This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

Integrated Minimum Mean Squared Error Algorithms for Combined Acoustic Echo Cancellation and Noise Reduction

Arnout Roebben, Toon van Waterschoot, *Member, IEEE*, Jan Wouters, and Marc Moonen, *Fellow, IEEE*

***Abstract*—In many speech recording applications, noise and acoustic echo corrupt the desired speech. Consequently, combined noise reduction (NR) and acoustic echo cancellation (AEC) is required. Generally, a cascade approach is followed, i.e., the AEC and NR are designed in isolation by selecting a separate signal model, separate cost function, and separate solution strategy. The AEC and NR are then cascaded one after the other, not accounting for their interaction. In this paper, an integrated approach is proposed to consider this interaction in a general multi-microphone/multi-loudspeaker setup. Therefore, a single signal model of either the microphone signal vector or the extended signal vector, obtained by stacking microphone and loudspeaker signals, is selected, a single mean squared error cost function is formulated, and a common solution strategy is used. Using this microphone signal model, a multi-channel Wiener filter (MWF) is derived. Using the extended signal model, it is shown that an extended MWF (MWF_{ext}) can be derived, and several equivalent expressions can be found, which are nevertheless shown to be interpretable as cascade algorithms. Specifically, the MWF_{ext} is shown to be equivalent to algorithms where the AEC precedes the NR (AEC-NR), the NR precedes the AEC (NR-AEC), and the extended NR (NR_{ext}) precedes the AEC and post-filter (PF) (NR_{ext}-AEC-PF). Under rank-deficiency conditions the MWF_{ext} is non-unique. Equivalence then amounts to the expressions being specific, not necessarily minimum-norm solutions, for this MWF_{ext}. The practical performances differ due to non-stationarities and imperfect correlation matrix estimation, with the AEC-NR and NR_{ext}-AEC-PF attaining best overall performance.**

***Index Terms*—Integrated algorithm design, Audio signal processing, Multi-channel, Acoustic echo cancellation (AEC), Noise reduction (NR), Multi-channel Wiener filter (MWF)**

I. Introduction

In many speech recording applications, e.g., hands-free telephony in a car or telecommunication using hearing instruments, noise and acoustic echo corrupt the desired speech as illustrated in Fig. 1 [1]. The noise originates from within the room (the so-called near-end room), while the echo originates from loudspeakers playing signals recorded in another room (the so-called far-end room). To suppress the near-end room noise and echo, noise reduction (NR) and acoustic echo cancellation (AEC) algorithms are required.

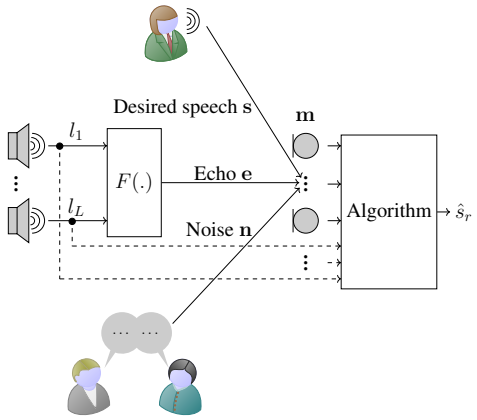


Fig. 1. Algorithms for combined acoustic echo cancellation (AEC) and noise reduction (NR) aim at providing an estimate of the desired speech $\hat{s}_r$ in reference microphone r by suppressing the near-end room noise signal vector $\mathbf{n}$ and echo signal vector $\mathbf{e}$, originating from the loudspeaker signals l_j, $j \in \{1, \cdots, L\}$, by means of the echo path map $F(.)$. To this end, the microphone signal vector $\mathbf{m}$ and possibly the loudspeaker signals l_j are utilised. The figurines have been generated using [2].

AEC aims at estimating the desired speech by suppressing the echo [1]. AEC algorithms traditionally exploit the availability of loudspeaker signals to compute an estimate of this echo, which can then be subtracted from the microphone signals [1], [3]. Other approaches exist as well, such as approaches based on single-channel gains [4], data-driven end-to-end neural networks (NNs) [5], and hybrid combinations of model-based and data-driven approaches [6]–[8].

NR aims at estimating the desired speech by suppressing the near-end room noise [9], [10]. NR algorithms traditionally exploit the availability of microphone signals. Examples of NR algorithms include beamformers, such as the multi-channel Wiener filter (MWF) and generalised sidelobe canceller (GSC) [9], [10], Kalman filters [11], single-channel gains [12], data-driven end-to-end NNs [13] and hybrid combinations of model-based and data-driven approaches [14].

Algorithms for combined AEC and NR consequently aim at estimating the desired speech by jointly suppressing the echo and near-end room noise [1]. Combined AEC and NR algorithms exploit the availability of microphone and loudspeaker signals as illustrated in Fig. 1. To achieve this combined AEC

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). The scientific responsibility is assumed by its authors.

Arnout Roebben, Toon van Waterschoot and Marc Moonen are with the Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, B-3001 Leuven, Belgium (e-mail: arnout.roebben@esat.kuleuven.be; toon.vanwaterschoot@esat.kuleuven.be; marc.moonen@esat.kuleuven.be).

Jan Wouters is with the Department of Neurosciences, Research Group ExpORL, KU Leuven, B-3000 Leuven, Belgium (e-mail: jan.wouters@med.kuleuven.be).

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

and NR, generally, a cascade approach is adhered to, i.e., the AEC and NR are designed in isolation by selecting a separate signal model, formulating a separate cost function, and using a separate solution strategy. The AEC and NR are then cascaded one after the other, leading to algorithms where the AEC precedes the NR (AEC-NR) [15]–[18], the NR precedes the AEC (NR-AEC) [19]–[22], or variations thereof [23]–[25].

Despite the AEC being less perturbed by near-end room noise in the NR-AEC, the AEC-NR tends to outperform the NR-AEC in terms of echo suppression. Indeed, the AEC in the NR-AEC needs to track (the adaptivity of) the NR, which is not the case for the AEC in the AEC-NR [17], [20], [26]. Further, although the AEC needs to process the noisy microphone signal vectors, the AEC-NR can use the NR to suppress residual echo [27]. The NR-AEC, on the other hand, is advantageous computational-complexity-wise as only one echo path estimation per loudspeaker is required, opposed to one echo path estimation for each loudspeaker-microphone pair for the AEC-NR [17], [20], [26].

In this paper, the cascade approach is contrasted to an integrated approach where a single algorithm for joint AEC and NR is designed by selecting a single signal model, formulating a single cost function, and using a common solution strategy. Whereas the cascade approach, thus, designs two separate algorithms with distinct parameters, the integrated approach designs a single algorithm with shared parameters.

In [28] and [29], [30], this integrated approach has been applied with a linearly constrained minimum variance (LCMV) cost function, together with a microphone signal model [28] or extended signal model of stacked microphone and loudspeaker signal vectors [29], [30]. In [31], a combined state-space model for the echo path and the autoregressive desired speech has been considered, using a cost function and solution strategy based on Kalman filtering and expectation maximisation (EM). Integrated single-channel gains have been studied in [32]–[34].

Data-driven NNs generally do not assume any particular signal model, but do optimise a single cost function to perform both the AEC and NR task [35]–[38]. There can either be one NN jointly performing both tasks [36], [37], or dedicated subnetworks can separately perform one single task each [38]. These NNs can be preceded by a model-based AEC [39]–[42], hence leading to hybrid methods. This preceding AEC itself can also be of hybrid nature [43].

For specific choices of the signal model and cost function, the algorithms resulting from the cascade and integrated approaches can nevertheless be shown to be equivalent. In [15], [23], [44]–[46], a mean squared error (MSE) cost function has been used together with an extended signal model. While it is stressed that this integrated approach leads to a single algorithm with shared parameters for the AEC and NR, it has been shown that the resulting algorithm can be interpreted as the AEC-NR cascade algorithm in the 1-microphone/1-loudspeaker setup [15], [23], in the 2-microphone/1-loudspeaker setup with uncorrelated near-end room noise in both microphones [23], in the multi-microphone/1-loudspeaker setup when using a linear signal model and invertibility assumptions [47], and in the multi-microphone/multi-loudspeaker setup when using a specific generalised eigenvalue decomposition (GEVD) implementation and (multiframe) linear signal model for the echo paths [46]. Similarly, in the multi-microphone/1-loudspeaker setup it has been shown that the resulting algorithm can be interpreted as an NR-AEC cascade algorithm [44], and in the multi-microphone/multi-loudspeaker as an extended MWF (MWF_{ext}), i.e., an MWF on an extended signal model when using the specific GEVD implementation and (multiframe) linear signal model [46]. Nevertheless, while both the AEC-NR, NR-AEC and MWF_{ext} are theoretically equivalent under these specific conditions, their practical performances differ due to parameter estimation differences, e.g., as described supra for the AEC-NR and NR-AEC.

The MSE cost has thus only been partially explored: only for specific choices of the signal model and cost function, and only showing equivalence for the AEC-NR, NR-AEC and MWF_{ext}. In this paper, the aim is to apply the integrated approach to the most general multi-microphone/multi-loudspeaker setup, where the loudspeaker and microphone signals are possibly linearly related, allowing to generalise previous work, to propose new algorithms, and to study practical performance differences across algorithms. Therefore, the main research questions, tackled in this paper are as follows:

1) What algorithms are retrieved by applying the integrated approach to a combined AEC and NR problem in the general multi-microphone/multi-loudspeaker setup, possibly with linear dependencies between the loudspeaker and microphone signals, and using an MSE cost?
2) How do non-stationarities and imperfect correlation matrix estimation influence the practical performance?

As for the first research question, a single signal model of either the microphone signal vector, or the extended signal vector obtained by stacking microphone and loudspeaker signals will be selected, a single MSE cost function will be formulated, and a common solution strategy will be used. Selecting the microphone signal model, an MWF will be derived, whereas selecting the extended signal model, an extended MWF (MWF_{ext}) will be derived. This MWF_{ext} will be shown to be theoretically equivalent to several expressions, which turn out to be interpretable as specific cascade algorithms. Specifically, the MWF_{ext} will be shown to be equivalent to the AEC-NR and NR-AEC, thus generalising [15], [23], [44]–[46]. Further, the MWF_{ext} will be shown to be equivalent to an algorithm where an extended NR (NR_{ext}) precedes the AEC and a post-filter (PF) (NR_{ext}-AEC-PF) under the assumption of the echo paths being additive maps, i.e., preserving the addition operation. This algorithm without PF, NR_{ext}-AEC, has originally been proposed as a cascade algorithm for combined acoustic feedback cancellation (AFC) and NR [48], and as a cascade algorithm for combined AEC and NR [22]. In this paper, the algorithm will be adapted to the general AEC and NR problem, and the PF will be shown to be necessary for MSE optimality. Additionally, under rank-deficiency conditions the MWF_{ext} is non-unique, such that it will be shown that under these rank-deficiency conditions the theoretical equivalence of the AEC-NR, NR-AEC and NR_{ext}-AEC-PF expressions to the MWF_{ext} amounts to the expressions being specific, not necessarily minimum-norm solutions, for this MWF_{ext}.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

As for the second research question, non-stationarities and imperfect correlation matrix estimation effects will be analysed theoretically, and validated experimentally. The AEC-NR and NR_{ext}-AEC-PF attain best overall performance.

In Section II the cascade approach is reviewed first, discussing the AEC and NR design in isolation. This is contrasted to the integrated approach in Section III. The practical performance differences due to non-stationarities and imperfect correlation matrix estimation are analysed theoretically in Section IV. A computational complexity comparison is provided in Section V. After introducing the simulation setup in Section VI, the practical performances are also experimentally analysed in Section VII. Finally, Section VIII draws the conclusions. MATLAB code is available in [49].

