[t] + \mathbf{s}[t] + \mathbf{n}[t], \tag{34c}$$

where the feedback signal is decomposed as the sum of the feedback signal originating from the desired speech source $\mathbf{e}^{(s)}[t] \in \mathbb{R}^{M\times 1}$ and the feedback signal originating from the noise source $\mathbf{e}^{(n)}[t] \in \mathbb{R}^{M\times 1}$. The corresponding loudspeaker signals are denoted as $l^{(s)}[t]$ and $l^{(n)}[t]$, respectively. After applying $\hat{\mathbf{W}}(q,t)$ and $\hat{\mathbf{F}}(q,t)$, the desired speech in the reference microphone, i.e., $s_r[t]$, is ideally retained, while the interfering noise, feedback of the desired speech, and feedback of the noise are reduced. Therefore, it is proposed to use the ratio between the desired speech power and the power of each interferer, in a system without $\hat{\mathbf{W}}(q,t)$ and $\hat{\mathbf{F}}(q,t)$ inserted (cfr. Fig. 2), compared to a system with both inserted (cfr. Fig. 1). In other words, it is proposed to use the signal-to-noise ratio (SNR) improvement (ΔSNR), the signal-to-speech-feedback ratio improvement ($\Delta\text{SFR}^\text{s}$) [38], the signal-to-noise-feedback ratio improvement ($\Delta\text{SFR}^\text{n}$) [38], and the speech distortion (SD) [36], defined as

$$\Delta\text{SNR[dB]} = 10\log_{10}\left(\frac{P^{(s)}_{\text{proc.}}}{P^{(n)}_{\text{proc.}}}\right) - 10\log_{10}\left(\frac{P^{(s)}_{\text{no-proc.}}}{P^{(n)}_{\text{no-proc.}}}\right), \tag{35a}$$

$$\Delta\text{SFR}^\text{s}\text{[dB]} = 10\log_{10}\left(\frac{P^{(s)}_{\text{proc.}}}{P^{(e^{(s)})}_{\text{proc.}}}\right) - 10\log_{10}\left(\frac{P^{(s)}_{\text{no-proc.}}}{P^{(e^{(s)})}_{\text{no-proc.}}}\right), \tag{35b}$$

$$\Delta\text{SFR}^\text{n}\text{[dB]} = 10\log_{10}\left(\frac{P^{(s)}_{\text{proc.}}}{P^{(e^{(n)})}_{\text{proc.}}}\right) - 10\log_{10}\left(\frac{P^{(s)}_{\text{no-proc.}}}{P^{(e^{(n)})}_{\text{no-proc.}}}\right), \tag{35c}$$

$$\text{SD[dB]} = 10\log_{10}\left(\frac{P^{(s)}_{\text{proc.}}}{P^{(s)}_{\text{no-proc.}}}\right), \tag{35d}$$

where $P_l^{(s)}$, $P_l^{(n)}$, $P_l^{(e^{(s)})}$, and $P_l^{(e^{(n)})}$ refer to the power of the desired speech, noise, feedback originating from the desired speech source, and feedback originating from the noise source, respectively. In addition, l = no-proc. refers to a system without processing inserted, and l = proc. refers to a system with processing inserted. The overall speech quality improvement is measured as the perceptual evaluation of speech quality (PESQ) improvement (ΔPESQ) as [19]

$$\Delta\text{PESQ} = \text{PESQ}(\hat{\mathbf{W}}(q,t)^{\top}\mathbf{y}[t], s_r[t]) - \text{PESQ}(m_{r,\text{no-proc}}[t], s_r[t]), \tag{36}$$

where the first argument is the signal under test and the second argument is the noise-free reference signal. Finally, the overall increase in stability due to the addition of $\hat{\mathbf{W}}(q,t)$ and $\hat{\mathbf{F}}(q,t)$ can be monitored through the added stable gain (ASG), with

$$\text{ASG [dB]} = -\text{MSG} - 20\log_{10}\left(\max_{\omega\in\mathcal{P}} |G(\omega)\hat{\mathbf{W}}(\omega)^{H}(\mathbf{F}(\omega) - \hat{\mathbf{F}}(\omega))|\right), \tag{37}$$

obtained by Fourier-transformation of the time-domain closed-loop filters. It should be noted that the set of frequencies for which the system can become unstable $\mathcal{P}$ in (32) and (37) may be different due to the insertion of the processing, and should be recalculated. To convert the STFT-domain NR-filters $\hat{\mathbf{w}}[f,k]$ to the time domain before computing the ASG, the distortion function equivalent of $\hat{\mathbf{w}}[f,k]$ is employed [8]. To focus on the steady-state performance, the ΔSNR, $\Delta\text{SFR}^{\text{s}}$, $\Delta\text{SFR}^{\text{n}}$, and SD metrics are calculated on the second half of the signals. The ASG is calculated using the final estimates $\hat{\mathbf{W}}(q,T)$ and $\hat{\mathbf{F}}(q,T)$ for a signal of length T, and is denoted as ASG^{f}. To give an indication of the convergence, the NESD (18) is assessed using both the final estimate $\hat{F}_i(q,T)$, denoted by NESD_i^{f}, and the average NESD obtained with $\hat{F}_i(q,t)$, $\forall t \in \{0,\cdots,T\}$, denoted by NESD_i^{m}.

6. Results And Discussion

The different combinations of AFC-filters and NR-filters are analysed systematically as follows. In Section 6.1, first, no AFC-filters are used ($\hat{\mathbf{F}}(q,t) = \mathbf{0}$), such that the NR-filters are tasked with cancelling both the feedback and the noise. Subsequently, in Section 6.2, only AFC-filters are employed, without NR-filters ($\hat{\mathbf{W}}(q,t) = \mathbf{t}_r$). While this does not allow to suppress the impinging noise signal $\mathbf{n}[t]$ as the AFC-filters can only cancel the feedback $\mathbf{F}(l[t])$, this does allow to study the properties of the AFC-filters. Section 6.3 combines the model-based AFC-filters with the NR-filters, while Section 6.4 combines the hybrid AFC-filters with the NR-filters. Finally, in Section 6.5, the effect of loudspeaker non-linearities, and finetuning is analysed, as well as the interpretability of the hybrid design. While the aim is to experimentally analyse different combinations, it is emphasised that this also encompasses previously unexplored combinations, such as the hybrid NR-filters with no prior AFC-filters and the combination of hybrid AFC-filters with hybrid NR-filters.

