KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

| | |
|---|---|
| **Citation/Reference** | Authors: Arnout Roebben, Toon van Waterschoot, and Marc Moonen<br>Title: CASCADED NOISE REDUCTION AND ACOUSTIC ECHO CANCELLATION BASED ON AN EXTENDED NOISE REDUCTION<br>Conference: Proc. 32nd European Signal Process. Conf. (EUSIPCO 2020)<br>Pages: 11-15<br>doi: 10.23919/EUSIPCO63174.2024.10714941 |
| **Archived version** | ☐ Submitted version<br>■ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher)<br>☐ Published<br>☐ Published (Open Access) |
| **Acknowledgements** | In this manuscript, the following funding is acknowledged:<br>☐ EU H22020 ☐ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ☐ Other |
| **Published version** | `https://ieeexplore.ieee.org/document/10714941` |
| **Author contact** | arnout.roebben@esat.kuleuven.be |



# CASCADED NOISE REDUCTION AND ACOUSTIC ECHO CANCELLATION BASED ON AN EXTENDED NOISE REDUCTION

Arnout Roebben, Toon van Waterschoot, and Marc Moonen

*Department of Electrical Engineering (ESAT)*
*STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics*

*KU Leuven*
Leuven, Belgium
{arnout.roebben, toon.vanwaterschoot, marc.moonen}@kuleuven.be

***Abstract*—In many speech recording applications, the recorded desired speech is corrupted by both noise and acoustic echo, such that combined noise reduction (NR) and acoustic echo cancellation (AEC) is called for. A common cascaded design corresponds to NR filters preceding AEC filters. These NR filters aim at reducing the near-end room noise (and possibly partially the echo) and operate on the microphones only, consequently requiring the AEC filters to model both the echo paths and the NR filters. In this paper, however, we propose a design with extended NR ($\text{NR}_{\text{ext}}$) filters preceding AEC filters under the assumption of the echo paths being additive maps, thus preserving the addition operation. Here, the $\text{NR}_{\text{ext}}$ filters aim at reducing both the near-end room noise and the far-end room noise component in the echo, and operate on both the microphones and loudspeakers. We show that the succeeding AEC filters remarkably become independent of the $\text{NR}_{\text{ext}}$ filters, such that the AEC filters are only required to model the echo paths, improving the AEC performance. Further, the degrees of freedom in the $\text{NR}_{\text{ext}}$ filters scale with the number of loudspeakers, which is not the case for the NR filters, resulting in an improved NR performance.**

***Index Terms*—Acoustic echo cancellation (AEC), Noise reduction (NR), Extended noise reduction ($\text{NR}_{\text{ext}}$), Multichannel**

## I. Introduction

In many speech recording applications, such as hands-free telephony in cars or hearing instruments, the recorded desired speech is corrupted by both noise and acoustic echo. This noise originates from within the room, the so-called near-end room, whereas the echo originates from loudspeakers playing signals recorded in another room, the so-called far-end room. To remove the noise and echo, combined noise reduction (NR) and acoustic echo cancellation (AEC) is called for.

Acoustic echo cancellation (AEC) aims at removing the echo while preserving the desired speech. AEC algorithms traditionally exploit the loudspeaker signals to compute an estimate of this echo, which can subsequently be subtracted from the microphone signals [1]. On the other hand, noise reduction (NR) aims at removing the near-end room noise while preserving the desired speech [2].

Algorithms for combined AEC and NR aim at jointly removing the echo and near-end room noise while preserving the desired speech [1], [3], [4]. Generally, a cascaded approach is adhered to, where either the AEC precedes the NR (AEC-NR) [3], [5], [6], or the NR precedes the AEC (NR-AEC) [4], [7], [8]. NR-AEC is beneficial computational complexity-wise, only requiring one AEC filter for each loudspeaker rather than for each microphone-loudspeaker pair. Further, the AEC filters in NR-AEC operate under reduced noise influence. However, as the microphone signals in NR-AEC are affected by the NR filters, the AEC filters do not solely model the echo paths between the loudspeakers and microphones, but rather the combination of echo paths and NR filters [4], [6], [9].

In this paper, to combat this drawback, we propose a design where an extended NR ($\text{NR}_{\text{ext}}$) precedes the AEC ($\text{NR}_{\text{ext}}$-AEC) under the assumption of the echo paths being additive maps, thus preserving the addition operation. Whereas the NR filters aim at reducing the near-end room noise (and possibly partially the echo) and operate on the microphones only, the $\text{NR}_{\text{ext}}$ filters aim at reducing both the near-end room noise and the far-end room noise component in the echo, and operate on both the microphones and loudspeakers. We show that the AEC filters remarkably become independent of the $\text{NR}_{\text{ext}}$ filters, such that the AEC filters are thus only required to model the echo paths, thereby improving the AEC performance. Additionally, the $\text{NR}_{\text{ext}}$ filters attain a better NR performance as the degrees of freedom in the $\text{NR}_{\text{ext}}$ filters scale with the number of loudspeakers, which is not the case for the NR filters. This improved performance is demonstrated by means of simulations, of which the code is available at [10].