II. Cascade approach

Section II-A and Section II-B describe the design of the isolated AEC and NR algorithms, optimal in the MSE sense for their isolated cost functions. The near-end room noise presence is thus neglected in the design of the AEC and the echo presence in the design of the NR. Section II-C considers the cascade approach for combined AEC and NR.

The signal model is presented in the z-domain to accommodate the duality between frequency domain and time domain. The conversion from z- to frequency domain is realised by replacing index z with frequency-bin index f, and possibly frame index k. The conversion from z- to time domain is realised by replacing the z-domain variables with time-lagged vectors, possibly multiplied with Toeplitz matrices, and replacing index z with time index t.

A. Acoustic echo cancellation (AEC)

AEC aims at estimating the desired speech by suppressing the echo originating from the loudspeakers. To this end, the echo signals are estimated from the loudspeaker signals and subtracted from the microphone signals [1] [50, Chapter 5].

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

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

This microphone signal vector can be decomposed into a desired speech signal vector $\mathbf{s}(z) \in \mathbb{C}^{M\times 1}$ and an echo signal vector $\mathbf{e}(z) \in \mathbb{C}^{M\times 1}$ as

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

where $\mathbf{e}(z)$ originates from the loudspeaker signal vector $\mathbf{l}(z)$

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

by means of a map $F(.) : \mathbb{C}^{L\times 1} \to \mathbb{C}^{M\times 1}$, i.e., $\mathbf{e}(z) = F(\mathbf{l}(z))$. While $F(.)$ can be a general map, it can also be restricted by, e.g., assuming $F(.)$ to be an additive or linear map. If the loudspeaker signal vector $\mathbf{l}(z)$ consists of a far-end room speech component $\mathbf{l}^s(z)$ and a far-end room noise component $\mathbf{l}^n(z)$, an additive map preserves this speech-noise relation as [22]

$$\begin{aligned} \mathbf{e}(z) &= F(\mathbf{l}^s(z) + \mathbf{l}^n(z)) = F(\mathbf{l}^s(z)) + F(\mathbf{l}^n(z)) && \text{(4a)} \\ &= \mathbf{e}^s(z) + \mathbf{e}^n(z), && \text{(4b)} \end{aligned}$$

with $\mathbf{e}^s(z)$ and $\mathbf{e}^n(z)$ the far-end room speech and noise components in the echo respectively. In a linear map,

$$\begin{aligned} \mathbf{e}(z) &= F\left(\mathbf{l}(z)\right) && \text{(5a)} \\ &= F_{\text{lin}}(z)\mathbf{l}(z), && \text{(5b)} \end{aligned}$$

with

$$F_{\text{lin}}(z) = \begin{bmatrix} f_1^1(z) & \cdots & f_1^L(z) \\ \vdots & \ddots & \vdots \\ f_M^1(z) & \cdots & f_M^L(z) \end{bmatrix}, \tag{6}$$

and $f_i^j(z)$ the transfer function of the echo path between the jth loudspeaker and the ith microphone. A continuous-time linear echo path combined with multi-rate digital signal processing (with aliasing), e.g., corresponds to an additive map rather than a linear map.

2) Cost function: The goal of the AEC is to minimise the MSE between the echo signal for a chosen reference microphone $r \in \{1, \cdots, M\}$ and the filtered loudspeaker signal vector $\mathbf{w}_{\text{AEC}}(z)^H\mathbf{l}(z)$, with $\mathbf{w}_{\text{AEC}}(z) \in \mathbb{C}^{L\times 1}$ [50, Chapter 5]:

$$\mathbf{w}_{\text{AEC}}(z) = \underset{\mathbf{w}(z)}{\operatorname{argmin}}\, \mathbb{E}\left\{ \left\| e_r(z) - \mathbf{w}(z)^H\mathbf{l}(z) \right\|_2^2 \right\}. \tag{7}$$

3) Solution strategy: Defining $R_{ll}(z) = \mathbb{E}\left\{\mathbf{l}(z)\mathbf{l}(z)^H\right\} \in \mathbb{C}^{L\times L}$ and $R_{le}(z) = \mathbb{E}\left\{\mathbf{l}(z)\mathbf{e}(z)^H\right\} \in \mathbb{C}^{L\times M}$ as the loudspeaker correlation matrix and the loudspeaker-echo cross-correlation matrix respectively, the solution to (7) is given as the solution to the Wiener-Hopf equations [50, Chapter 5]

$$R_{ll}(z)\mathbf{w}_{\text{AEC}}(z) = R_{le}(z)\mathbf{t}_r, \tag{8}$$

with $\mathbf{t}_r \in \mathbb{C}^{M\times 1}$ a unit vector with 1 at position r and 0 elsewhere. The solution to (8) is given as

$$\mathbf{w}_{\text{AEC}}(z) = R_{ll}(z)^g R_{le}(z)\mathbf{t}_r, \tag{9}$$

with $.^g$ denoting the generalised inverse[1,2] [51]. This generalised inverse is non-unique if $R_{ll}(z)$ is rank-deficient, although specific choices can be made such as the pseudo-inverse $.^\dagger$ which is uniquely defined, and with which (9) corresponds to the minimum-norm solution [52]. If $R_{ll}(z)$ is full-rank, i.e., the loudspeaker signals are linearly independent, then $.^g$ can be replaced with the inverse $.^{-1}$.

The desired speech estimate $\hat{s}_{r,\text{AEC}} \in \mathbb{C}$ in microphone r is given as [50, Chapter 5]

$$\hat{s}_{r,\text{AEC}}(z) = \mathbf{t}_r^H\mathbf{m}(z) - \mathbf{w}_{AEC}(z)^H\mathbf{l}(z). \tag{10}$$

[1]A generalised inverse $.^g$ only needs to satisfy the property $R_{ll}(z)R_{ll}(z)^g R_{ll}(z) = R_{ll}(z)$ [51]. A pseudo-inverse additionally needs to satisfy the properties $R_{ll}(z)^g R_{ll}(z)R_{ll}(z)^g = R_{ll}(z)^g$, $(R_{ll}(z)R_{ll}(z)^g)^H = R_{ll}(z)R_{ll}(z)^g$, and $(R_{ll}(z)^g R_{ll}(z))^H = R_{ll}(z)^g R_{ll}(z)$ [52].

[2]This generalised inverse is atypical in the literature as commonly the inverse or pseudo-inverse is used, e.g., [50, Chapter 1]. Nevertheless, the generalised inverse is already introduced for consistency with Section III.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

Depending on possible post-processing of the signals after applying the AEC, this procedure can be repeated for multiple choices of reference microphone.

B. Noise reduction (NR)

NR aims at estimating the desired speech by suppressing the near-end room noise through filtering the microphone signals [9].

1) Signal model: Considering an M-microphone setup, the microphone signal vector $\mathbf{m}(z) \in \mathbb{C}^{M\times 1}$ consists of a desired speech signal vector $\mathbf{s}(z) \in \mathbb{C}^{M\times 1}$ and a near-end room noise signal vector $\mathbf{n}(z) \in \mathbb{C}^{M\times 1}$

$$\mathbf{m}(z) = \mathbf{s}(z) + \mathbf{n}(z). \tag{11}$$

2) Cost function: The goal of the NR is to minimise the MSE between the desired speech signal for a chosen reference microphone $r \in \{1, \cdots, M\}$ and the filtered microphone signal vector $\mathbf{w}_{\text{NR}}(z)^H\mathbf{m}(z)$, with $\mathbf{w}_{\text{NR}}(z) \in \mathbb{C}^{M\times 1}$ [9], [53]:

$$\mathbf{w}_{\text{NR}}(z) = \underset{\mathbf{w}(z)}{\operatorname{argmin}}\, \mathbb{E}\left\{\left\|s_r(z) - \mathbf{w}(z)^H\mathbf{m}(z)\right\|_2^2\right\}. \tag{12}$$

3) Solution strategy: Defining $R_{mm}(z) = \mathbb{E}\left\{\mathbf{m}(z)\mathbf{m}(z)^H\right\} \in \mathbb{C}^{M\times M}$ and $R_{ss}(z) = \mathbb{E}\left\{\mathbf{s}(z)\mathbf{s}(z)^H\right\} \in \mathbb{C}^{M\times M}$ as the microphone correlation matrix and the desired speech correlation matrix respectively, and assuming $\mathbf{s}(z)$ and $\mathbf{n}(z)$ to be uncorrelated, the solution to (12) is given as [9]

$$\mathbf{w}_{\text{NR}}(z) = R_{mm}(z)^g R_{ss}(z)\mathbf{t}_r. \tag{13}$$

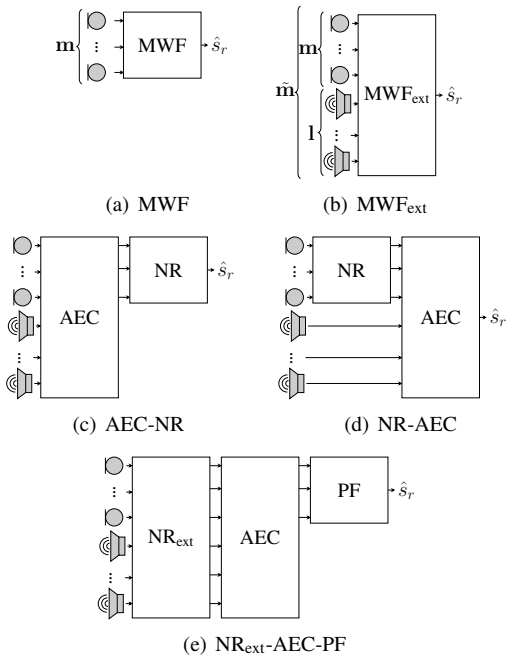
If $R_{mm}(z)$ is rank-deficient, then $.^g$ can again be chosen as $.^\dagger$, such that (13) corresponds to the minimum-norm solution. Indeed, $R_{mm}(z)$ can be rank-deficient, e.g., modelling $< M$ localised noise sources. If $R_{mm}(z)$ is full-rank, then $.^g$ can be replaced with $.^{-1}$. The desired speech estimate $\hat{s}_{r,\text{NR}}(z) \in \mathbb{C}$ in microphone r is given as [9]

$$\hat{s}_{r,\text{NR}}(z) = \mathbf{w}_{\text{NR}}(z)^H\mathbf{m}(z). \tag{14}$$

Depending on possible post-processing of the signals after applying the NR, this procedure can be repeated for multiple choices of reference microphone.

C. Combined AEC and NR

To resolve the combined AEC and NR problem, both algorithms are cascaded one after the other, where the AEC precedes the NR (AEC-NR) [15]–[18], the NR precedes the AEC (NR-AEC) [19]–[22] or variations thereof are considered [23]–[25]. As such, the cascade approach may suffer from algorithmic conflicts as the interaction between the isolated algorithms is not taken into account. Indeed, (residual) echo and near-end room noise may leak between the different stages of the cascade algorithms, thereby reducing performance.



(a) MWF (b) MWF$_{\text{ext}}$ (c) AEC-NR (d) NR-AEC (e) NR$_{\text{ext}}$-AEC-PF

Fig. 2. Cascade interpretation of the integrated algorithms. (a) The MWF is obtained using the microphone signal model. (b) The MWF$_{\text{ext}}$ is obtained using the extended signal model, which can be shown to be theoretically equivalent to (c) the AEC-NR, (d) the NR-AEC, and (e) the NR$_{\text{ext}}$-AEC-PF.