6.1. NR-filters Without AFC-filters

Fig. 4 shows the performance of the model-based (Section 3.2.1), hybrid (Section 3.2.2), and data-driven (Section 3.2.3) NR-filters $\hat{\mathbf{W}}(q,t)$, when no AFC-filters are used, i.e., $\hat{\mathbf{F}}(q,t) = \mathbf{0}$. The NR-filters are then tasked with cancelling both the feedback and the noise.

When using an oracle VAD, the model-based NR-filters (MWF_*) achieve both feedback cancellation, given the positive ASG^{f}, $\Delta\text{SFR}^{\text{s}}$ and $\Delta\text{SFR}^{\text{n}}$, and noise reduction, given the positive ΔSNR and SD close to zero. Despite the feedback originating from the desired speech being correlated with the desired speech, as detailed in Section 3.2.3, an SFR^{s} improvement is observed. This improvement can be attributed to the spatial separation of the loudspeaker and the desired speech source, which together with the rank-1 approximation selecting the highest SNR eigenmode as detailed in Section 3.2.1, seems to lead to this performance improvement. When using the WebRTC VAD the performance decreases compared to using the oracle VAD, but a performance improvement remains with respect to the system without any processing, such that the model-based NR-filters seem viable, even under imperfect VAD-knowledge.

Both the hybrid NR-filters (Corr_*) and data-driven NR-filters (Mask_*) achieve effective feedback cancellation, given the positive ASG^{f}, $\Delta\text{SFR}^{\text{s}}$ and $\Delta\text{SFR}^{\text{n}}$. Similarly, the hybrid and data-driven NR-filters also achieve effective

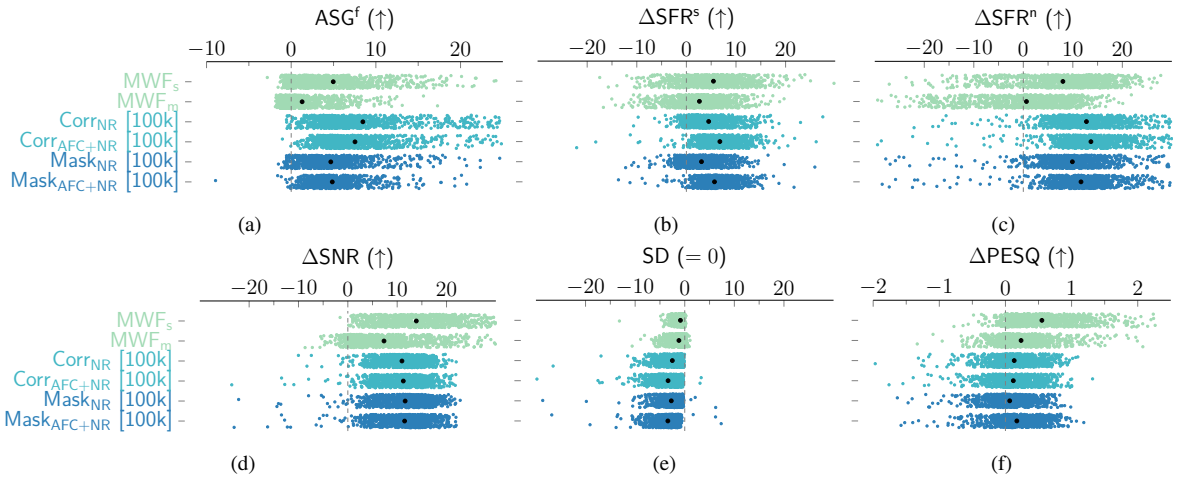

Figure 4: Performance comparison of the model-based (MWF$_*$), hybrid (Corr$_*$) and data-driven (Mask$_*$) NR-filters with $\hat{\mathbf{F}}(q,t) = \mathbf{0}$. The mean performance is represented with a black dot, and the individual datapoints are represented with coloured dots. The symbols (↑), (↓), and (= 0) indicate that higher values, lower values, or values closer to zero are desired.

noise reduction, given the positive ΔSNR and SD close to zero. Yet, the hybrid NR-filters achieve a higher feedback cancellation performance than both the model-based and data-driven NR-filters, as evidenced by a higher ASG^{f}, ΔSFR^{s}, and ΔSFR^{n}, despite the model-mismatch of the GEVD-MWF model structure and despite the data-driven NR-filters not requiring such a model structure, as detailed in Section 3.2.2. Possibly, this model-structure, although neglecting the correlation between the desired speech and the feedback thereof, aids the optimisation with respect to the data-driven NR-filters, and as the neural network is optimised for overall performance according to the MSE cost (20), the parameter estimation has adapted to this model-mismatch. This observation is consistent with [27, 39, 49], where hybrid algorithms are also found to outperform data-driven algorithms. The data-driven and hybrid NR-filters achieve a similar noise reduction performance, given the similar ΔSNR and SD scores.

As for the difference between only training for noise interferers ($\cdot_{\text{NR}}$) and training for joint noise and feedback interferers ($\cdot_{\text{AFC+NR}}$), the NR-filters only optimised for a noise reduction task can effectively cancel the noise and acoustic feedback. This is evidenced by a positive ASG^{f}, ΔSFR^{s} ΔSFR^{n} and ΔSNR, and an SD close to zero, thereby resulting in an overall positive ΔPESQ. Further optimising the neural networks for joint noise reduction and feedback cancellation following the procedure of Section 4 does, nevertheless, improve the feedback cancellation performance further, as also observed in [50]. Indeed, the ΔSFR^{s} and ΔSFR^{n} increase for both NR-filters after this optimisation for joint noise reduction and feedback cancellation, and similarly, the ASG^{f} increases for the data-driven NR-filters. The ASG^{f} for the hybrid NR-filters drops slightly as the MSE cost (20) aims at removing the feedback and noise, and thus mainly targets the ΔSFR^{s} and ΔSFR^{n}. The noise reduction performance is not altered by this optimisation for joint noise and feedback interferers.

Conclusion: Model-based, hybrid and data-driven NR-filters, without prior AFC-filters, prove effective for joint feedback cancellation and noise reduction. Yet, hybrid NR-filters achieve the highest improvement in the system's stability, with a mean ASG^{f} of 7.53 dB. Optimising for feedback and noise interferers during training improves performance over solely optimising for noise interferers during training.