A design similar to $\text{NR}_{\text{ext}}$-AEC has previously been proposed for the related, but distinct, problem of combined NR and acoustic feedback cancellation (AFC) [11]. However, to the best of our knowledge, it has not been proposed for

This research was carried out at the ESAT Laboratory of KU Leuven, in the frame of Research Council KU Leuven C14-21-0075 "A holistic approach to the design of integrated and distributed digital signal processing algorithms for audio and speech communication devices", and Aspirant Grant 11PDH24N (for A. Roebben) from the Research Foundation - Flanders (FWO).

combined NR and AEC, which in particular leads to the additional observation proven in this paper that the AEC filters become independent of the $\text{NR}_{\text{ext}}$ filters.

## II. Signal model

The signal model is presented in the $z$-domain to accommodate the duality between frequency- and time-domain. The $z$-to frequency-domain conversion is realised by replacing index $z$ with frequency-bin index $f$ (and possibly frame index $k$). The $z$- to time-domain conversion is realised by replacing the $z$-domain variables with time-lagged vectors.

Considering an $L$-loudspeaker/$M$-microphone setup, the microphone signals $m_m(z)$, $m \in \{1,\dots,M\}$, can be stacked into the microphone signal vector $\mathbf{m}(z) \in \mathbb{C}^{M\times 1}$ as

$$\mathbf{m}(z) = \begin{bmatrix} m_1(z) & m_2(z) & \dots & m_M(z) \end{bmatrix}^\top, \tag{1}$$

and decomposed into the desired speech signal vector $\mathbf{s}(z) \in \mathbb{C}^{M\times 1}$, the near-end room noise signal vector $\mathbf{n}(z) \in \mathbb{C}^{M\times 1}$, and the echo signal vector $\mathbf{e}(z) \in \mathbb{C}^{M\times 1}$ as

$$\mathbf{m}(z) = \mathbf{s}(z) + \mathbf{n}(z) + \mathbf{e}(z). \tag{2}$$

Similarly, the loudspeaker signals, $l_l(z)$, $l \in \{1,\dots,L\}$, can be stacked into the loudspeaker signal vector $\mathbf{l}(z) \in \mathbb{C}^{L\times 1}$ as

$$\mathbf{l}(z) = \begin{bmatrix} l_1(z) & l_2(z) & \dots & l_L(z) \end{bmatrix}^\top, \tag{3}$$

and split in a far-end room speech component $\mathbf{l}^s(z)$ and far-end room noise component $\mathbf{l}^n(z)$, i.e., $\mathbf{l}(z) = \mathbf{l}^s(z) + \mathbf{l}^n(z)$, thus modelling far-end room speech and noise respectively.

We assume the echo paths to be additive maps $F(.) : \mathbb{C}^{L\times 1} \to \mathbb{C}^{M\times 1}$, where $\mathbf{e}(z) = F(\mathbf{l}(z))$ can thus be decomposed into a speech component in the echo $\mathbf{e}^s(z) = F(\mathbf{l}^s(z))$ and a noise component in the echo $\mathbf{e}^n(z) = F(\mathbf{l}^n(z))$, i.e.,

$$\mathbf{e}(z) = F(\mathbf{l}^s(z)) + F(\mathbf{l}^n(z)) = \mathbf{e}^s(z) + \mathbf{e}^n(z). \tag{4}$$

The addition operation is thus preserved by the echo paths. An example of an echo path satisfying this additivity assumption is a linear echo path, possibly followed by sub-Nyquist sampling.

In addition, leveraging both the available microphone and loudspeaker signal vectors, an extended microphone signal vector $\tilde{\mathbf{m}}(z) \in \mathbb{C}^{(M+L)\times 1}$ can be defined as

$$\begin{aligned} \tilde{\mathbf{m}}(z) &= \begin{bmatrix} \mathbf{m}(z)^\top & \mathbf{l}(z)^\top \end{bmatrix}^\top && \text{(5a)}\\ &= \tilde{\mathbf{s}}(z) + \tilde{\mathbf{n}}(z) + \tilde{\mathbf{e}}^s(z) + \tilde{\mathbf{e}}^n(z) && \text{(5b)}\\ &= \begin{bmatrix} \mathbf{s}(z) \\ \mathbf{0}_{L\times 1} \end{bmatrix} + \begin{bmatrix} \mathbf{n}(z) \\ \mathbf{0}_{L\times 1} \end{bmatrix} + \begin{bmatrix} \mathbf{e}^s(z) \\ \mathbf{l}^s(z) \end{bmatrix} + \begin{bmatrix} \mathbf{e}^n(z) \\ \mathbf{l}^n(z) \end{bmatrix}. && \text{(5c)} \end{aligned}$$

We also make the following assumptions:

- $\mathbf{s}(z)$, $\mathbf{n}(z)$, $\mathbf{e}^s(z)$ and $\mathbf{e}^n(z)$ are uncorrelated, $\mathbf{l}^s(z)$ and $\mathbf{l}^n(z)$ are uncorrelated, and $\mathbf{l}^s(z)$ and $\mathbf{l}^n(z)$ are uncorrelated with $\mathbf{s}(z)$ and $\mathbf{n}(z)$.
- The microphone correlation matrix $R_{mm}(z) = \mathbb{E}\{\mathbf{m}(z)\mathbf{m}(z)^H\} = R_{ss}(z) + R_{nn}(z) + R_{ee}(z) \in \mathbb{C}^{M\times M}$ is of full rank, with $R_{ss}(z) = \mathbb{E}\{\mathbf{s}(z)\mathbf{s}(z)^H\}$, $R_{nn}(z) = \mathbb{E}\{\mathbf{n}(z)\mathbf{n}(z)^H\}$ and $R_{ee}(z) = \mathbb{E}\{\mathbf{e}(z)\mathbf{e}(z)^H\}$. Here, $R_{ee}(z)$ can be decomposed into $R_{e^se^s}(z) = \mathbb{E}\{\mathbf{e}^s(z)\mathbf{e}^s(z)^H\}$ and $R_{e^ne^n}(z) = \mathbb{E}\{\mathbf{e}^n(z)\mathbf{e}^n(z)^H\}$. The loudspeaker correlation matrix $R_{ll}(z) = \mathbb{E}\{\mathbf{l}(z)\mathbf{l}(z)^H\} = R_{l^sl^s}(z) + R_{l^nl^n}(z) \in \mathbb{C}^{L\times L}$ is of full rank, with $R_{l^sl^s}(z) = \mathbb{E}\{\mathbf{l}^s(z)\mathbf{l}^s(z)^H\}$ and $R_{l^nl^n}(z) = \mathbb{E}\{\mathbf{l}^n(z)\mathbf{l}^n(z)^H\}$. Similarly, $R_{l^sl^s}(z)$ is of full rank, e.g., due to multiple far-end room speech sources or due to non-linearities that are additive maps in the transmission from far-end to near-end room. The general case where the correlation matrices can be rank-deficient will not be considered here for the sake of conciseness but will be considered elsewhere.
- The mean squared error (MSE) optimal estimates of the echo paths between the loudspeakers and microphones are equal for the echo and the far-end room speech component in the echo, i.e., $R_{ll}(z)^{-1}R_{le}(z) = R_{l^sl^s}(z)^{-1}R_{l^se^s}(z)$. Here, $R_{le}(z) = \mathbb{E}\{\mathbf{l}(z)\mathbf{e}(z)^H\} \in \mathbb{C}^{L\times M}$, and $R_{l^se^s}(z) = \mathbb{E}\{\mathbf{l}^s(z)\mathbf{e}^s(z)^H\}$.

Similarly, the extended microphone correlation matrix can be defined as $R_{\tilde{m}\tilde{m}} = \mathbb{E}\{\tilde{\mathbf{m}}(z)\tilde{\mathbf{m}}(z)^H\} = R_{\tilde{s}\tilde{s}}(z) + R_{\tilde{n}\tilde{n}}(z) + R_{\tilde{e}\tilde{e}}(z) \in \mathbb{C}^{(M+L)\times(M+L)}$ with $R_{\tilde{s}\tilde{s}} = \mathbb{E}\{\tilde{\mathbf{s}}(z)\tilde{\mathbf{s}}(z)^H\}$, $R_{\tilde{n}\tilde{n}} = \mathbb{E}\{\tilde{\mathbf{n}}(z)\tilde{\mathbf{n}}(z)^H\}$, and $R_{\tilde{e}\tilde{e}} = \mathbb{E}\{\tilde{\mathbf{e}}(z)\tilde{\mathbf{e}}(z)^H\}$. Here, $R_{\tilde{e}\tilde{e}}(z)$ can be decomposed into $R_{\tilde{e}^s\tilde{e}^s}(z) = \mathbb{E}\{\tilde{\mathbf{e}}^s(z)\tilde{\mathbf{e}}^s(z)^H\}$ and $R_{\tilde{e}^n\tilde{e}^n}(z) = \mathbb{E}\{\tilde{\mathbf{e}}^n(z)\tilde{\mathbf{e}}^n(z)^H\}$.

While $\mathbf{n}(z)$ and $\mathbf{e}^n(z)$ can be assumed to be stationary always-on signal vectors, $\mathbf{s}(z)$ and $\mathbf{e}^s(z)$ are non-stationary on-off signal vectors. A voice activity detector (VAD) is assumed available to distinguish between these on-off periods. To this end, $\text{VAD}_s(z)$ differentiates between on-off periods in $\mathbf{s}(z)$, and $\text{VAD}_{e^s}(z)$ in $\mathbf{e}^s(z)$. The following regimes can be defined:

- $\text{VAD}_s(z) = 1$, $\text{VAD}_{e^s}(z) = 1$: $\mathbf{m}(z) = \mathbf{s}(z) + \mathbf{n}(z) + \mathbf{e}(z)$;
- $\text{VAD}_s(z) = 1$, $\text{VAD}_{e^s}(z) = 0$: $\mathbf{m}(z) = \mathbf{s}(z) + \mathbf{n}(z) + \mathbf{e}^n(z)$;
- $\text{VAD}_s(z) = 0$, $\text{VAD}_{e^s}(z) = 1$: $\mathbf{m}(z) = \mathbf{n}(z) + \mathbf{e}(z)$;
- $\text{VAD}_s(z) = 0$, $\text{VAD}_{e^s}(z) = 0$: $\mathbf{m}(z) = \mathbf{n}(z) + \mathbf{e}^n(z)$.