III. Integrated approach

In this section, the integrated approach is applied by selecting a signal model, including both the near-end room noise and echo, of either the microphone signal vector only or of an extended signal vector obtained by stacking microphone and loudspeaker signals (Section III-A), formulating an MSE cost function (Section III-B), and using a common solution strategy (Section III-C). Using the microphone signal model, an MWF will be obtained, whereas using the extended signal model an MWF$_{\text{ext}}$ will be obtained. Several theoretically equivalent expressions to the MWF$_{\text{ext}}$ are derived, which turn out to be nevertheless interpretable as specific cascade algorithms, as illustrated in Fig. 2. Specifically, the MWF$_{\text{ext}}$ is shown to be theoretically equivalent to algorithms where the AEC precedes the NR (AEC-NR), the NR precedes the AEC (NR-AEC), and the extended NR (NR$_{\text{ext}}$) precedes the AEC and a post-filter (PF) (NR$_{\text{ext}}$-AEC-PF). Under rank-deficiency conditions the MWF$_{\text{ext}}$ is non-unique, such that this theoretical equivalence amounts to the expressions being specific, not necessarily minimum-norm solutions, for this MWF$_{\text{ext}}$. For conciseness, the z-indices will be omitted from now on.

A. Signal model

The microphone signal vector $\mathbf{m} \in \mathbb{C}^{M\times 1}$ is defined as a mixture of a desired speech signal vector $\mathbf{s} \in \mathbb{C}^{M\times 1}$, an echo signal vector $\mathbf{e} \in \mathbb{C}^{M\times 1}$ and a near-end room noise signal vector $\mathbf{n} \in \mathbb{C}^{M\times 1}$:

$$\mathbf{m} = \mathbf{s} + \mathbf{e} + \mathbf{n}. \tag{15}$$

This signal model (15) will be used to derive the MWF in Section III-C1.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

An extended signal model can be constructed as well, stacking the microphone signal vector $\mathbf{m}$ and loudspeaker signal vector $\mathbf{l}$ into $\tilde{\mathbf{m}} \in \mathbb{C}^{(M+L)\times 1}$ as

$$\tilde{\mathbf{m}} = \begin{bmatrix}\mathbf{m}^\top & \mathbf{l}^\top\end{bmatrix}^\top \tag{16a}$$
$$= \tilde{\mathbf{s}} + \tilde{\mathbf{e}} + \tilde{\mathbf{n}} \tag{16b}$$
$$= \begin{bmatrix}\mathbf{s}\\ \mathbf{0}_{L\times 1}\end{bmatrix} + \begin{bmatrix}\mathbf{e}\\ \mathbf{l}\end{bmatrix} + \begin{bmatrix}\mathbf{n}\\ \mathbf{0}_{L\times 1}\end{bmatrix}, \tag{16c}$$

Herein, $\mathbf{0}_{L\times 1} \in \mathbb{C}^{L\times 1}$ corresponds to a vector containing only zero elements. This signal model (16) will be used to derive the MWF and MWF$_{\text{ext}}$ in Section III-C1 and Section III-C2 respectively, and the AEC-NR and NR-AEC in Section III-C3 and Section III-C4 respectively.

If the echo path is an additive map as well, $\mathbf{e}$ can be decomposed into the sum of far-end room speech and noise components, i.e., $\mathbf{e}^s$ and $\mathbf{e}^n$, originating from $\mathbf{l}^s$ and $\mathbf{l}^n$ respectively, such that (15) and (16) can be decomposed into

$$\mathbf{m} = \mathbf{s} + \mathbf{e}^s + \mathbf{e}^n + \mathbf{n}, \tag{17}$$

and

$$\tilde{\mathbf{m}} = \tilde{\mathbf{s}} + \tilde{\mathbf{e}}^s + \tilde{\mathbf{e}}^n + \tilde{\mathbf{n}} \tag{18a}$$
$$= \begin{bmatrix}\mathbf{s}\\ \mathbf{0}_{L\times 1}\end{bmatrix} + \begin{bmatrix}\mathbf{e}^s\\ \mathbf{l}^s\end{bmatrix} + \begin{bmatrix}\mathbf{e}^n\\ \mathbf{l}^n\end{bmatrix} + \begin{bmatrix}\mathbf{n}\\ \mathbf{0}_{L\times 1}\end{bmatrix}. \tag{18b}$$

The signal model (16) will be used to derive the NR$_{\text{ext}}$-AEC-PF in Section III-C5.

As $\mathbf{s}$ and $\mathbf{e}^s$ represent speech signal vectors, they can be assumed to attain an on-off behaviour. To detect these on-off periods, voice activity detectors (VADs) are assumed to be available, i.e., VAD$_s$ and VAD$_{e^s}$ differentiate between on-off periods in $\mathbf{s}$ and $\mathbf{e}^s$ respectively. As $\mathbf{n}$ and $\mathbf{e}^n$, on the other hand, represent noise signal vectors, they can be assumed to be always-on. Thus, the following regimes can be defined:

- VAD$_s = 1$, VAD$_{e^s} = 1$: $\mathbf{m} = \mathbf{s} + \mathbf{e} + \mathbf{n}$ recorded,
- VAD$_s = 1$, VAD$_{e^s} = 0$: $\mathbf{s} + \mathbf{e}^n + \mathbf{n}$ recorded,
- VAD$_s = 0$, VAD$_{e^s} = 1$: $\mathbf{e} + \mathbf{n}$ recorded,
- VAD$_s = 0$, VAD$_{e^s} = 0$: $\mathbf{e}^n + \mathbf{n}$ recorded.

For general non-additive echo path maps, VAD$_{e^s}$ then does not detect on-off periods of $\mathbf{e}^s$ itself as possibly only non-linear combinations of the loudspeaker signals can be observed, but VAD$_{e^s}$ then still discriminates between activity of all components in the echo signal and activity of only partial components. In the literature (e.g., [30], [45]), the echo is often assumed to only contain a stationary always-on signal, which corresponds to only keeping $\mathbf{e}^n$ and discarding VAD$_{e^s}$.

Further, the following assumptions are made:

- $\mathbf{s}$, $\mathbf{n}$, $\mathbf{e}^s$ and $\mathbf{e}^n$ are uncorrelated, $\mathbf{l}^s$ and $\mathbf{l}^n$ are uncorrelated, and $\mathbf{l}^s$ and $\mathbf{l}^n$ are uncorrelated with $\mathbf{s}$ and $\mathbf{n}$.
- For the theoretical derivation all (speech) signals will be assumed stationary during their on periods. However, in Section IV, this assumption will be relaxed by considering the influence of non-stationary signal vectors. Indeed, in practice, $\mathbf{s}$ (and $\mathbf{e}^s$) are non-stationary, whereas $\mathbf{n}$ (and $\mathbf{e}^n$) are nevertheless considered to be stationary.

The general M-microphone/L-loudspeaker setup is considered, and no assumptions are made regarding the rank of R_{ss}, R_{nn}, R_{mm} or R_{ll}, thus possibly modelling localised desired speech and near-end room noise sources and linearly related loudspeaker and microphone signals.

B. Cost function

The goal of the integrated AEC and NR algorithm is to minimise the MSE between the desired speech signal $s_r \in \mathbb{C}$ and a filtered signal vector, simultaneously suppressing the echo signal vector $\mathbf{e}$ and near-end room noise signal vector $\mathbf{n}$, by either considering the microphone signal model (15), i.e.,

$$\mathbf{w}_{\text{int}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}\left\{\left\|s_r - \mathbf{w}^H\mathbf{m}\right\|_2^2\right\}, \tag{19}$$

to design $\mathbf{w}_{\text{int}} \in \mathbb{C}^{M\times 1}$, or the extended signal model (16), i.e.,

$$\tilde{\mathbf{w}}_{\text{int}} = \underset{\tilde{\mathbf{w}}}{\operatorname{argmin}}\, \mathbb{E}\left\{\left\|s_r - \tilde{\mathbf{w}}^H\tilde{\mathbf{m}}\right\|_2^2\right\}, \tag{20}$$

to design $\tilde{\mathbf{w}}_{\text{int}} \in \mathbb{C}^{(M+L)\times 1}$. As illustrated in Fig. 1, the loudspeaker signals are only filtered by $\tilde{\mathbf{w}}_{\text{int}}$ when estimating s_r, thus without affecting the playback in the near-end room.

C. Solution strategy

1) Multi-channel Wiener filter (MWF): Using the microphone signal model (15), the solution to (19) is obtained as

$$\boxed{\mathbf{w}_{\text{int}} = R_{mm}^g R_{ss}\mathbf{t}_r,} \tag{21}$$

hence corresponding to the multi-channel Wiener filter (MWF) (Fig. 2(a)). $R_{mm} = R_{ss} + R_{ee} + R_{nn}$ can be computed during periods of desired speech and echo activity (VAD$_s = 1 \wedge$ VAD$_{e^s} = 1$). R_{ss} can be computed by constructing $R_{nn}+R_{ee}$ during periods of simultaneous desired speech inactivity and echo activity (VAD$_s = 0 \wedge$ VAD$_{e^s} = 1$), and subtracting the result from R_{mm}.

2) Extended multi-channel Wiener filter (MWF$_{ext}$): Similar to the MWF, using the extended signal model (16), the solution to (20) is obtained as an MWF with extended correlation matrices $R_{\tilde{m}\tilde{m}} = \mathbb{E}\left\{\tilde{\mathbf{m}}\tilde{\mathbf{m}}^H\right\} \in \mathbb{C}^{(M+L)\times(M+L)}$ and $R_{\tilde{s}\tilde{s}} = \mathbb{E}\left\{\tilde{\mathbf{s}}\tilde{\mathbf{s}}^H\right\} \in \mathbb{C}^{(M+L)\times(M+L)}$

$$\boxed{\tilde{\mathbf{w}}_{\text{int}} = R_{\tilde{m}\tilde{m}}^g R_{\tilde{s}\tilde{s}}\tilde{\mathbf{t}}_r,} \tag{22}$$

hence, named the extended MWF (MWF$_{\text{ext}}$) (Fig. 2(b)). $R_{\tilde{m}\tilde{m}}$ can be computed when VAD$_s = 1 \wedge$ VAD$_{e^s} = 1$, while $R_{\tilde{s}\tilde{s}}$ can be computed by constructing $R_{\tilde{e}\tilde{e}} + R_{\tilde{n}\tilde{n}}$ when VAD$_s = 0 \wedge$ VAD$_{e^s} = 1$ and subtracting the result from $R_{\tilde{m}\tilde{m}}$.

3) AEC precedes NR (AEC-NR): A closed-form expression for a valid generalised matrix inverse in (22) is as follows (cfr. Supplementary material Section IX)

$$R_{\tilde{m}\tilde{m}}^{g,g} = \begin{bmatrix}\Sigma_{mm}^\dagger & -\Sigma_{mm}^\dagger R_{el}R_{ll}^\dagger\\ -R_{ll}^\dagger R_{le}\Sigma_{mm}^\dagger & R_{ll}^\dagger + R_{ll}^\dagger R_{le}\Sigma_{mm}^\dagger R_{el}R_{ll}^\dagger\end{bmatrix}, \tag{23}$$

with $\Sigma_{mm} = R_{mm} - R_{el}R_{ll}^\dagger R_{le}$. Herein, $R_{\tilde{m}\tilde{m}}^{g,g} \in \mathbb{C}^{(M+L)\times(M+L)}$ is thus one possible valid expression for all of the generally non-unique generalised inverses $R_{\tilde{m}\tilde{m}}^g$ (cfr. Supplementary material Section IX). If Σ_{mm} is full-rank,

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

i.e., $\text{rank}(\Sigma_{mm}) = M$, then (23) corresponds to the pseudo-inverse. If additionally $\text{rank}(R_{ll}) = L$, then (23) corresponds to the inverse.