6.2. AFC-filters Without NR-filters

Fig. 5 shows the performance of the model-based (Section 3.1.1) and hybrid (Section 3.1.2) AFC-filters without NR-filters, i.e., $\hat{\mathbf{W}}(q,t) = \mathbf{t}_r$. As the AFC-filters can only cancel $\mathbf{F}(l[t])$, and do not alter $\mathbf{n}[t]$, ΔSNR and SD are omitted. A continuous adaptive feedback canceller (CAF) is included as well, i.e, a standard adaptive filter that shares the update equation of the PEM-AFROW algorithm (10) but with no prewhitening filter, i.e., $\hat{A}_i(q,t) = 1$ [30, 37]. The inclusion of the CAF allows to confirm the benefit of this prewhitening.

As per [30, 37], the model-based AFC-filters (PEM-AFROW) converge to negative NESD values, corresponding to an increase in ASG^{f}, and outperform the CAF, highlighting the importance of the prewhitening, even despite

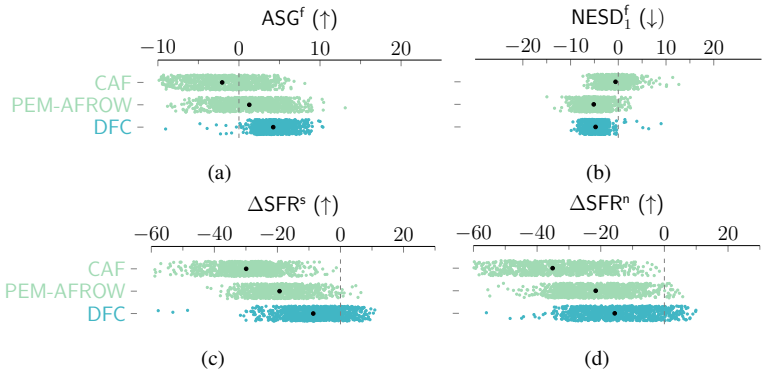

Figure 5: Performance comparison of the model-based (CAF and PEM-AFROW), and hybrid (DFC) AFC-filters with $\hat{\mathbf{W}}(q,t) = \mathbf{t}_r$. The mean performance is represented with a black dot, and the individual datapoints are represented with coloured dots. The symbols (↑) and (↓) indicate that higher values, or lower values are desired.

assumption (15b) not holding true exactly. Nevertheless, in agreement with [24], the hybrid AFC-filters (DFC) achieve higher performance than the model-based AFC-filters (PEM-AFROW) as characterised by a higher ASG^{f}.

Despite the increase in ASG^{f} for both the model-based and hybrid AFC-filters, a seemingly contradictory decrease in SFR^{s} and SFR^{n} is observed. For the model-based AFC-filters this can be explained as follows. As illustrated in Fig. 6, which shows the ground-truth and estimated feedback paths and prewhitening filters for one of the simulations using the PEM-AFROW algorithm, as well as the error $y[t]$ and its prewhitened equivalent $\hat{A}_1(q,t)y[t]$ used to update the AFC-filters, the prewhitening filter $\hat{A}_1(q,t)$, tends to have a highpass characteristic, particularly for speech inputs [37]. This reduces the influence of the lower frequency region in the error $\hat{A}_1(q,t)y[t]$ as used in (10). This, in turn, leads to a good fit of the estimated feedback path in the higher frequency region, where instability is most likely to occur due to the magnitude of the ground-truth feedback path, and hence to an improvement in stability as characterised by the ASG^{f}. Yet, the desired speech input, in particular, contains substantial energy in those lower frequency regions, such that the mismatch between estimated and ground-truth feedback path in the lower frequency regions leads to an increase in the power of the feedback of the desired speech and noise in the loop. Consequently, the model-based AFC-filters achieve an increase in stability, as characterised by the ASG^{f}, but may also increase the low-frequency feedback, as characterised by the $\Delta\text{SFR}^{\text{s}}$ and $\Delta\text{SFR}^{\text{n}}$. Only applying the AFC-filters in the loop for the higher frequency range prone to instability would resolve this issue.

A similar effect is observed for the hybrid AFC-filters, as visualised in Fig. 7, and can be clarified as follows. The NESD as defined in (18) minimises the power of the error between the estimated feedback path and the ground-truth feedback path, normalised with respect to the total power of the ground truth feedback path, thereby leading to a good approximation of the feedback path in the time-domain view, as illustrated in Fig. 7. Through Parseval's theorem, this error can be equivalently defined in the frequency domain, such that it can be seen that the NESD focuses on matching the estimated feedback path to the ground-truth feedback path in the high-power frequency regions. While in principle this is desired as these frequency regions are the primary source of instability, the error in the, often low-power, lower frequency regions receives reduced emphasis, potentially leading to an estimated feedback path with larger power, as illustrated in Fig. 7. However, again, due to the desired speech input, in particular, having the majority of its energy in the lower frequency regions, this large mismatch between estimated and ground-truth feedback path in the lower frequency regions leads to an increase in the power of the feedback of the desired speech and noise. A per-frequency normalisation, or only applying the AFC-filters in the higher frequency range prone to instability, would resolve this issue, possibly at the expense of the increase in stability.

Conclusion: The model-based and hybrid AFC-filters increase the system's stability, but may increase low-frequency feedback. A mean ASG^{f} of 1.28 dB is obtained for the model-based AFC-filters, and a mean of 4.23 dB for the hybrid AFC-filters.

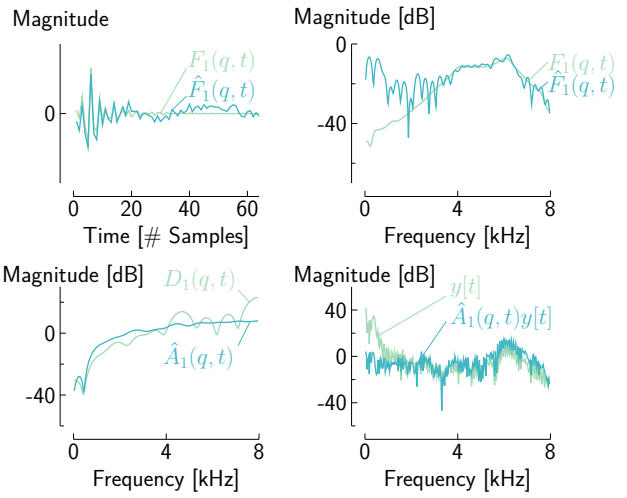

Figure 6: Time and frequency domain ground-truth ($F_1(q,t)$) and estimated ($\hat{F}_1(q,t)$) feedback path using model-based AFC-filters (PEM-AFROW) after processing all of the data. The $y[t]$ error signal and its prewhitened equivalent $\hat{A}_1(q,t)y[t]$ are shown, as well as the estimated prewhitening filter $\hat{A}_1(q,t)$ and the ground-truth prewhitening filter $D_1(q,t)$ estimated on the desired speech input $s[t]$.