## III. Filters

Combined NR and AEC aims at designing a filter $\tilde{\mathbf{w}}_r(z) \in \mathbb{C}^{(M+L)\times 1}$ to estimate the desired speech in reference microphone $r \in \{1,\cdots,M\}$. As illustrated in Fig. 1(a), commonly, $\tilde{\mathbf{w}}_r(z)$ is designed by cascading NR filters to reduce the near-end room noise (and possibly partially the echo) in the microphones with AEC filters to reduce the echo. This design, referred to as NR-AEC, is discussed in Section III-A.

To combat dependence of the AEC filters on the NR filters in NR-AEC, we propose to precede the AEC filters with $\text{NR}_{\text{ext}}$ filters as illustrated in Fig. 1(b) and referred to as $\text{NR}_{\text{ext}}$-AEC. Contrary to NR-AEC, the AEC filters remarkably become independent of the $\text{NR}_{\text{ext}}$ filters as discussed in Section III-B.

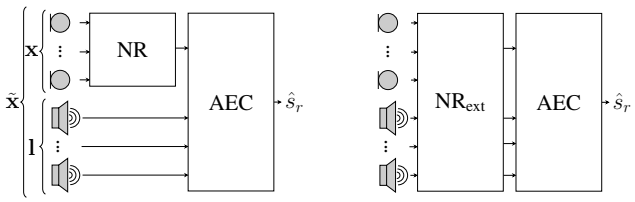
### A. NR first, AEC second (NR-AEC)

In NR-AEC, $\tilde{\mathbf{w}}_r(z)$ is designed to reconstruct the desired speech in the $r$th microphone as $\hat{s}_r(z) = \tilde{\mathbf{w}}_r(z)^H\tilde{\mathbf{m}}(z)$ with:

$$\tilde{\mathbf{w}}_r(z) = \underbrace{\begin{bmatrix} W_{NR}(z) & \mathbf{0}_{M\times L} \\ \mathbf{0}_{L\times M} & \mathbb{I}_{L\times L} \end{bmatrix}}_{NR} \underbrace{\begin{bmatrix} \mathbb{I}_{M\times M} \\ -W^\star_{AEC}(z) \end{bmatrix}}_{AEC} \tilde{\mathbf{t}}_r, \text{ with} \tag{6a}$$

$$W_{NR}(z) = R_{mm}(z)^{-1}R_{ss}(z) \tag{6b}$$

$$W^\star_{AEC}(z) = R_{ll}(z)^{-1}R_{le^\star}(z). \tag{6c}$$




(a) NR-AEC (b) $\mathrm{NR}_{\mathrm{ext}}$-AEC

Fig. 1. (a) The NR filters aim at reducing the near-end room noise (and possibly partially the echo), and the AEC filters aim at reducing the echo. (b) The $\mathrm{NR}_{\mathrm{ext}}$ filters aim at reducing both the near-end room noise and far-end room noise component in the echo, and the AEC filters aim at reducing the far-end room speech (and residual noise) component in the echo.

Here, $\mathbf{e}^{\star}(z)$ is defined as $W_{NR}(z)^H\mathbf{e}(z)$, i.e., as the echo signal vector after applying $W_{NR}(z)$, and $R_{le^{\star}}(z) = \mathbb{E}\{\mathbf{l}(z)\mathbf{e}^{\star}(z)^H\}$ corresponds to the loudspeaker-echo correlation matrix after applying $W_{NR}(z)$. $\tilde{\mathbf{t}}_r \in \mathbb{C}^{(M+L)\times 1}$ is a unit vector with 1 at position $r$ and 0 elsewhere.

$W_{NR}(z)$ is a multichannel Wiener filter (MWF) that operates on the microphones, aimed at suppressing the near-end room noise (and echo) while preserving the desired speech [12, Section 2.4]. $W^{\star}_{AEC}(z)$ corresponds to an MSE optimal estimate of the echo paths after applying $W_{NR}(z)$, by means of which the echo is suppressed [13, Section 2.1]. As the AEC filters are applied after the NR filters, the AEC filters here do not solely model the echo paths, but rather the combination of the echo paths and the NR filters, i.e., $R_{ll}(z)^{-1}R_{le^{\star}}(z) \neq R_{ll}(z)^{-1}R_{le}(z)$.

In [14], (6) has been considered in an $L$-loudspeaker/1-microphone setup by assuming linear echo paths and a sufficiently long $\tilde{\mathbf{w}}_r(z)$ to model the echo paths. Under these conditions, (6) was found optimal in MSE sense, except that $R_{mm}(z)$ should be replaced with $R_{ss}(z)+R_{nn}(z)$. $W_{NR}(z)$ is then a true NR filter only aimed at reducing the near-end room noise [14]. Nevertheless, $R_{mm}(z)$ is used both in [14] and in this paper, as $R_{ss}(z)+R_{nn}(z)$ cannot be estimated due to $\mathbf{e}(z)$ being always-on. A similar reasoning can be applied to an $L$-loudspeaker/$M$-microphone scenario, with undermodelling of the echo paths, but is omitted here for conciseness.