Plugging (23) into (22), the following expression is obtained, with $\mathbb{I}_{M\times M} \in \mathbb{C}^{M\times M}$ the identity matrix:

$$\tilde{\mathbf{w}}_{\text{int}} = \underbrace{\begin{bmatrix} \mathbb{I}_{M\times M} & 0_{M\times L} \\ -R_{ll}^{\dagger}R_{le} & 0_{L\times L} \end{bmatrix}}_{\text{AEC}} \underbrace{\begin{bmatrix} \Sigma_{mm}^{\dagger}R_{ss} \\ 0_{L\times M} \end{bmatrix}}_{\text{NR}} \mathbf{t}_r, \tag{24}$$

which can be interpreted as an AEC preceding an NR (Fig. 2(c)). Indeed, $R_{ll}^{\dagger}R_{le}$ corresponds to the minimum-norm estimate of the echo paths as described in Section II-A. Furthermore, Σ_{mm} can be interpreted as the microphone correlation matrix after applying the AEC, such that $\Sigma_{mm}^{\dagger}R_{ss}$ corresponds to an MWF aimed at suppressing near-end room noise and residual echo after the AEC, and thus acts as a multi-channel post-filter. It will nevertheless be referred to as NR for consistency with the cascaded design definitions in Section II, and for consistency with the literature [23], [45], [46], [54]. To show this equivalence, the microphone correlation matrix after applying the AEC is defined as

$$\begin{aligned} &\mathbb{E}\left\{\left(\mathbf{m} - R_{el}R_{ll}^{\dagger}\mathbf{l}\right)\left(\mathbf{m} - R_{el}R_{ll}^{\dagger}\mathbf{l}\right)^{H}\right\} \\ &= R_{mm} - 2R_{el}R_{ll}^{\dagger}R_{le} + R_{el}R_{ll}^{\dagger}R_{ll}R_{ll}^{\dagger}R_{le}, \end{aligned} \tag{25}$$

wherein, $R_{ll}^{\dagger}R_{ll}R_{ll}^{\dagger} = R_{ll}^{\dagger}$ as per definition of the pseudo-inverse [52], such that (25) is indeed equal to Σ_{mm}. Consequently, if $\text{rank}(\Sigma_{mm}) = M$, then the MWF_{ext} (22) is uniquely defined and equivalent to the AEC-NR (24) as (23) then corresponds to the uniquely-defined pseudo-inverse. If $\text{rank}(\Sigma_{mm}) < M$, (22) is not uniquely defined as $R_{\tilde{m}\tilde{m}}^{g}$ is not unique. Nevertheless, (24) then still corresponds to one of the solutions of the Wiener-Hopf equations $R_{\tilde{m}\tilde{m}}\tilde{\mathbf{w}} = R_{\tilde{s}\tilde{s}}\tilde{\mathbf{t}}_r$, be it not necessarily to the minimum-norm solution.

To compute the correlation matrices in (24), R_{ll} and R_{le} can be readily constructed when $\text{VAD}_{e^s} = 1$ as $\mathbf{s}$, $\mathbf{n}$ and $\mathbf{e}$ are uncorrelated, although in practice R_{le} is computed when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$ to reduce excess error from $\mathbf{s}$ [50, Chapter 8]. Σ_{mm} can then be constructed by computing the microphone correlation matrix after applying the AEC when $\text{VAD}_s = 1 \wedge \text{VAD}_{e^s} = 1$, whereas R_{ss} can be computed by subtracting the microphone correlation matrix after applying the AEC when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$, i.e., subtracting $R_{ee} + R_{nn} - R_{el}R_{ll}^{\dagger}R_{le}$, from Σ_{mm}.

If the echo path is linear and the AEC filters are chosen sufficiently long to model this echo path, i.e., $\mathbf{e} = F_{\text{lin}}\mathbf{l}$ and $R_{le} = R_{ll}F_{\text{lin}}^{H}$, $\tilde{\mathbf{w}}_{\text{int}}$ (24) is equivalent to $\tilde{\mathbf{w}}_{\text{int,lin}} \in \mathbb{C}^{(M+L)\times 1}$:

$$\tilde{\mathbf{w}}_{\text{int,lin}} = \underbrace{\begin{bmatrix} \mathbb{I}_{M\times M} & 0_{M\times L} \\ -R_{ll}^{\dagger}R_{ll}F_{\text{lin}}^{H} & 0_{L\times L} \end{bmatrix}}_{\text{AEC}} \underbrace{\begin{bmatrix} \left(R_{ss} + R_{nn}\right)^{-1}R_{ss} \\ 0_{L\times M} \end{bmatrix}}_{\text{NR}} \mathbf{t}_r, \tag{26}$$

where the NR does not need to suppress residual echo as the AEC already fully cancels the echo. If R_{ll} is full-rank, then $R_{ll}^{\dagger}R_{ll}F_{\text{lin}} = F_{\text{lin}}$, corresponding to straightforward AEC. If R_{ll} is rank-deficient, i.e., if the loudspeaker signals are linearly dependent, the AEC $R_{ll}^{\dagger}R_{ll}F^{H}$ does not correspond to the true echo path matrix F_{lin}, but to a weighted version thereof. This linear weighting results in a far-end room dependence of the AEC, as R_{ll} depends on the far-end room scenario. The AEC, therefore, also needs to adapt to changes in the far-end room, such that (26) provides a generalised expression for the multi-channel AEC problem, e.g., extending [55].

4) NR precedes AEC (NR-AEC): By reordering the block matrices in (24), the following expression is obtained

$$\tilde{\mathbf{w}}_{\text{int}} = \underbrace{\begin{bmatrix} \Sigma_{mm}^{\dagger}R_{ss} & 0_{M\times L} \\ 0_{L\times M} & \mathbb{I}_{L\times L} \end{bmatrix}}_{\text{NR}} \underbrace{\begin{bmatrix} \mathbb{I}_{M\times M} \\ -R_{ll}^{\dagger}R_{le}\left(\Sigma_{mm}^{\dagger}R_{ss}\right) \end{bmatrix}}_{\text{AEC}} \mathbf{t}_r. \tag{27}$$

Here, the NR aimed to suppress the noise and residual echo now precedes the AEC (Fig. 2(d)), as opposed to the AEC-NR (24). As the echo signal vector after applying the NR is affected by the NR, the AEC not only has to model the echo paths but rather the combination of the echo paths and the NR, which leads to the additional factor $\Sigma_{mm}^{\dagger}R_{ss}$ in the AEC.

The advantage performance-wise of the NR-AEC over the AEC-NR is that the AEC operates under reduced near-end room noise influence, thereby using a cascade implementation where the NR precedes the AEC. However, to do so, the loudspeaker correlation matrix and the loudspeaker-echo cross-correlation matrix are only computed after applying an NR, such that Σ_{mm} is generally replaced by R_{mm}, resulting in an expression $\tilde{\mathbf{w}}_{\text{mod}} \in \mathbb{C}^{(M+L)\times 1}$ with a modified NR and AEC, NR_{mod} and AEC_{mod} respectively [19]–[22],

$$\tilde{\mathbf{w}}_{\text{mod}} = \underbrace{\begin{bmatrix} R_{mm}^{\dagger}R_{ss} & 0_{M\times L} \\ 0_{L\times M} & \mathbb{I}_{L\times L} \end{bmatrix}}_{\text{NR}_{\text{mod}}} \underbrace{\begin{bmatrix} \mathbb{I}_{M\times M} \\ -R_{ll}^{\dagger}R_{le}\left(R_{mm}^{\dagger}R_{ss}\right) \end{bmatrix}}_{\text{AEC}_{\text{mod}}} \mathbf{t}_r, \tag{28}$$

which then loses its MSE optimality. R_{mm} is computed when $\text{VAD}_s = 1 \wedge \text{VAD}_{e^s} = 1$, and R_{ss} is computed by subtracting $R_{ee} + R_{nn}$, collected when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$, from R_{mm}. R_{ll} and $R_{le}\left(R_{mm}^{\dagger}R_{ss}\right)$ are then computed as the loudspeaker correlation matrix and the loudspeaker-echo cross-correlation matrix after applying the NR_{mod}.

5) Extended NR precedes AEC and PF (NR_{ext}-AEC-PF): The additional assumption is now made that the echo path is an additive map, hence using the signal model (18). As detailed in Section II-A, this assumption is more general than the assumption of a linear echo path as it only assumes that the plus operator is preserved by the map (Section II-A1). This makes it possible to decouple the effect of $\mathbf{l}^s$ and $\mathbf{l}^n$ in the microphone into $\mathbf{e}^s$ and $\mathbf{e}^n$. Time-varying echo paths are also allowed, as time-variance does not affect this property. Furthermore, assume that $R_{ll}^{\dagger}R_{le}$ is a solution of the Wiener-Hopf equations $R_{l^sl^s}W = R_{l^se^s}$, i.e., $R_{l^sl^s}R_{ll}^{\dagger}R_{le} = R_{l^se^s}$. The latter assumption corresponds to the minimum-norm echo path estimate using the loudspeaker signals also being a solution of the Wiener-Hopf equations for the echo path estimate using only the far-end room speech component in

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

the loudspeaker signals, thereby extending the assumption in [22] to the case where R_{ll} and $R_{l^s l^s}$ can be rank-deficient.

Under these assumptions, an NR_{ext} filter $W_{\text{NR}_{\text{ext}}} \in \mathbb{C}^{(M+L)\times(M+L)}$ can be defined as the MSE-optimal filter to suppress both $\tilde{\mathbf{n}}$ and $\tilde{\mathbf{e}}^n$ while preserving both $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{e}}^s$:

$$W_{\text{NR}_{\text{ext}}} = R_{\tilde{m}\tilde{m}}^{g,g} \left(R_{\tilde{s}\tilde{s}} + R_{\tilde{e}^s\tilde{e}^s}\right). \tag{29}$$

Applying $W_{\text{NR}_{\text{ext}}}$ to $\tilde{\mathbf{m}}$ leads to an extended signal vector,

$$\tilde{\mathbf{m}}^{\text{`}} = \begin{bmatrix} \mathbf{m}^{\text{`}} \\ \mathbf{l}^{\text{`}} \end{bmatrix} = W_{\text{NR}_{\text{ext}}}^H \tilde{\mathbf{m}}. \tag{30}$$

Using (29) and the assumption that $R_{l^s l^s} R_{ll}^\dagger R_{le} = R_{l^s e^s}$, the following filter can be shown to be equivalent to (22) if $R_{\tilde{m}\tilde{m}}^{g,g}$ is chosen as defined in (23) for $R_{\tilde{m}\tilde{m}}^{g}$ both in (22) and (31) (cfr. Supplementary material Section X)

$$\tilde{\mathbf{w}}_{\text{int}} = \underbrace{R_{\tilde{m}\tilde{m}}^{g,g} \left(R_{\tilde{s}\tilde{s}} + R_{\tilde{e}^s\tilde{e}^s}\right)}_{\text{NR}_{\text{ext}}} \underbrace{\begin{bmatrix} \mathbb{I}_{M\times M} & 0_{M\times L} \\ -R_{l^{\text{`}}l^{\text{`}}}^\dagger R_{l^{\text{`}}m^{\text{`}}} & 0_{L\times L} \end{bmatrix}}_{\text{AEC}} \cdot \underbrace{\begin{bmatrix} \Sigma_{m^{\text{`}}m^{\text{`}}}^\dagger R_{s^{\text{`}}s} \\ 0_{L\times M} \end{bmatrix}}_{\text{PF}} \mathbf{t}_r, \tag{31}$$