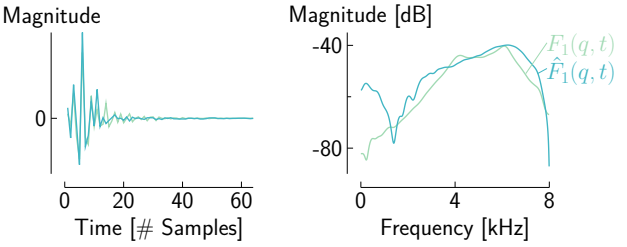

Figure 7: Time and frequency domain ground-truth ($F_1(q,t)$) and estimated ($\hat{F}_1(q,t)$) feedback path using hybrid AFC-filters (DFC), optimised with NESD cost, after processing all of the data.

6.3. Cascaded Model-based AFC-filters With NR-filters

Fig. 8 shows the performance of the combination of model-based AFC-filters (PEM-AFROW) with model-based (MWF$_*$), hybrid (Corr$_*$), and data-driven (Mask$_*$) NR-filters. The combination of the model-based AFC-filters with either design of NR-filters leads to an overall reduction in performance as indicated through a negative ΔPESQ (Fig. 8(i)). This performance reduction is caused by the inability of the model-based AFC-filters to convergence to the ground-truth feedback paths. Indeed, the AFC-filters only achieve negative NESD values in combination with the model-based NR-filters, and not in combination wih the hybrid and data-driven NR-filters. This is coupled to a decrease in ASG^{f} with respect to only using the NR-filters, and a larger spread in SD than in Section 6.1. These results are in contrast to the convergence of the AFC-filters without NR-filters as in Section 6.2. As the multi-microphone NR-filters influence the convergence of the AFC-filters through the closed loop, the adaptivity and frequency-selectivity of these NR-filters seems to prevent the AFC-filters from converging.

While the combination of model-based AFC-filters with model-based NR-filters can be viable as, e.g., in [31], it seems to require careful parameter tuning since the AFC- and NR-filters influence each other through the closed loop. Indeed, the parameter tuning as used in [31] has been tailored to each specific scenario considered there. In contrast, in Fig. 8, despite using commonly employed parameters for the AFC- and NR- filters, a performance reduction is observed. Similarly, the combination of model-based AFC-filters and data-driven NR-filters has been applied in [22, 46, 50], yet, the multi-microphone data-driven NR-filters seem to prevent reliable convergence of the model-based AFC-filters in the scenarios considered.

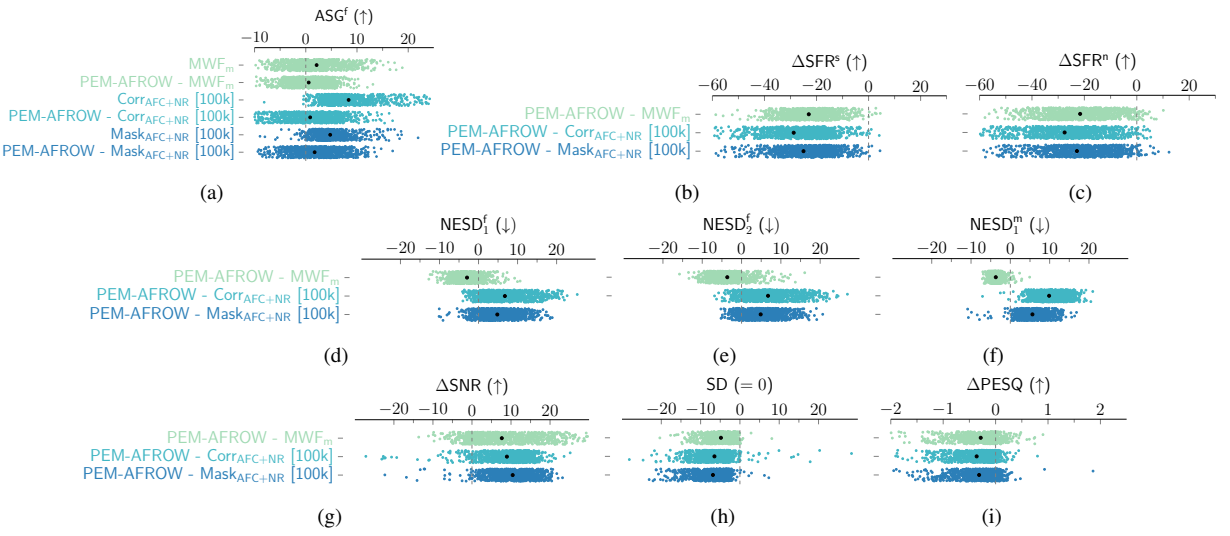

Figure 8: Performance comparison of the combination of model-based (PEM-AFROW) AFC-filters with model-based (MWF$_*$), hybrid (Corr$_*$), and data-driven (Mask$_*$) NR-filters. The mean performance is represented with a black dot, and the individual datapoints are represented with coloured dots. The symbols (↑), (↓), and (= 0) indicate that higher values, lower values, or values closer to zero are desired. The ASG^{f} with only the NR-filters as label represents the ASG^{f} when only inserting the NR-filters, i.e., allowing to characterise the benefit of the AFC-filters. The number in square brackets corresponds to the number of parameters in the neural network associated with the NR-filters.

Conclusion: The multi-microphone NR-filters seem to prevent reliable convergence of the model-based AFC-filters due to the influence of the NR-filters on the AFC-filters given the closed loop, thereby leading to a performance reduction with a mean ΔPESQ between −0.36 and −0.28.