### B. $\mathrm{NR}_{ext}$ first, AEC second ($\mathrm{NR}_{ext}$-AEC)

In $\mathrm{NR}_{\mathrm{ext}}$-AEC, $\tilde{\mathbf{w}}_r(z)$ is designed to reconstruct the desired speech in the $r$th microphone as $\hat{s}_r(z) = \tilde{\mathbf{w}}_r(z)^H\tilde{\mathbf{m}}(z)$ with:

$$\tilde{\mathbf{w}}_r(z) = \underbrace{W_{NR_{ext}}(z)}_{NR_{ext}} \underbrace{\begin{bmatrix}\mathbb{I}_{M\times M}\\ -W^{\star\star}_{AEC}(z)\end{bmatrix}}_{AEC}\tilde{\mathbf{t}}_r, \text{ with} \tag{7a}$$

$$W_{NR_{ext}}(z) = R_{\tilde{m}\tilde{m}}(z)^{-1}\left(R_{\tilde{s}\tilde{s}}(z) + R_{\tilde{e}^s\tilde{e}^s}(z)\right) \tag{7b}$$

$$W^{\star\star}_{AEC}(z) = R_{l^{\star\star}l^{\star\star}}(z)^{-1}R_{l^{\star\star}m^{\star\star}}(z). \tag{7c}$$

Here, $\mathbf{m}^{\star\star}(z)$ and $\mathbf{l}^{\star\star}(z)$ are defined as $\begin{bmatrix}\mathbf{m}^{\star\star}(z)^\top & \mathbf{l}^{\star\star}(z)^\top\end{bmatrix}^\top = W_{NR_{ext}}(z)^H\tilde{\mathbf{m}}(z)$, i.e., as the microphone and loudspeaker signal vectors after applying $W_{NR_{ext}}(z)$, and $R_{l^{\star\star}l^{\star\star}}(z) = \mathbb{E}\{\mathbf{l}^{\star\star}(z)\mathbf{l}^{\star\star}(z)^H\}$ and $R_{l^{\star\star}m^{\star\star}}(z) = \mathbb{E}\{\mathbf{l}^{\star\star}(z)\mathbf{m}^{\star\star}(z)^H\}$.

$W_{NR_{ext}}(z)$ is an MWF that operates on the extended signal model, aimed at suppressing the near-end room noise and far-end room noise component in the echo while preserving the desired speech and far-end room speech component in the echo. $W^{\star\star}_{AEC}(z)$ corresponds to an MSE optimal estimate of the echo paths between the loudspeakers and microphones, by means of which the far-end room speech component in the echo and the residual far-end room noise component in the echo are suppressed. Although the AEC filters are computed after the $\mathrm{NR}_{\mathrm{ext}}$ filters, contrary to NR-AEC, the AEC filters are not affected by the $\mathrm{NR}_{\mathrm{ext}}$ filters. Indeed, rewriting $W_{NR_{ext}}(z)$ using the $2\times 2$ block inverse formula [15, (2.3)], $R_{l^{\star\star}l^{\star\star}}(z)$ and $R_{l^{\star\star}m^{\star\star}}(z)$ are given as:

$$R_{l^{\star\star}l^{\star\star}}(z) = R_{l^sl^s}(z)R_{ll}(z)^{-1}R_{l^sl^s}(z) \tag{8a}$$

$$R_{l^{\star\star}m^{\star\star}}(z) = R_{l^sl^s}(z)R_{ll}(z)^{-1}R_{l^se^s}(z). \tag{8b}$$

Using (8), $W^{\star\star}_{AEC}(z)$ corresponds to:

$$W^{\star\star}_{AEC}(z) = R_{l^sl^s}(z)^{-1}R_{l^se^s}(z), \tag{9}$$

which is indeed the MSE optimal estimate of the echo paths, and so remarkably independent of the $\mathrm{NR}_{\mathrm{ext}}$ filters. The optimality of (7) will be discussed elsewhere.

## IV. Experimental procedures

### A. Acoustic scenarios

In a $5\,\mathrm{m}\times 5\,\mathrm{m}\times 3\,\mathrm{m}$ room with a reflection coefficient of $0.15$, source-to-mic and loudspeaker-to-mic impulse responses of length $128$ samples at a sampling rate of $16\,\mathrm{kHz}$ are generated using the randomised image method (RIM) with randomised distances of $0.13\,\mathrm{m}$ [16]. One desired speech source, one near-end room noise source and $L = 2$ loudspeakers are placed at congruent angles in a circle with a $0.2\,\mathrm{m}$ radius around the mean microphone position. $M = 2$ microphones are positioned at $\begin{bmatrix}2 & 1.9 & 1\end{bmatrix}\mathrm{m}$ and $\begin{bmatrix}2 & 1.8 & 1\end{bmatrix}\mathrm{m}$. Five scenarios with varying desired speech source, near-end room noise source and loudspeaker positions are examined.

Sentences of the hearing in noise test (HINT) database, concatenated with $5\,\mathrm{s}$ of silence, are used as a desired speech source and far-end room speech component in the loudspeakers [17]. Babble noise is used as the near-end room noise source to model competing talkers [18], and white noise as the far-end room component in the loudspeakers to model, e.g., sensor and far-end room noise. All signals are $10\,\mathrm{s}$ long.