In the post-filter (PF), $\Sigma_{m^{\text{`}}m^{\text{`}}} = R_{m^{\text{`}}m^{\text{`}}} - R_{m^{\text{`}}l^{\text{`}}} R_{l^{\text{`}}l^{\text{`}}}^\dagger R_{l^{\text{`}}m^{\text{`}}}$ corresponds to the microphone correlation matrix after applying the NR_{ext} and the AEC. Consequently, (31) can be interpreted as an NR_{ext} preceding an AEC and a PF (Fig. 2(e)), where the NR_{ext} is aimed at suppressing the near-end room noise and the far-end room noise component in the echo, and the AEC is aimed at removing the far-end room speech and residual noise components in the echo. The PF is aimed at suppressing residual noise and echo, while preserving the desired speech. $R_{\tilde{m}\tilde{m}}$ can be collected when both the desired speech and the far-end room speech component in the echo are active, i.e., when $\text{VAD}_s = 1 \wedge \text{VAD}_{e^s} = 1$. $R_{\tilde{s}\tilde{s}} + R_{\tilde{e}^s\tilde{e}^s}$ can be computed by collecting $R_{\tilde{e}^n\tilde{e}^n} + R_{\tilde{n}\tilde{n}}$ when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 0$, and subtracting the result from $R_{\tilde{m}\tilde{m}}$. $R_{l^{\text{`}}l^{\text{`}}}$ and $R_{l^{\text{`}}m^{\text{`}}}$ can be readily collected after applying the NR_{ext} when $\text{VAD}_{e^s} = 1$ as $\mathbf{s}^{\text{`}}$, $\mathbf{e}^{\text{`}}$ and $\mathbf{n}^{\text{`}}$ are uncorrelated. In practice, excess error due to $\mathbf{s}^{\text{`}}$ can nevertheless be avoided by computing $R_{l^{\text{`}}m^{\text{`}}}$ when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$. For the PF, $\Sigma_{m^{\text{`}}m^{\text{`}}}$ can be collected when $\text{VAD}_s = 1 \wedge \text{VAD}_{e^s} = 1$ after applying the NR_{ext} and AEC. $R_{s^{\text{`}}s}$ can be computed by subtracting $R_{e^{\text{`}}e} - R_{m^{\text{`}}l^{\text{`}}} R_{l^{\text{`}}l^{\text{`}}}^\dagger R_{l^{\text{`}}e} + R_{n^{\text{`}}n}$, collected when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$, from $\Sigma_{m^{\text{`}}m}$, collected when $\text{VAD}_s = 1 \wedge \text{VAD}_{e^s} = 1$. Alternatively, when $\begin{bmatrix} \mathbb{I}_{M\times M} & 0_{M\times L} \end{bmatrix} W_{\text{NR}_{\text{ext}}} \begin{bmatrix} \mathbb{I}_{M\times M} \\ 0_{L\times M} \end{bmatrix}$ is of full-rank, $R_{s^{\text{`}}s}$ can be computed by subtracting $R_{e^{\text{`}}e^{\text{`}}} - R_{m^{\text{`}}l^{\text{`}}} R_{l^{\text{`}}l^{\text{`}}}^\dagger R_{l^{\text{`}}m^{\text{`}}} + R_{n^{\text{`}}n^{\text{`}}}$ from $\Sigma_{m^{\text{`}}m^{\text{`}}}$ and post-multiplying with $\left(\begin{bmatrix} \mathbb{I}_{M\times M} & 0_{M\times L} \end{bmatrix} W_{\text{NR}_{\text{ext}}} \begin{bmatrix} \mathbb{I}_{M\times M} \\ 0_{L\times M} \end{bmatrix} \right)^{-1}$.

If $\text{rank}(\Sigma_{mm}) = M$, then the MWF_{ext} is uniquely defined and equivalent to the NR_{ext}-AEC-PF (31) as (23) then corresponds to the uniquely-defined pseudo-inverse. If $\text{rank}(\Sigma_{mm}) < M$, then (23) is not uniquely defined as $R_{\tilde{m}\tilde{m}}^{g}$ is not unique. Nevertheless, (31) then still corresponds to one of the solutions of the Wiener-Hopf equations $R_{\tilde{m}\tilde{m}}\tilde{\mathbf{w}} = R_{\tilde{s}s}\mathbf{t}_r$, be it not necessarily the minimum-norm solution.

The difference between the NR-AEC (27) and the first two filters (NR_{ext} and AEC) in the NR_{ext}-AEC-PF (31) can be seen as follows [22]: Whereas the NR in (27) operates solely on the microphones, aimed at suppressing the near-end room noise, the NR_{ext} in (31) operates on both the microphones and the loudspeakers (without affecting the playback), aimed at suppressing both the near-end room noise and the far-end room noise component in the echo. Consequently, the AEC in (27) aims at removing the entire echo signal vector, while the AEC in (31) mostly aims at suppressing the far-end room speech component in the echo (and the residual far-end room noise component in the echo). Further, whereas the AEC in (27) needs to adapt to the preceding NR, the AEC in (31) actually does not need to adapt to the preceding NR_{ext}, because (cfr. Supplementary material Section X),

$$R_{l^{\text{`}}l^{\text{`}}}^\dagger R_{l^{\text{`}}m^{\text{`}}} = R_{l^s l^s}^\dagger R_{l^s e^s}, \tag{32}$$

which corresponds to the MSE-optimal AEC based on the far-end room speech component in the echo, which is thus independent of the NR_{ext}. Supplementary material Section X thus generalises [22], where this independence of NR_{ext} and AEC was shown when considering a full-rank R_{mm} and R_{ll}.

The NR_{ext}-AEC-PF can also be interpreted as the AEC-NR to which a preceding NR_{ext} is added. This NR_{ext} then already aims at partially reducing the near-end room noise and far-end room noise component in the echo before applying the AEC-NR. The PF in the NR_{ext}-AEC-PF has the same functionality as the NR in the AEC-NR, but will be referred to as PF to avoid confusion with this NR_{ext}.

If the echo path is linear and the AEC filters are chosen sufficiently long to model this echo path, i.e., $\mathbf{e} = F_{\text{lin}}\mathbf{l}$ and $R_{le} = R_{ll}F_{\text{lin}}^H$, (31) simplifies to $\tilde{\mathbf{w}}_{\text{int,lin}}$ (cfr. Supplementary material Section XI)

$$\tilde{\mathbf{w}}_{\text{int,lin}} = \underbrace{R_{\tilde{m}\tilde{m}}^{g,g} \left(R_{\tilde{s}\tilde{s}} + R_{\tilde{e}^s\tilde{e}^s}\right)}_{\text{NR}_{\text{ext}}} \underbrace{\begin{bmatrix} \mathbb{I}_{M\times M} \\ -R_{l^s l^s}^\dagger R_{l^s l^s} F_{\text{lin}}^H \end{bmatrix}}_{\text{AEC}} \mathbf{t}_r, \tag{33}$$

where the PF is reduced to $\begin{bmatrix} \mathbb{I}_{M\times M} \\ 0_{L\times M} \end{bmatrix}$ and hence omitted.

IV. Practical considerations

Although theoretical equivalence exists between the expressions in Section III, the practical performances of their cascade algorithm implementations differ due to non-stationarities and imperfect correlation matrix estimation. In practice the expressions of Section III are either implemented in the frequency domain, replacing index z with frequency-bin index f and possibly frame index k, or in the time domain, replacing the z-domain variables with time-lagged vectors and replacing index z with time index t. To illustrate the implications of the practical considerations, the frequency-domain expressions will be given next, although a similar reasoning can be made for the time-domain expressions.

Regarding the non-stationarities, (short-term) stationarity of the spatial position of the desired speech, near-end room noise

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

and loudspeakers can be assumed [45]. Spectral (short-term) stationarity of the near-end room noise and the far-end room noise component in the loudspeakers and echo can also be assumed as they model, e.g., sensor and diffuse noise sources. Correspondingly, the effect of spectral non-stationarities in the desired speech and (the far-end room speech component in) the loudspeakers and echo is focused upon. These non-stationarities, e.g., lead to a different contribution of the loudspeaker correlation matrix depending on the regime, such that $R_{ll}^{\{\text{s=1},e^s=1\}}(f)$ recorded when $\text{VAD}_s = 1 \wedge \text{VAD}_{e^s} = 1$ and $R_{ll}^{\{\text{s=0},e^s=1\}}(f)$ recorded when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$ differ, i.e., $R_{ll}^{\{\text{s=1},e^s=1\}}(f) \neq R_{ll}^{\{\text{s=0},e^s=1\}}(f)$. However, due to the assumed (short-term) stationarity of the near-end room noise, $R_{nn}^{\{\text{s=1},\text{e}^s=1\}}(f) = R_{nn}^{\{\text{s=1},\text{e}^s=0\}}(f) = R_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f) = R_{nn}^{\{\text{s=0},\text{e}^s=0\}}(f) = R_{nn}(f)$, and a similar property holds for the far-end room noise component in the echo.

Regarding the correlation matrix estimation, as there is only access to a finite number of samples and only during the regimes as specified by $\text{VAD}_s(k,f)$ and $\text{VAD}_{e^s}(k,f)$ (Section III-A), the correlation matrices are estimated using time averaging during these regimes. To illustrate this procedure, the microphone correlation matrix $\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f) \in \mathbb{C}^{M\times M}$ in the MWF (21) can be estimated across the K frames where $\text{VAD}_s(k,f) = 1 \wedge \text{VAD}_{e^s}(k,f) = 1$ as

$$\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{m}(k,f)\mathbf{m}(k,f)^H. \tag{34}$$

Herein, $\hat{\cdot}$ is used to denote estimated variables. As during the regimes specified by $\text{VAD}_s(k,f)$ and $\text{VAD}_{e^s}(k,f)$ there is, e.g., no access to desired speech only, $\hat{R}_{ss}^{\{\text{s=1},\text{e}^s=1\}}$ cannot be readily estimated by time averaging. Nevertheless, $\hat{R}_{ss}^{\{\text{s=1},\text{e}^s=1\}}(f)$ can be estimated by subtracting the correlation matrices $\hat{R}_{ee}^{\{\text{s=0},\text{e}^s=1\}}(f) + \hat{R}_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f)$ estimated when $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$ from $\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f)$.