6.4. Cascaded Hybrid AFC-filters With NR-filters

Fig. 9 shows the performance of the combination of hybrid AFC-filters (DFC) with model-based (MWF$_*$), hybrid (Corr$_*$), and data-driven (Mask$_*$) NR-filters. As opposed to the combination of model-based AFC-filters with the NR-filters (Section 6.3), the combination of hybrid AFC-filters with the NR-filters results in an overall performance improvement, as evidenced by the positive ΔPESQ. Indeed, the hybrid AFC-filters convergence to negative NESD values. This negative NESD is coupled to an increase in stability given the increase in ASG^{f} when inserting the AFC-filters. Given that in Fig. 9 the data points of the NESD$^1_{\text{f}}$, NESD$^2_{\text{f}}$, and NESD$^1_{\text{m}}$ are centered around a similar negative mean and have a similar spread, the feedback path estimates converge similarly for both microphones on a dataset level. Further, all types of NR-filters yield similar improvement in noise reduction, indicated by similar ΔSNR and SD.

Similar to Section 6.1, the combination of hybrid AFC-filters with hybrid NR-filters (DFC - Corr$_{\text{AFC+NR}}$ [100k]) outperforms the combination of hybrid AFC-filters with data-driven NR-filters (DFC - Mask$_{\text{AFC+NR}}$) and hybrid AFC-filters with model-based NR-filters (DFC - MWF$_{\text{m}}$) in terms of improved stability. As the NESD of the AFC-filters in the former combination is slightly higher, this increase in ASG^{f} can mainly be attributed to the improved stability due to the hybrid NR-filters as in Section 6.1.

As inserting the hybrid AFC-filters increases the total number of neural network parameters, Fig. 10 shows the ASG^{f} for data-driven NR-filters with no prior AFC-filters, the combination of data-driven NR-filters with hybrid AFC-filters, and the combination of hybrid NR-filters with hybrid AFC-filters, with an approximate equal total number of neural network parameters. To this end, the number of parameters for the neural network estimating the NR-filters is increased from 108 940 to 1 007 190 parameters when no hybrid AFC-filters are employed, as detailed in Section 5.2, to have around 1M neural network parameters in total for each method. The ASG^{f} is higher when employing the hybrid AFC-filters even though the neural network associated with the NR-filters employs fewer parameters. Thus, it seems beneficial ASG^{f}-wise to reallocate neural network parameters from the NR-filters to the AFC-filters. Additionally, the

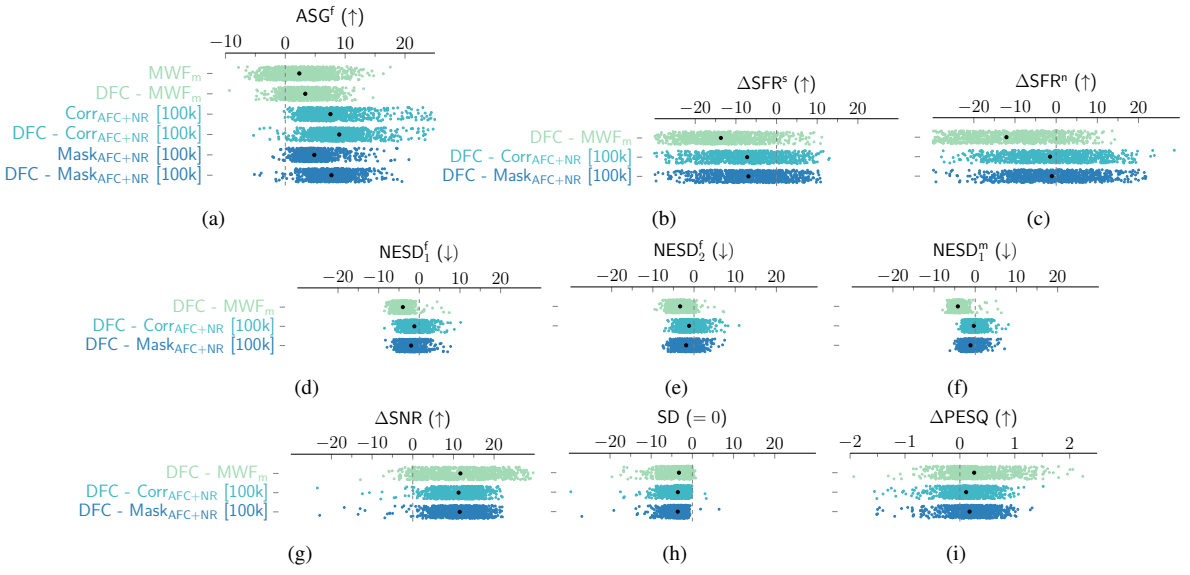

Figure 9: Performance comparison of the combination of hybrid (DFC) AFC-filters with model-based (MWF$_*$), hybrid (Corr$_*$), and data-driven (Mask$_*$) NR-filters. The mean performance is represented with a black dot, and the individual datapoints are represented with coloured dots. The symbols (↑), (↓), and (= 0) indicate that higher values, lower values, or values closer to zero are desired. The ASG^{f} with only the NR-filters as label represents the ASG^{f} when only inserting the NR-filters, i.e., allowing to characterise the benefit of the AFC-filters. The number in square brackets corresponds to the number of parameters in the neural network associated with the NR-filters.

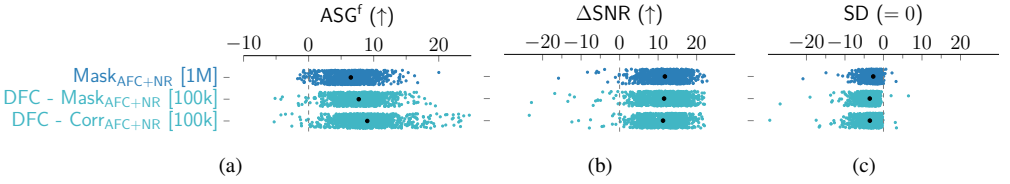

Figure 10: Performance comparison of the data-driven NR-filters (Mask$_*$) with no prior AFC-filters, and the combination of data-driven NR-filters or hybrid NR-filters (Corr$_*$) with prior hybrid AFC-filters (DFC). The total number of neural network parameters associated with the NR-filters is varied such that the total number of neural network parameters is fixed around 1M. The number in square brackets corresponds to the number of parameters in the neural network associated with the NR-filters.

noise reduction performance (ΔSNR and SD) does not alter substantially for the NR-filters with 1M or 100k parameters, potentially preceded by hybrid AFC-filters, even though this in principle increases the training-inference mismatch.

Conclusion: The hybrid AFC-filters increase the system's stability, such that its combination with either type of NR-filters result in a stability improvement between 1.47 dB and 2.86 dB ASG^{f}. The combination of hybrid AFC-filters with hybrid NR-filters generally outperforms the combinations of the hybrid AFC-filters with the data-driven and model-based NR-filters. For a fixed number of neural network parameters, using hybrid AFC-filters with hybrid or data-driven NR-filters with a smaller number of parameters seems beneficial over solely employing NR-filters with a larger number of parameters.