The power ratio between the echo signals is set to $0\,\mathrm{dB}$. The power ratio between $\mathbf{l}^s(z)$ and $\mathbf{l}^n(z)$ equals $0\,\mathrm{dB}$. The input signal-to-noise ratio ($\mathrm{SNR}^{\mathrm{in}}$) and signal-to-echo ($\mathrm{SER}^{\mathrm{in}}$) ratio in microphone $r$ are varied between $-15\,\mathrm{dB}$ and $15\,\mathrm{dB}$.

### B. Algorithm settings

The filters $W_{NR}(z)$ and $W_{NR_{ext}}(z)$ are implemented in the short-time Fourier transform- (STFT) domain as frequency-domain filters attain larger SNR improvements than time-domain filters [12, Chapter 2]. To this end, a squared root Hann window of size $512$ samples with a window shift of $256$ samples is used. The correlation matrices are estimated

by averaging across frames. As one desired speech source is considered, the rank of the desired speech correlation matrix at frequency-bin $f$, $R_{ss}(f)$, is enforced to equal one by using a generalised eigenvalue decomposition (GEVD) [2]. Similarly, as one desired speech source and two independent loudspeaker sources are considered, the rank of $R_{\tilde{s}\tilde{s}}(f) + R_{\tilde{e}^s\tilde{e}^s}(f)$ is enforced to equal three by using a GEVD. The NR and $\text{NR}_{\text{ext}}$ filters are nevertheless applied in the time-domain by converting the STFT-domain filters to an equivalent distortion function of $2N-1$ samples [19], as the AEC filters need to model the NR filters and frequency-domain filters cannot be modelled exactly by the AEC filters [20].

The AEC filters $W^{\star}_{AEC}(z)$ and $W^{\star\star}_{AEC}(z)$ are implemented in the time-domain as STFT-domain filters cannot model the echo path impulse responses exactly [20]. To this end, a normalised least mean squares (NLMS) implementation is used with step size $0.1$ and regularisation $10^{-6}$ [13, Section 2.2]. To reduce excess error in the NLMS updates, the AEC filters are updated only when the desired speech is inactive.

For NR-AEC, the number of coefficients in the AEC filter between the $r$th microphone and each loudspeaker $L_{\hat{F}}$ is varied between the length of the echo path impulse responses (128) and the length of the convolution of the NR filters and the echo path impulse responses ($(2{\cdot}512{-}1){+}128{-}1 = 1150$). For $\text{NR}_{\text{ext}}$-AEC, the coefficients exceeding coefficient index 128 are explicitly set to 0 as the AEC filters only aim at modelling the echo path impulse responses.

First, to study the converged filters, the correlation matrices in the NR and $\text{NR}_{\text{ext}}$ filters are calculated across the entire data, and the final NLMS updated AEC filters are used for the entire data. Second, the filters are adapted through time. For the NR and $\text{NR}_{\text{ext}}$ filters, to this end, exponential smoothing with a weight for the previous estimate of $0.995$ is used. Ideal VADs are assumed to isolate the influence of VAD errors, but practical implementations can be found, e.g., in [21].

### C. Performance measures

Performance is evaluated using the intelligibility-weighted SNR improvement $\Delta\text{SNR}^{\text{I}}$, the intelligibility-weighted SER improvement $\Delta\text{SER}^{\text{I}}$ and the intelligibility-weighted speech distortion (SD) $\text{SD}^{\text{I}}$ [12], [22].

## V. Results and Discussion

Fig. 2 and Fig. 3 show the NR and AEC performance of NR-AEC and $\text{NR}_{\text{ext}}$-AEC as a function of $L_{\hat{F}}$ when using the converged filters for the entire data. As $L_{\hat{F}}$ does not influence the NR and $\text{NR}_{\text{ext}}$ filters, Fig. 2 is independent of $L_{\hat{F}}$.

Regarding the NR performance, when $\text{SER}^{\text{in}}$ is lower than or comparable to $\text{SNR}^{\text{in}}$, $\text{NR}_{\text{ext}}$-AEC attains a better NR performance than NR-AEC. Indeed, the NR can be interpreted as an MWF aiming at cancelling the one near-end room noise and two echo sources. However, as there are two microphones, the NR filters only have one degree of freedom, which is insufficient to cancel the two dominant echo signals, resulting in a low $\text{SNR}^{\text{I}}$ improvement and high $\text{SD}^{\text{I}}$ degradation. $\text{NR}_{\text{ext}}$ does not suffer from this phenomenon, as the degrees of freedom