This subtraction can also be performed using a generalised eigenvalue decomposition (GEVD) on this matrix pencil $\{\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f), \hat{R}_{ee}^{\{\text{s=0},\text{e}^s=1\}}(f) + \hat{R}_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f)\}$ to enforce a rank constraint on $\hat{R}_{ss}^{\{\text{s=1},\text{e}^s=1\}}$ [53], [56]. This rank constraint can be imposed as the ground truth R_{ss} generally models a limited number of desired speech sources and consequently generally has rank $R_s < M$. To this end, $\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f)$ and $\hat{R}_{ee}^{\{\text{s=0},\text{e}^s=1\}}(f) + \hat{R}_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f)$ are jointly diagonalised, and only the R_s modes with highest signal-to-noise ratio (SNR) are retained when subtracting both. Collecting the generalised eigenvectors in the columns of $\hat{Q}(f) \in \mathbb{C}^{M\times M}$, denoting the generalised eigenvalues of $\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f)$ by $\hat{\lambda}_{m_i}(f)$, and denoting the generalised eigenvalues of $\hat{R}_{e}^{\{\text{s=0},\text{e}^s=1\}}(f) + \hat{R}_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f)$ by $\hat{\lambda}_{(e+n)_i}(f)$, the GEVD adheres to the following equations [53]

$$\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f) = \hat{Q}(f)\hat{\Lambda}_m(f)\hat{Q}(f)^H \tag{35a}$$

$$= \hat{Q}(f)\text{diag}\Big(\hat{\lambda}_{m_1}(f), \cdots, \hat{\lambda}_{m_M}(f)\Big)\hat{Q}(f)^H \tag{35b}$$

$$\hat{R}_{ee}^{\{\text{s=0},\text{e}^s=1\}}(f) + \hat{R}_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f) = \hat{Q}(f)\hat{\Lambda}_{(e+n)}(f)\hat{Q}(f)^H \tag{35c}$$

$$= \hat{Q}(f)\text{diag}\Big(\hat{\lambda}_{(e+n)_1}(f), \cdots, \hat{\lambda}_{(e+n)_M}(f)\Big)\hat{Q}(f)^H. \tag{35d}$$

Using this GEVD, the rank-R_s GEVD approximation $\hat{R}_{ss}^{\{\text{s=1},\text{e}^s=1\}}(f)$ can then be computed as [53]

$$\hat{R}_{ss}^{\{\text{s=1},\text{e}^s=1\}} = \hat{Q}(f)\cdot\text{diag}\big(\hat{\lambda}_{m_1}(f) - \hat{\lambda}_{(e+n)_1}(f), \cdots, \hat{\lambda}_{m_{R_s}}(f) - \hat{\lambda}_{(e+n)_{R_s}(f)}, 0, \cdots, 0\big)\hat{Q}(f)^H. \tag{36}$$

In what follows, the influence of non-stationarities and imperfect correlation matrix estimation is discussed in more detail for each of the algorithms. Afterwards, the connection is made with recent advances in data-driven approaches.

1) Multi-channel Wiener filter (MWF): Regarding the (far-end speech component in the) loudspeakers and echo, non-stationary echo (together with imperfect correlation matrix estimation) leads to the same contribution of the noise correlation matrix but a different contribution of the echo correlation matrix in $\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f)$ and $\hat{R}_{ee}^{\{\text{s=0},\text{e}^s=1\}}(f) + \hat{R}_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f)$, such that by subtracting both, next to the desired speech correlation matrix, the echo correlation matrix is partially retained. The MWF then aims at estimating the desired speech and partially the echo. A GEVD approximation does not resolve this issue.

Regarding the desired speech, the non-stationarity effect is more contained than for the loudspeakers and echo. Indeed, while the echo correlation matrix is present on both sides of the matrix pencil $\{\hat{R}_{mm}^{\{\text{s=1},\text{e}^s=1\}}(f), \hat{R}_{ee}^{\{\text{s=0},\text{e}^s=1\}}(f) + \hat{R}_{nn}^{\{\text{s=0},\text{e}^s=1\}}(f)\}$, the desired speech correlation matrix is only present on the left-hand side. Consequently, the effect of desired speech non-stationarities is effectively limited to the MWF only being optimal with respect to the desired speech's average temporal characteristics.

2) Extended multi-channel Wiener filter (MWF_{ext}): Regarding the (far-end speech component in the) loudspeakers and echo, in [46], it is argued that plain subtraction of $R_{\tilde{m}\tilde{m}}^{\{\text{s=1},\text{e}^s=1\}}(f)$ by $R_{\tilde{e}\tilde{e}}^{\{\text{s=0},\text{e}^s=1\}}(f) + R_{\tilde{n}\tilde{n}}^{\{\text{s=0},\text{e}^s=1\}}(f)$ leads to partial reconstruction of the echo due to non-stationary loudspeaker and echo signal vectors having a different contribution in the extended echo correlation matrix on both sides. Nevertheless, it is argued that the GEVD-implementation of the MWF_{ext} using a linear signal model with sufficiently long filters to model the echo path is not affected by these non-stationarities [46]. Indeed, referring to the extended signal model (16), a vector $\mathbf{0}_{L\times 1}$ is stacked under $\mathbf{s}(k,f)$ and $\mathbf{n}(k,f)$, while $\mathbf{l}(k,f)$ is stacked under $\mathbf{e}(k,f)$. This difference in structure is reflected in the generalised eigenvectors, as the GEVD of $\{R_{\tilde{m}\tilde{m}}^{\{\text{s=1},\text{e}^s=1\}}(f), R_{\tilde{e}\tilde{e}}^{\{\text{s=0},\text{e}^s=1\}}(f) + R_{\tilde{n}\tilde{n}}^{\{\text{s=0},\text{e}^s=1\}}(f)\}$ is defined as

$$R_{\tilde{m}\tilde{m}}^{\{\text{s=1},\text{e}^s=1\}}(f) = \tilde{Q}(f)\tilde{\Lambda}_{\tilde{m}}(f)\tilde{Q}(f)^H \tag{37a}$$

$$R_{\tilde{e}\tilde{e}}^{\{\text{s=0},\text{e}^s=1\}}(f) + R_{\tilde{n}\tilde{n}}^{\{\text{s=0},\text{e}^s=1\}}(f) = \tilde{Q}(f)\tilde{\Lambda}_{(\tilde{e}+\tilde{n})}(f)\tilde{Q}(f)^H. \tag{37b}$$

with $\tilde{Q}(f), \tilde{\Lambda}_{\tilde{m}}(f), \tilde{\Lambda}_{(\tilde{e}+\tilde{n})}(f) \in \mathbb{C}^{(M+L)\times(M+L)}$ containing the generalised eigenvectors and eigenvalues respectively. As for a linear echo path with sufficiently long filters to model this echo path $\text{rank}(R_{\tilde{e}\tilde{e}}) \leq L$, $\tilde{Q}(f)$ reflects the zero-structure as [46]

$$\tilde{Q}(f) = \left[\begin{array}{c|c} Q_1(f) & \tilde{Q}_2(f) \\ \hline 0_{M\times M} & \end{array}\right]. \tag{38}$$

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

Herein, $\begin{bmatrix} Q_1(f)^\top & 0_{M\times M} \end{bmatrix}^\top \in \mathbb{C}^{(M+L)\times M}$ is related to the desired speech and near-end room noise, and $\tilde{Q}_2(f) \in \mathbb{C}^{(M+L)\times L}$ to the echo, such that only retaining $\tilde{Q}_2(f)$ allows for suppressing the echo even given its non-stationarity [46].

While this property is true for a linear signal model with sufficiently long filters to model the echo paths, in practice correlation matrix estimation using time averaging likely leads to the extended echo correlation matrices being of full-rank, i.e., being of rank $M+L$, due to imperfect correlation matrix estimation or due to the application of a general echo path (on linearly independent loudspeaker signals). The contribution of this extended echo correlation matrix on each side of the matrix pencil $\{\hat{\tilde{R}}_{\tilde{m}\tilde{m}}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f), \hat{\tilde{R}}_{\tilde{e}\tilde{e}}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f) + \hat{\tilde{R}}_{\tilde{n}\tilde{n}}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f)\}$ will then differ, such that all generalised eigenvectors will be influenced by the loudspeaker and echo signal vectors, and the echo will be partially retained by the MWF_ext.

As for the MWF (Section IV-1), the effect of desired speech non-stationarities is limited, leading to an MWF_ext optimal to the desired speech's average temporal characteristics.

3) AEC precedes NR (AEC-NR): As both $\hat{R}_{ll}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f)$ and $\hat{R}_{le}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f)$ are updated under the same conditions, i.e., $\text{VAD}_s = 0 \wedge \text{VAD}_{e^s} = 1$, both matrices are similarly affected by loudspeaker and echo non-stationarities. Consequently, while it is true that in general the estimated echo path may differ based on the loudspeaker and echo non-stationarities, i.e., $\hat{R}_{ll}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)^\dagger \hat{R}_{le}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)$ not necessarily equalling $\hat{R}_{ll}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f)^\dagger \hat{R}_{le}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f)$, the estimated echo path is nevertheless optimal with respect to the loudspeaker's and echo's average temporal characteristics during that regime. Additionally, under additional assumptions of the echo path, stronger statements can be made. For example, modelling a volume increase α jointly for all loudspeakers when $\text{VAD}_s = 1$ (i.e., $\alpha \mathbf{1}(k,f)$) with respect to $\text{VAD}_s = 0$, leads to $\alpha^2 \hat{R}_{ll}^{\{\text{s}=1,\text{e}^\text{s}=1\}} = \hat{R}_{ll}^{\{\text{s}=0,\text{e}^\text{s}=1\}}$. Under this condition, when the echo path is linear with possible undermodelling, it nevertheless holds that $\hat{R}_{ll}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)^\dagger \hat{R}_{le}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f) = \hat{R}_{ll}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f)^\dagger \hat{R}_{le}^{\{\text{s}=0,\text{e}^\text{s}=1\}}(f)$. As the AEC is not updated when $\text{VAD}_s = 1$, non-stationarities of the desired speech do not affect the AEC.

The non-stationarities of the loudspeaker and echo affect the NR similarly to the MWF as described in Section IV-1, although to a reduced extent as the AEC already partially reduces the echo. In the limit case where the echo path is linear, the filters are sufficiently long to model echo paths, and there is no noise presence, the AEC can already remove the entire echo, such that the NR is not affected by loudspeaker and echo non-stationarities. The effect of desired speech non-stationarities is the same as for the MWF (Section IV-1).

4) NR precedes AEC (NR-AEC): In its modified form (28), the NR corresponds to the MWF (21), such that the effect of non-stationary echo, loudspeaker and desired speech for the NR is the same as described in Section IV-1.

While the NR thus partially reconstructs the echo due to loudspeaker and echo non-stationarities, the AEC further reduces the echo using an estimated echo path optimal with respect to the loudspeaker's and echo's average temporal characteristics, such that the NR-AEC in its modified form can be interpreted as an MWF adjusted with an AEC to suppress residual echo. As described in Section IV-3, the AEC is not affected by desired speech non-stationarities due to the AEC only being updated when $\text{VAD}_s(k,f) = 0$.

5) NR_{ext} precedes AEC and PF (NR_{ext}-AEC-PF): The NR_ext is less affected by non-stationarities in the far-end room speech component in the loudspeakers and echo than the MWF_ext as these components only appear in $\hat{R}_{\tilde{m}\tilde{m}}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)$ when computing $\hat{R}_{\tilde{s}\tilde{s}}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f) + \hat{R}_{\tilde{e}^s\tilde{e}^s}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)$ by subtracting $\hat{R}_{\tilde{e}^n\tilde{e}^n}^{\{\text{s}=0,\text{e}^\text{s}=0\}}(f) + \hat{R}_{\tilde{n}\tilde{n}}^{\{\text{s}=0,\text{e}^\text{s}=0\}}(f)$ from $\hat{R}_{\tilde{m}\tilde{m}}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)$, possibly using a GEVD. The effect of desired speech non-stationarities is thus limited as $\hat{R}_{\tilde{s}\tilde{s}}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)$ only appears in $\hat{R}_{\tilde{m}\tilde{m}}^{\{\text{s}=1,\text{e}^\text{s}=1\}}(f)$.

The AEC is similarly affected by non-stationarities as in the AEC-NR (Section IV-3) and the NR-AEC (Section IV-4).

The PF takes the form of an MWF (after applying the NR_ext and the AEC), such that the PF is affected by echo, loudspeaker and desired speech non-stationarities similar to the MWF (Section IV-1). Nevertheless, as the NR_ext-AEC already partially removes the near-end room noise and echo, the PF is less affected by the loudspeaker and echo non-stationarities than the MWF.