6.5. Exploratory Analysis

6.5.1. Loudspeaker Non-linearities

Fig. 11 shows the performance of a scenario with and without loudspeaker non-linearities, as governed by (33). Both the data-driven NR-filters (Mask$_{\text{AFC+NR}}$ [1M]), and the hybrid AFC-filters with hybrid NR-filters (DFC - Corr$_{\text{AFC+NR}}$

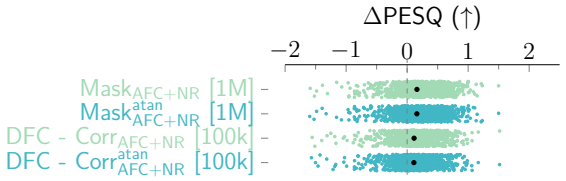

Figure 11: The ΔPESQ performance is not altered significantly for (mild) loudspeaker non-linearities, where $.^{\text{atan}}$ denotes the system with non-linearities.

[100k]) achieve similar performance across both scenarios, showing robustness to (mild) non-linearities, despite the hybrid AFC-filters being a priori optimised to approximate linear feedback paths.

Conclusion: The hybrid AFC-filters seem robust against (mild) loudspeaker non-linearities.

6.5.2. Finetuning

Fig. 12 shows the performance when finetuning the combination of hybrid AFC-filters with hybrid NR-filters (DFC - Corr$_{\text{AFC+NR}}$ [100k]).

In a first finetuning strategy (FIN$_1$), the closed-loop training signals generated with $\hat{\mathbf{W}}(q,t) = \mathbf{t}_r$ and $\hat{\mathbf{F}}(q,t) = \mathbf{0}$, are passed through both $\hat{\mathbf{F}}(q,t)$ and $\hat{\mathbf{W}}(q,t)$, such that a cost can be defined to jointly finetune both filters. Using the NESD as a cost function for updating the AFC-filters and the MSE for the NR-filters results in an improved ASG^{f} as the hybrid AFC-filters can optimise for a noisy multi-microphone scenario. This comes with a slight increase in ΔSFR^{s} and ΔSFR^{n} as the NR-filters can adjust for the low-frequency mismatch of the AFC-filters optimising the NESD (Section 6.2). Yet, only optimising the NR-filters with an MSE cost does not replicate this effect. Optimising both filters with an MSE cost further improves the ΔSFR^{s} and ΔSFR^{n} due to the increased emphasis on reducing the high-power low-frequency component of the feedback signal, at the expense of a decrease in ASG^{f}. The noise reduction performance (ΔSNR and SD) is not affected by the finetuning.

In a second finetuning strategy (FIN$_2$), the NR-filters are added in the closed loop system when generating the training signals, which allows to adapt the AFC-filters to the NR-filters. This leads to an increased stability as the ASG^{f} increases, and to an increase in ΔSFR^{s} and ΔSFR^{n}. Again, the noise reduction performance (ΔSNR and SD) is not affected by this finetuning.

Conclusion: Finetuning the neural networks of the NR-filters and AFC-filters can lead to a performance improvement, depending on the finetuning cost function and strategy.

6.5.3. Interpretability

The correlation matrices as estimated by a neural network for the hybrid NR-filters are solely constrained to be valid correlation matrices (i.e., Hermitian and positive semi-definite), and optimised on the training signals using the MSE cost (20). Yet, despite the mismatch of the model-structure (Section 3.2.1), the estimated desired speech power shows a remarkable correspondence with the ground-truth desired speech power after an initial convergence, as shown in Fig. 13. As noted in [27], possibly, to optimally make use of the, potentially mismatched, model structure, the estimated parameters are in correspondence with this structure. The stripe-artefacts, visible along the frequency axis of Fig. 13, likely arise due to the specific neural network architecture employed with transposed convolution operations, prone to so-called checkerboard artefacts [25].

Conclusion: The parameters of the hybrid NR-filters show correspondence with the desired speech spectrogram, possibly due to the effective usage of the model structure.

7. Conclusion

The performance of different combinations of cascaded AFC-filters, constructed following a model-based or hybrid design approach, and NR-filters, constructed following a model-based, hybrid, or data-driven design approach, was experimentally analysed in a general multi-microphone scenario for joint feedback cancellation and noise reduction

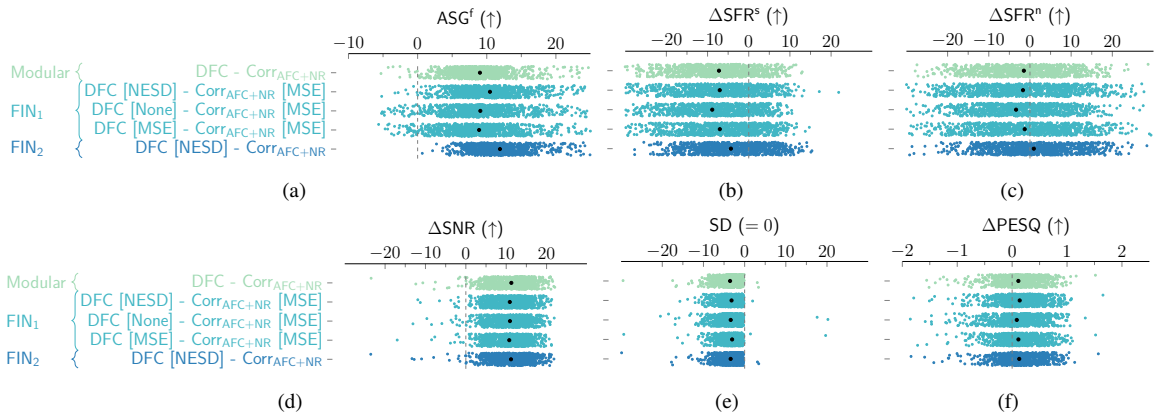

Figure 12: Performance comparison of finetuning strategies for combining AFC-filters with NR-filters. The upper point cloud represents the modular combination without finetuning (Modular). The middle three point clouds represent a first finetuning strategy (FIN_1), where the training signals that are generated with $\hat{\mathbf{W}}(q,t) = \mathbf{t}_r$ and $\hat{\mathbf{F}}(q,t) = \mathbf{0}$ are passed both through $\hat{\mathbf{F}}(q,t)$ and $\hat{\mathbf{W}}(q,t)$. The lower point cloud represents a second finetuning strategy (FIN_2) where the training signals are generated with presence of $\hat{\mathbf{W}}(q,t)$. The cost used to finetune each filter is enclosed in square brackets.