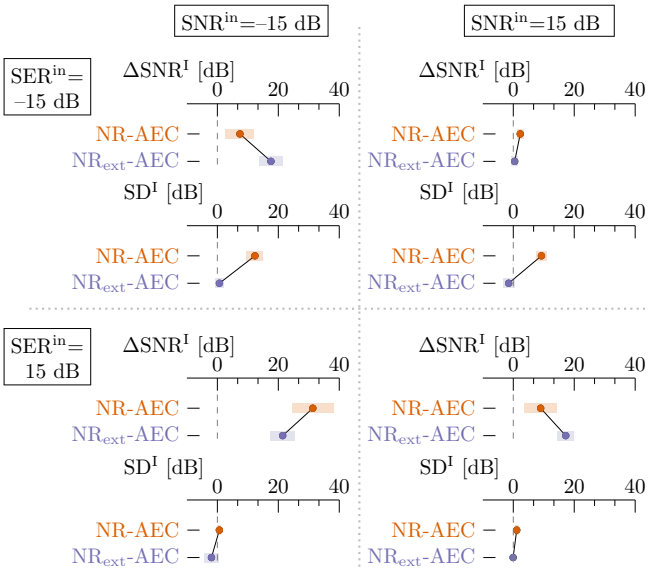



Fig. 2. NR performance using the converged filters for the entire data. Dots show mean performance and shading the standard deviation. At low $\text{SER}^{\text{in}}$, $\text{NR}_{\text{ext}}$-AEC has better NR performance as the $\text{NR}_{\text{ext}}$ filters scale with the number of loudspeakers opposed to the NR filters. Only at high $\text{SER}^{\text{in}}$ and low $\text{SNR}^{\text{in}}$, NR-AEC has better NR performance as the NR filters use a lower rank-approximation than the $\text{NR}_{\text{ext}}$ filters, limiting the noise from each mode.

in the $\text{NR}_{\text{ext}}$ filters scale with the number of loudspeakers, thus attaining a larger $\text{SNR}^{\text{I}}$ improvement and lower $\text{SD}^{\text{I}}$ degradation. Only at high $\text{SER}^{\text{in}}$ combined with low $\text{SNR}^{\text{in}}$, better NR performance is achieved in NR-AEC. As $\text{NR}_{\text{ext}}$ uses a larger rank-approximation in the GEVD (3 rather than 1), noise along an increased number of modes in the GEVD is retained in the $\text{NR}_{\text{ext}}$ filters than in the NR filters.

$\text{NR}_{\text{ext}}$-AEC also attains a better AEC performance than NR-AEC at lower $L_{\hat{F}}$ in Fig. 3. The AEC filters in $\text{NR}_{\text{ext}}$-AEC only need to model the 128 coefficient long echo path impulse responses, while the AEC filters in NR-AEC need to model the 1150 coefficients in the convolution of the echo path impulse responses and the NR filters. Only at high $\text{SER}^{\text{in}}$ combined with low $\text{SNR}^{\text{in}}$ the top AEC performance of NR-AEC exceeds that of $\text{NR}_{\text{ext}}$-AEC as the AEC filters in NR-AEC are less perturbed by the noise; and as imperfect $\text{NR}_{\text{ext}}$ filters do influence the AEC filters to some extent, but these $\text{NR}_{\text{ext}}$ filters cannot be trivially modelled by the AEC filters as the $\text{NR}_{\text{ext}}$ filters also alter the loudspeaker signals. However, when considering adaptive filters (Fig. 4), this advantage of NR-AEC is ineffective, as the AEC filters then need to track the adaptivity of the NR filters, unlike in $\text{NR}_{\text{ext}}$-AEC.

## VI. Conclusion

We have introduced a cascaded design for combined NR and AEC under the assumption of the echo paths being additive maps, thus preserving the addition operation. To this end, we have proposed an $\text{NR}_{\text{ext}}$ preceding an AEC ($\text{NR}_{\text{ext}}$-AEC), and compared this $\text{NR}_{\text{ext}}$-AEC to the traditional design with an NR preceding an AEC (NR-AEC). Whereas NR filters aim at reducing the near-end room noise (and possibly partially

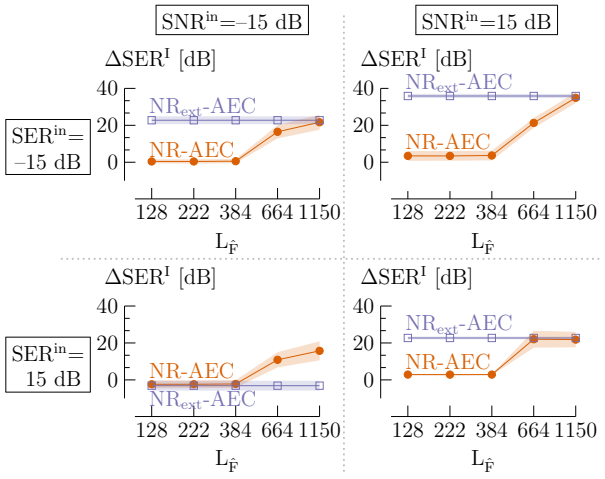



Fig. 3. AEC performance in function of $L_{\hat{F}}$ using the converged filters for the entire data. Dots show mean performance and shading the standard deviation. As the AEC filters in $\text{NR}_{\text{ext}}$-AEC are independent of the $\text{NR}_{\text{ext}}$ filters, as opposed to NR-AEC, $\text{NR}_{\text{ext}}$-AEC performance exceeds NR-AEC. Only at high $\text{SER}^{\text{in}}$ and low $\text{SNR}^{\text{in}}$ the top performance is higher in NR-AEC, yet this advantage is lost with adaptive filters (Fig. 4).