6) Connection to data-driven approaches: The expressions (21), (22), (24), (28), and (31) can also be aligned with advances in data-driven approaches, as these expressions provide model structures optimal with respect to the MSE cost function, of which the model-based parameter estimation in terms of correlation matrices can be replaced by data-driven parameter estimation, e.g., using NNs.

For example, the AEC can be equivalently implemented with a recursive least squares (RLS) adaptive filter [57], which can be unified with Kalman filtering [57]. As there exist many variants of hybrid Kalman filters for AEC with NN-based parameter estimation [7], [8], the AEC-NR, NR-AEC and NR_ext-AEC-PF can make use of these hybrid Kalman filters. Another example would be to supply the MWF, NR, or PF through a NN (a so-called mask), optimised through an MSE or other cost function, or to replace the MWF, NR, or PF with a NN. This leads to algorithms with linear/hybrid AEC filters followed by a NN as in [39]–[43] with the AEC-NR of (24).

V. Computational complexity

Next to the different practical considerations discussed in Section IV, the different algorithms also have a different computational complexity. To this end, a floating-point operation (FLOP) count is performed using big $\mathcal{O}$ notation as a function of the dimensions of the correlation matrices involved, for the cascade algorithm implementations of the expressions as specified in Section III-C. More specifically, the expressions (21), (22), (24), (28), and (31) in function of correlation matrices and pseudo-inverses thereof are considered in their frequency-domain formulation (cfr. Section IV). Nevertheless, similar expressions exist for the time-domain expressions, and more numerically efficient computational implementations do

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

Algorithm	Computational complexity		
	MWF_{ext}/NR_{ext}	**AEC**	**MWF/NR/PF**
MWF			$\mathcal{O}\left(M^3\right)$
MWF_{ext}	$\mathcal{O}\left((M+L)^3\right)$		
AEC-NR		$\mathcal{O}\left(L^3\right)$	$\mathcal{O}\left(M^3\right)$
NR-AEC		$\mathcal{O}\left(L^3\right)$	$\mathcal{O}\left(M^3\right)$
NR_{ext}-AEC-PF	$\mathcal{O}\left((M+L)^3\right)$	$\mathcal{O}\left(L^3\right)$	$\mathcal{O}\left(M^3\right)$

TABLE I
THE COMPUTATIONAL COMPLEXITY IN BIG $\mathcal{O}$ NOTATION OF (21), (22), (24), (28), AND (31) AS A FUNCTION OF THE DIMENSIONS OF THE CORRELATION MATRICES INVOLVED. THE NR_{EXT}-AEC-PF HAS THE HIGHEST COMPLEXITY, WHILE THE MWF HAS THE LOWEST.

exist to compute these expressions as, e.g., detailed in [1]. The complexity of the different stages making up the algorithms is detailed first, wherein stages with similar complexity have been grouped together, leading the an overall computational complexity comparison. The resulting computational complexity for the different algorithms is also given in Table I.

1) MWF_{ext}/NR_{ext}: The MWF_{ext} and NR_{ext} consider both the microphone and loudspeaker signals. Both algorithms require the updating of the correlation matrices using outer products ($\mathcal{O}((M+L)^2)$), the computation of $\hat{R}_{\tilde{s}\tilde{s}}$ (MWF_{ext}) or $\hat{R}_{\tilde{s}\tilde{s}} + \hat{R}_{\tilde{e}^s\tilde{e}^s}$ (NR_{ext}) either using plain matrix subtraction ($\mathcal{O}((M+L)^2)$) or subtraction guided by a GEVD ($\mathcal{O}((M+L)^3)$), the computation of a pseudo-inverse ($O((M+L)^3)$), and a matrix-matrix multiplication ($\mathcal{O}((M+L)^3)$) [58]. The filter application requires an inner product ($\mathcal{O}(M+L)$) [58]. Thus, the MWF_{ext} and NR_{ext} have a complexity of $\mathcal{O}((M+L)^3)$, although the explicit computation of the pseudo-inverse could be avoided using the Woodbury identity [58].

2) AEC: The AEC requires the updating of the loudspeaker-loudspeaker and loudspeaker-echo correlation matrices using outer products ($\mathcal{O}(L^2)$ and $O(LM)$), the computation of the pseudo-inverse of this loudspeaker-loudspeaker correlation matrix ($\mathcal{O}(L^3)$), and a matrix-matrix multiplication of this pseudo-inverse with the loudspeaker-echo correlation matrix ($\mathcal{O}(L^2M)$). The filter application requires an inner product and vector subtraction ($\mathcal{O}(2L+M)$). Thus, a complexity of $\mathcal{O}(L^3)$ is obtained, although the explicit computation of the pseudo-inverse could again be avoided using the Woodbury identity, thereby resorting to RLS or other adaptive filter implementations [58].

3) MWF/NR/PF: The MWF, NR, and PF require similar computations as the MWF_{ext} and NR_{ext} but only consider the microphone signals, resulting in a complexity of $\mathcal{O}(M^3)$.

4) Comparison: As the MWF only involves the microphone signals, it is computationally the cheapest. The NR_{ext}-AEC-PF also involves loudspeaker signals, and requires three stages, such that it is computationally the most expensive. The computational complexity of the MWF_{ext}, AEC-NR, and NR-AEC is intermediate, e.g., the NR_{ext}-AFC-Pf can be interpreted as the AEC-NR with prior NR_{ext} stage. When using efficient adaptive filtering implementations to compute the AEC, and when there is only one reference microphone, the NR-AEC is more computationally efficient than the AEC-NR as only one AEC filter per loudspeaker is required rather than one AEC filter for each loudspeaker-microphone pair [17], [20], [26].

VI. EXPERIMENT DESIGN

Although the MWF_{ext}, AEC-NR, NR-AEC and NR_{ext}-AEC-PF are theoretically equivalent, they differ in their practical estimation as described in Section IV, such that their practical performances will differ. To experimentally validate these performance differences, the algorithms are compared to one another in the acoustic scenarios described in Section VI-A, using the algorithm settings described in Section VI-B, and the performance measures described in Section VI-C. As the main focus of this paper is on algorithms derived from signal models, i.e., (15) and (16), the comparison will focus on these model-based algorithms. The reader is referred to [8], for a comparison of the benefits and downsides of model-based, hybrid and data-driven algorithms. A relative comparison between these algorithms, and the impact of variations in signal-to-noise ratio (SNR) and signal-to-echo ratio (SER) is focussed upon, rather than aiming at maximal performance.

A. Acoustic scenarios

1) Setup-1: Five scenarios with varying desired speech source, near-end room noise source and loudspeaker positions are considered in a $5\,\text{m} \times 5\,\text{m} \times 3\,\text{m}$ room with a wall reflection coefficient of 0.15 ($T_{60} = 0.11\,\text{s}$), corresponding to the scenarios examined in [22]. To this end, the source-to-microphone and loudspeaker-to-microphone impulse responses of length 128 samples are created using the randomised image method (RIM) [59] with a $16 \cdot 10^3$ Hz sampling rate and randomised distances of $0.13\,\text{m}$. $M = 2$ microphones are positioned at $\begin{bmatrix}2 & 1.9 & 1\end{bmatrix}\text{m}$ and $\begin{bmatrix}2 & 1.8 & 1\end{bmatrix}\text{m}$, while the position of the one desired speech source, one near-end room noise source and $L = 2$ loudspeakers are varied by placing these sources at congruent angles in a circle with a $0.2\,\text{m}$ radius around the mean microphone position. The desired speech source in the loudspeakers consists of sentences of the hearing in noise test (HINT) database, concatenated with $5\,\text{s}$ of silence [60], and the near-end room noise source consists of babble noise to model competing speakers [61]. Additionally, the far-end room speech component in the loudspeakers consists of HINT sentences, and the far-end room noise component in the loudspeakers consists of white noise, e.g., to model sensor and far-end room noise, of which the power ratio is set to $0\,\text{dB}$. The relative power ratio between the echo signals in the microphones is set to $0\,\text{dB}$. All signals are $30\,\text{s}$ long. The power ratio between the far-end room speech and noise components in the loudspeakers, and the power ratio between the echo signals both equal $0\,\text{dB}$. The input SNR (SNR^{in}) and SER (SER^{in}) ratio in the reference microphone are varied between $-15\,\text{dB}$ and $15\,\text{dB}$.

2) Setup-2: As in Setup-1, five scenarios with varying desired speech source, near-end room noise source and loudspeaker positions are considered. However, rather than using simulated impulse responses, measured impulse responses from the meeting room in the Aachen impulse response (AIR) database are selected [62]. As desired speech source and speech component in the loudspeakers $30\,\text{s}$ long sequences of HINT sentences are used [60], and as noise source and noise component in the loudspeakers office noise from the

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

diverse environments multi-channel acoustic noise database (DEMAND) is used [63]. One noise source and one loudspeaker are considered, for which the SNR^{in} and SER^{in} are set to $5\,\text{dB}$ and $0\,\text{dB}$ respectively. The power ratio between the far-end room speech and noise components in the loudspeakers is set to $0\,\text{dB}$ before feeding them through a (mild) non-linearity $\frac{\arctan(\alpha(\mathbf{l}^s+\mathbf{l}^n))}{\alpha}$ with $\alpha = 1$ to model loudspeaker non-linearities as per [8]. Two microphones are considered.

B. Algorithm settings

In order to avoid the correlation matrix initialisation influencing the performance of the algorithms, and to focus on the steady-state performance, the correlation matrices are estimated across the entire $30\,\text{s}$ of data.

Furthermore, the filters are calculated in the short-time Fourier transform (STFT) domain, the algorithms can also be implemented in the time domain, thereby replacing the index z with frequency-bin index f and frame index k, although as noted in Section II and Section IV, the algorithms can be equivalently implemented in the time domain. A squared root Hann window, window size of 512 samples, window shift of 256 samples and a sampling rate of $16 \cdot 10^3$ Hz are used in Setup-1. To accommodate for the larger impulse response length in the AIR database, a window size of 2048 samples and window shift of 1024 samples is used in Setup-2.

As one desired speech source is assumed, the rank of $\hat{R}_{ss}^{\{\text{s=1,e}^\text{s}\text{=1}\}}(f)$ in the MWF_{ext}, in the NR of the AEC-NR and the NR-AEC, and in the PF of the NR_{ext}-AEC-PF is enforced to be equal to one by using a GEVD [53]. Similarly, as one desired speech source and two independent loudspeaker sources are considered, the rank of $\hat{R}_{\tilde{s}\tilde{s}}^{\{\text{s=1,e}^\text{s}\text{=1}\}}(f) + \hat{R}_{\tilde{e}^s\tilde{e}^s}^{\{\text{s=1,e}^\text{s}\text{=1}\}}(f)$ in the NR_{ext} of the NR_{ext}-AEC-PF is enforced to equal three by using a GEVD. Indeed, in [53], it is shown that the rank of these estimated correlation matrices should be enforced to equal the number of sources using GEVD approximations. As by the assumption $\hat{R}_{l^sl^s}(f)\hat{R}_{ll}(f)^\dagger\hat{R}_{le}(f) = \hat{R}_{l^se^s}(f)$, $\begin{bmatrix}\mathbb{I}_{M\times M} & 0_{M\times L}\end{bmatrix}\hat{W}_{\text{NR}_{\text{ext}}}(f)\begin{bmatrix}0_{M\times L}\\ \mathbb{I}_{L\times L}\end{bmatrix} = 0_{M\times M}$ (Section III-C5), this zero-structure is enforced. Corresponding to the literature, e.g., [15], [20], [46], the algorithms are implemented in their cascade configuration, feeding the processed signals from one stage in the cascade to the next. The NR-AEC is thus implemented in its modified form (28).