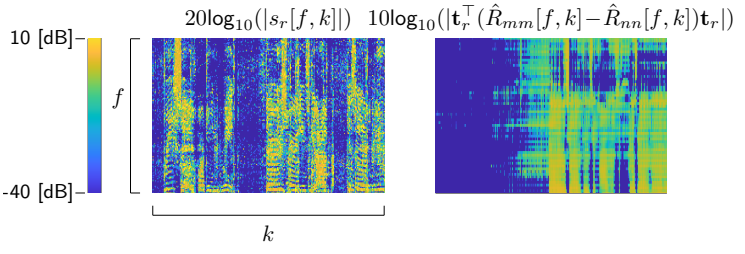

Figure 13: The estimated desired speech power (right), obtained by subtracting the correlation matrices estimated by the neural network associated with the hybrid NR-filters, shows remarkable correspondence with the spectrogram of the clean speech (left), after an initial converge time. Both variables are normalised with respect to the (estimated) microphone power.

with a general focus on the steady-state performance. Although the aim was not to propose novel AFC- and NR-filters, it did allow studying previously unexplored combinations, including the combination of hybrid AFC-filters with hybrid NR-filters, and hybrid NR-filters with no prior AFC-filters. Both the data-driven and hybrid NR-filters, possibly preceded by hybrid AFC-filters were shown to be effective combinations. Although the neural network associated with the hybrid NR-filters was optimised end-to-end on the training set, the estimated model-parameters showed a degree of similarity with the spectrogram of the clean speech. The DeepMFC training framework was adapted for multi-microphone scenarios, and strategies for joint finetuning of the neural networks associated with the AFC- and NR-filters were shown to be effective. Finally, an analysis framework was proposed to study the performance, and to study the effect of loudspeaker non-linearities.

CRediT authorship contribution statement

Arnout Roebben: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Software, Visualization, Writing — original draft. **Jesper Jensen:** Conceptualization, Methodology, Validation, Formal analysis, Supervision, Writing — review & editing. **Toon van Waterschoot:** Conceptualization, Methodology, Validation, Formal analysis, Supervision, Writing — review & editing. **Jan Wouters:** Conceptualization, Methodology, Validation, Formal analysis, Supervision, Writing — review & editing. **Marc Moonen:** Conceptualization, Methodology, Validation, Formal analysis, Supervision, Writing — review & editing.

References

[1] Avargel, Y., Cohen, I., 2007. System identification in the short-time fourier transform domain with crossband filtering. IEEE Trans. Audio Speech Lang. Process. 15, 1305–1319.

[2] Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H., 2000. Templates for the solution of algebraic eigenvalue problems. Software, Environments, and Tools. 1st ed., SIAM, Philadelphia, PA, U.S.A.

[3] Bastari, A., Squartini, S., Piazza, F., 2008. Joint acoustic feedback cancellation and noise reduction within the prediction error method framework, in: Proc. IEEE Hands-Free Speech Commun. Microphone Arrays (HSCMA), Trentino, Italy. pp. 228–231.

[4] Benesty, J., Chen, J., Huang, Y.A., Doclo, S., 2005. Study of the Wiener filter for noise reduction, in: Benesty, J., Makino, S., Chen, J. (Eds.), Speech enhancement. Springer, Berlin, Heidelberg. Signals and Communication Technology, pp. 9–41.

[5] Benesty, J., Makino, S., Chen, J., 2006. Speech enhancement. Signals and Communication Technology. 1st ed., Springer Berlin, Heidelberg, Berlin, Heidelberg.

[6] Dai, W., Li, X., Politis, A., Virtanen, T., 2024. Reference channel selection by multi-channel masking for end-to-end multi-channel speech enhancement, in: Proc. 2024 32nd Eur. Signal Process. Conf. (EUSIPCO 2024), Lyon, France. pp. 241–245.

[7] Denk, F., Kollmeier, B., 2021. The Hearpiece database of individual transfer functions of an openly available in-the-ear earpiece for hearing device research. Acta Acust. 5, 2.

[8] Didier, P., van Waterschoot, T., Doclo, S., Moonen, M., 2023. Sampling rate offset estimation and compensation for distributed adaptive node-specific signal estimation in wireless acoustic sensor networks. IEEE Open J. Signal Process. 4, 71–79.

[9] Dillon, H., 2012. Hearing Aids. 2nd ed., Thieme, Sydney, Australia.

[10] Doclo, S., Kellermann, W., Makino, S., Nordholm, S.E., 2015. Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. IEEE Signal Process. Mag. 32, 18–30.

[11] Durbin, J., 1960. The fitting of time-series models. Revue de l'Institut International de Statistique 28, 233–244.

[12] Google WebRTC Team, 2011–. Webrtc: Real-time communication for the web. https://webrtc.org/.

[13] Guo, M., 2013. Analysis, Design, and Evaluation of Acoustic Feedback Cancellation Systems for Hearing Aids – A Novel Approach to Unbiased Feedback Estimation. Ph.D. thesis. Aalborg University. Aalborg, Denmark.

[14] Haeb-Umbach, R., Nakatani, T., Delcroix, M., Boeddeker, C., Ochiai, T., 2024. Microphone array signal processing and deep learning for speech enhancement: Combining model-based and data-driven approaches to parameter estimation and filtering. IEEE Signal Process. Mag. 41, 12–23.

[15] Hänsler, E., Schmidt, G., 2006. Topics in acoustic echo and noise control. Signals and Communication Technology. 1st ed., Springer Berlin, Heidelberg, Berlin, Heidelberg, Germany.

[16] Hao, F., Moore, B.C., Zhang, H., Li, X., Zheng, C., 2025. L3C-DeepMFC: Low-latency low-complexity deep marginal feedback cancellation with closed-loop fine tuning for hearing aids, in: Proc. Interspeech 2025, Rotterdam, The Netherlands. pp. 838–842.

[17] Hellgren, J., Urban, F., 2001. Bias of feedback cancellation algorithms in hearing aids based on direct closed loop identification. IEEE Trans. Speech Audio Process. 9, 906–913.