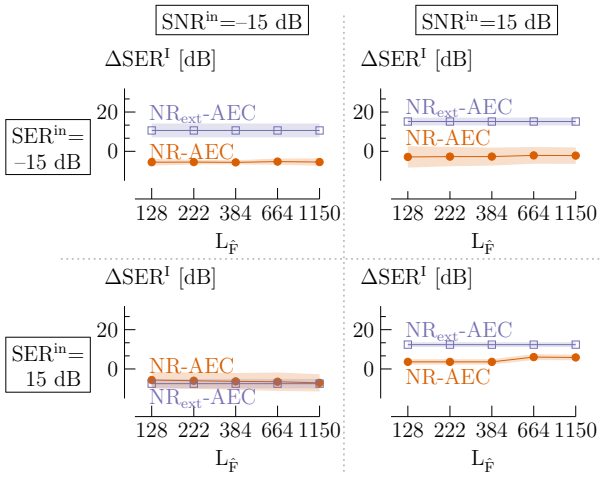



Fig. 4. AEC performance in function of $L_{\hat{F}}$, when adapting the filters through time. NR-AEC shows decreased performance compared to Fig. 3 as the AEC filters in NR-AEC need to track the adaptivity of the NR filters.

the echo) and operate on microphones only, the $\text{NR}_{\text{ext}}$ filters aim at reducing both the near-end room noise and the far-end room noise component in the echo, and operate on both the microphones and the loudspeakers. $\text{NR}_{\text{ext}}$-AEC outperforms NR-AEC in terms of both the NR and AEC performance, as the AEC filters in $\text{NR}_{\text{ext}}$-AEC remarkably become independent of the $\text{NR}_{\text{ext}}$ filters whereas the AEC filters in NR-AEC are dependent on the NR filters, and as the $\text{NR}_{\text{ext}}$ filters scale with the number of loudspeakers while the NR filters do not.

## References


[1] E. Hänsler and G. Schmidt, *Topics in Acoustic Echo and Noise Control*. Berlin, Heidelberg: Springer-Verlag, 2006.

[2] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[3] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21–32, Jan. 1998.

[4] S. Doclo, M. Moonen, and E. de Clippel, "Combined acoustic echo and noise reduction using GSVD-based optimal filtering," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, Jun. 2000, pp. II1061–II1064 vol.2.

[5] A. Cohen, A. Barnov, S. Markovich-Golan, and P. Kroon, "Joint Beamforming and Echo Cancellation Combining QRD Based Multichannel AEC and MVDR for Reducing Noise and Non-Linear Echo," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 6–10.

[6] M. Luis Valero and E. A. P. Habets, "Low-Complexity Multi-Microphone Acoustic Echo Control in the Short-Time Fourier Transform Domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 595–609, Mar. 2019.

[7] R. Martin, S. Gustafsson, and M. Moser, "Acoustic echo cancellation for microphone arrays using switched coefficient vectors," in *1997 International Workshop on Acoustic Echo and Noise Control (IWAENC)*,, London, United Kingdom, 1997, pp. I–85.

[8] M. Schrammen, A. Bohlender, S. Kühl, and P. Jax, "Change Prediction for Low Complexity Combined Beamforming and Acoustic Echo Cancellation," in *2019 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019, pp. 1–5.

[9] G. Reuven, S. Gannot, and I. Cohen, "Joint acoustic echo cancellation and transfer function GSC in the frequency domain," in *23rd IEEE Convention of Electrical and Electronics Engineers in Israel*, Tel-Aviv, Israel, Sep. 2004, pp. 412–415.

[10] A. Roebben, "Github repository: Cascaded noise reduction and acoustic echo cancellation based on an extended noise reduction," https://github.com/Arnout-Roebben/NRAEC_vs_NRextAEC, 2024.

[11] S. Ruiz, T. van Waterschoot, and M. Moonen, "Cascade Algorithms for Combined Acoustic Feedback Cancellation and Noise Reduction," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, p. 37, Sep. 2023.

[12] A. Spriet, "Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2004.

[13] G. Rombouts, "Adaptive filtering algorithms for acoustic echo and noise cancellation," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2003.

[14] S. Doclo, "Multi-microphone noise reduction and dereverberation techniques for speech applications," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2003.

[15] T.-T. Lu and S.-H. Shiou, "Inverses of 2 × 2 block matrices," *Computers & Mathematics with Applications*, vol. 43, no. 1-2, pp. 119–129, Jan. 2002.

[16] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the Modeling of Rectangular Geometries in Room Acoustic Simulations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 774–786, Apr. 2015.

[17] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, Feb. 1994.

[18] Auditec, "Auditory Tests (Revised), Compact Disc, Auditec," St. Louis, MO, 1997.

[19] P. Didier, T. van Waterschoot, S. Doclo, and M. Moonen, "Sampling Rate Offset Estimation and Compensation for Distributed Adaptive Node-Specific Signal Estimation in Wireless Acoustic Sensor Networks," *IEEE Open Journal of Signal Processing*, vol. 4, pp. 71–79, 2023.

[20] Y. Avargel and I. Cohen, "System Identification in the Short-Time Fourier Transform Domain With Crossband Filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[21] Z. Zhu, L. Zhang, K. Pei, and S. Chen, "A robust and lightweight voice activity detection algorithm for speech enhancement at low signal-to-noise ratio," *Digital Signal Processing*, vol. 141, p. 104151, Sep. 2023.

[22] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.