[18] Hettler, F., Denk, F., Jürgens, T., Husstedt, H., 2024. Acoustic feedback tendency in hearing aids for different types and couplings in relation to insertion gain. Zenodo. doi:10.5281/zenodo.10835180.

[19] International Telecommunication Union, T.S.S.I.T., 2007. Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs. P.862.2.

[20] Kates, J.M., 2008. Digital hearing aids. 1st ed., Plural Publications, San Diego, California, USA.

[21] Lee, H.W., Jeon, M.Y., 2011. A combined feedback and noise cancellation algorithm for binaural hearing aids. Adv. Electr. Comput. Eng. 11, 35–40.

[22] Li, Z., Guo, H., Ge, X., Chen, K., Lu, J., 2026. Joint multichannel acoustic feedback cancellation and speaker extraction via Kalman filter and deep non-linear spatial filter, in: Proc. 2026 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Barcelona, Spain. pp. 15307–15311.

[23] Lombard, A., Reindl, K., Kellermann, W., 2009. Combination of adaptive feedback cancellation and binaural adaptive filtering in hearing aids. EURASIP J. Adv. Signal Process. 2009, 968345.

[24] Lydaki, E., Tan, Z.H., Jensen, J., Guo, M., 2025. Deep feedback cancellation for hearing aids with improved system stability and sound quality, in: Proc. 2025 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Hyderabad, India. pp. 1–5.

[25] Odena, A., Dumoulin, V., Olah, C., 2017. Deconvolution and checkerboard artifacts. Distill 1, e3. URL: http://distill.pub/2016/deconv-checkerboard.

[26] Roebben, A., 2025. Analysis of design approaches for sequential multi-microphone feedback cancellation and noise reduction in hearing instruments. GitHub repository. URL: https://cloud.esat.kuleuven.be/index.php/s/mECjdzA9rnHa4jY.

[27] Roebben, A., Van Peer, J., Jesper, J., van Waterschoot, T., Wouters, J., Moonen, M., 2025a. A hybrid design for noise reduction with the generalised eigenvalue decomposition multi-channel Wiener filter. Submitted to IEEE Trans. Audio Speech Lang. Process. URL: https://cloud.esat.kuleuven.be/index.php/s/6afRWzx4SLGpogN.

[28] Roebben, A., van Waterschoot, T., Wouters, J., Moonen, M., 2025b. Integrated minimum mean squared error algorithms for combined acoustic echo cancellation and noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. 34, 512–528.

[29] Rombouts, G., Spriet, A., Moonen, M., 2008. Generalized sidelobe canceller based combined acoustic feedback- and noise cancellation. Signal Process. 88, 571–581.

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

[31] Ruiz, S., van Waterschoot, T., Moonen, M., 2023. Cascade algorithms for combined acoustic feedback cancellation and noise reduction. EURASIP J. Audio Speech Music Process. 2023, 37.

[32] Sankowsky-Rothe, T., Blau, M., Schepker, H., Doclo, S., 2015. Reciprocal measurement of acoustic feedback paths in hearing aids. J. Acoust. Soc. Amer. 138, EL399–EL404.

[33] Seidel, E., Mowlaee, P., Fingscheidt, T., 2024. Convergence and performance analysis of classical, hybrid, and deep acoustic echo control. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 1–15.

[34] Serizel, R., Moonen, M., Van Dijk, B., Wouters, J., 2014. Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 785–799.

[35] Soleimani, B., Schepker, H., Mirbagheri, M., 2023. Neural-AFC: Learning-based step-size control for adaptive feedback cancellation with closed-loop model training, in: Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Rhodes Island, Greece. pp. 1–5.

[36] Spriet, A., 2004. Adaptive Filtering Techniques for Noise Reduction and Acoustic Feedback Cancellation in Hearing Aids. Ph.D. thesis. KU Leuven. Leuven, Belgium. Chapter 2.

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

[38] Spriet, A., Rombouts, G., Moonen, M., Wouters, J., 2007. Combined feedback and noise suppression in hearing aids. IEEE Trans. Audio Speech Lang. Process. 15, 1777–1790.

[39] Tammen, M., Doclo, S., 2023. Parameter estimation procedures for deep multi-frame MVDR filtering for single-microphone speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 3237–3248.

[40] Tan, K., Wang, D., 2018. A convolutional recurrent neural network for real-time speech enhancement, in: Proc. Interspeech 2018, Hyderabad, India. pp. 3229–3233.

[41] Thiemann, J., Ito, N., Vincent, E., 2013. The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings, in: Proc. Meet. Acoust., p. 035081.

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

[43] Wu, Y.K., Azcarreta, J., Patel, K., Xu, B., Lee, J.S., Lee, S., Pandey, A., 2025. A novel deep learning framework for efficient multichannel acoustic feedback control, in: Proc. Interspeech 2025, Rotterdam, The Netherlands. pp. 788–792.

[44] Yamagishi, J., Veaux, C., MacDonald, K., 2019. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. Dataset available at: https://doi.org/10.7488/ds/2645.

[45] Zhan, X., Moore, B.C.J., Li, X., Zheng, C., 2026. Deep learning-based joint optimization of adaptive feedback cancellation and residual feedback suppression for hearing aids, in: Proc. 2026 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Barcelona, Spain. pp. 15592–15596.

[46] Zhang, H., Yu, M., Wu, Y., Yu, T., Yu, D., 2023a. Hybrid AHS: A hybrid of Kalman filter and deep learning for Acoustic Howling Suppression, in: Proc. Interspeech 2023, Dublin, Ireland. pp. 834–838.

[47] Zhang, H., Yu, M., Yu, D., 2023b. Deep AHS: A deep learning approach to acoustic howling suppression, in: Proc. 2023 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Rhodes Island, Greece. pp. 1–5.

[48] Zhang, H., Zhang, Y., Yu, M., Yu, D., 2024. Advancing acoustic howling suppression through recursive training of neural networks, in: Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Seoul, Republic of Korea. pp. 711–715.

[49] Zhang, Y., Zhang, H., Yu, M., Yu, D., 2023c. Neural network augmented Kalman filter for robust acoustic howling suppression, in: Proc. Interspeech 2023, Dublin, Ireland. pp. 1715–1719.

[50] Zheng, C., Wang, M., Li, X., Moore, B.C.J., 2022. A deep learning solution to the marginal stability problems of acoustic feedback systems for hearing aids. J. Acoust. Soc. Amer. 152, 3616–3634.