Ideal VADs for the desired speech and the far-end room speech component in the echo are assumed in Setup-1 to restrain the influence of VAD errors. In Setup-2, ideal VADs are compared to a NN-based VAD from [64]. Loudspeaker signals are fed to the NN to obtain the VAD flags related to VAD_{e^s}, and microphone signals are fed to the NN to obtain the VAD flags related to VAD_s.

C. Performance measures

Noise reduction performance is calculated using the intelligibility-weighted SNR improvement $\left(\Delta\text{SNR}^{\text{I}}\right)$, echo reduction performance using the intelligibility-weighted SER improvement $\left(\Delta\text{SER}^{\text{I}}\right)$ and speech distortion (SD) performance using the intelligibility-weighted speech distortion $\left(\text{SD}^{\text{I}}\right)$ [65], [66], which are calculated as

$$\Delta\text{SNR}^{\text{I}} = \sum_{n_o=1}^{N_o} I_{n_o}\left(\text{SNR}_{n_o}^{\text{out}} - \text{SNR}_{n_o}^{\text{in}}\right) \tag{39a}$$

$$\Delta\text{SER}^{\text{I}} = \sum_{n_o=1}^{N_o} I_{n_o}\left(\text{SER}_{n_o}^{\text{out}} - \text{SER}_{n_o}^{\text{in}}\right) \tag{39b}$$

$$\text{SD}^{\text{I}} = \sum_{n_o=1}^{N_o} I_{n_o}\text{SD}_{n_o}, \tag{39c}$$

with,

$$\text{SNR}_{n_o} = 10\log_{10}\left(\frac{P_{n_o}^s}{P_{n_o}^n}\right),\ \text{SER}_{n_o} = 10\log_{10}\left(\frac{P_{n_o}^s}{P_{n_o}^e}\right)$$
$$\text{and } \text{SD}_{n_o} = 10\log_{10}\left(\frac{P_{n_o}^{s,\text{out}}}{P_{n_o}^{s,\text{in}}}\right). \tag{40}$$

Herein, $P_{n_0}^s$, $P_{n_0}^n$ and $P_{n_0}^e \in \mathbb{R}$ refer to the desired speech, near-end room noise and echo signal powers in the n_0th one-third octave band $n_0 \in \{1, \cdots, N_0\}$. The SNR, SER and SD measures are intelligibility-weighted using the band importance $I_{n_o} \in \mathbb{R}$ of [67, Table 3]. A higher $\Delta\text{SER}^{\text{I}}$, $\Delta\text{SNR}^{\text{I}}$, and an SD^{I} closer to zero indicates higher performance.

The extended short-time objective intelligibility (ESTOI) improvement $\left(\Delta\text{ESTOI} = \text{ESTOI}^{\text{out}}\text{-ESTOI}^{\text{in}}\right)$ is applied as it predicts intelligibility when desired speech is corrupted with modulated interferers, such as echo and noise [68]. Speech quality is evaluated with the perceptual evaluation of speech quality (PESQ) [69]. To support these results using a different model, intelligibility and quality are also evaluated with the hearing-aid speech perception index (HASPI) version 2 improvement $\left(\Delta\text{HASPI} = \text{HASPI}^{\text{out}}\text{-HASPI}^{\text{in}}\right)$ and hearing-aid speech quality index (HASQI) version 2 improvement $\left(\Delta\text{HASQI} = \text{HASQI}^{\text{out}}\text{-HASQI}^{\text{in}}\right)$ [70], [71]. For HASPI and HASQI no hearing loss is assumed. Higher ΔESTOI, ΔPESQ, ΔHASPI, and ΔHASQI indicates higher performance.

Finally, in Setup-2, the NN-based acoustic echo cancellation mean opinion score (AECMOS) is also evaluated [72]. To this end, $\Delta\text{MOS}_{\text{echo}}$ and $\Delta\text{MOS}_{\text{other}}$ measure improvement in AEC performance similar to $\Delta\text{SER}^{\text{I}}$ and improvement in degradation due to other causes such as noise and speech distortion, thereby bundling $\Delta\text{SNR}^{\text{I}}$ and SD^{I}. Higher $\Delta\text{MOS}_{\text{echo}}$ and $\Delta\text{MOS}_{\text{other}}$ indicates higher performance.

VII. Results and Discussion

As the NR_{ext}-AEC-PF constitutes a new algorithm, Section VII-A first separately studies the performance of this algorithm, detailing the performance implications of each filter in the cascade. Thereafter, the performance of the integrated algorithms is compared to one another in Section VII-B.

A. NR_{ext} precedes AEC and PF (NR_{ext}-AEC-PF)

Fig. 3 shows the performance attained by each filter in the NR_{ext}-AEC-PF cascade algorithm implementation to investigate the improvement offered by each filter in Setup-1.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

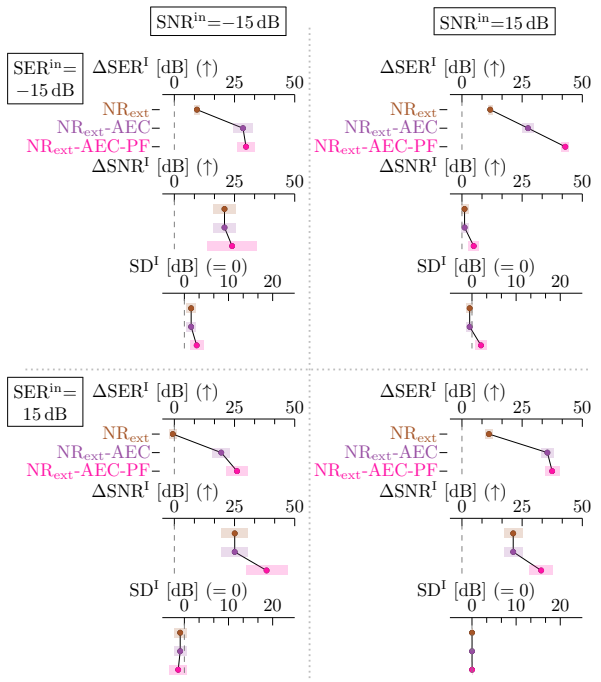


Fig. 3. Performance attained by each filter in the NR_{ext}-AEC-PF by means of the intelligibility-weighted signal-to-echo ratio improvement $(\Delta\text{SER}^{\text{I}})$, intelligibility-weighted signal-to-noise ratio improvement $(\Delta\text{SNR}^{\text{I}})$, and intelligibility weighted speech distortion. The mean performance is represented by the solid dots and the standard deviation by the shading. The performance generally increases with each filter in the cascade.

The NR_{ext} already improves the SER^{I} and SNR^{I} as the NR_{ext} aims at jointly suppressing the far-end room noise component in the echo and the near-end room noise. This SER^{I} improvement is larger for echo-dominant than for noise-dominant settings, and vice versa for the SNR^{I} improvement, since the NR_{ext} attributes more degrees of freedom towards suppressing the power-dominant interferer.

The AEC further improves the SER^{I} as the AEC aims at suppressing the echo, and as the excess error in the AEC due to noise perturbations is diminished by prior application of the NR_{ext} filter. The SNR^{I} and SD^{I} remain unaffected by the AEC, as the AEC does not filter the microphone signals, and so does not alter the near-end room noise or the desired speech.

The PF aims at further suppressing the near-end room noise and the echo, such that the PF improves both the SNR^{I} and SER^{I} with respect to the NR_{ext}-AEC. This SNR^{I} and SER^{I} improvement comes at the expense of a slightly increased SD^{I} as both the NR_{ext} and the PF distort the desired speech.

B. Comparison

1) Setup-1: Fig. 4 compares the performance of the integrated algorithms, as according to Section IV the theoretically equivalent algorithms of Section III differ practically due to non-stationarities and imperfect correlation matrix estimation. To this end, Fig. 4(a) shows the echo cancellation, noise reduction and speech distortion performance separately, which are amalgamated into the instrumental measures of Fig. 4(b). The AEC-NR and NR_{ext}-AEC-PF generally attain top performance.

The MWF attains lower performance than the AEC-NR, as the MWF does not take the loudspeaker information into account. Indeed there are only two microphones, such that the MWF can only completely cancel one interfering source. However, as there are two loudspeakers and one near-end room noise source, there are three interfering sources, such that the MWF lacks the degrees of freedom to cancel these interfering sources. This effect is mostly apparent at $\text{SER}^{\text{in}} = -15\,\text{dB}$, as the two echo sources then dominate the one near-end room noise source, leading to low SNR^{I}, SER^{I} and instrumental measure improvements, and high SD^{I}. While the NR in the AEC-NR can also be interpreted as an MWF after preprocessing by an AEC, this NR suffers less from a lack of degrees of freedom as the AEC already partially suppresses the echo, such that the AEC-NR scales with the number of loudspeakers.

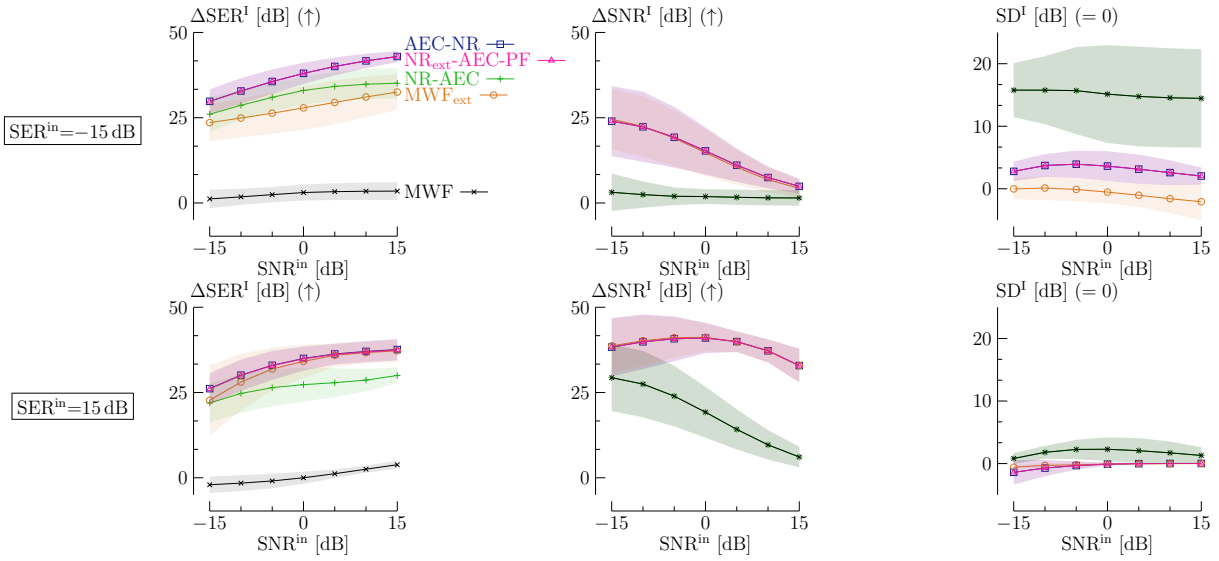
The MWF_{ext} also scales with the number of loudspeakers, in this way increasing the SNR^{I}, SER^{I} and instrumental measures and decreasing the SD^{I} with respect to the MWF. Nevertheless, the MWF_{ext} SER^{I} improvement is lower than for the AEC-NR. As described in Section IV, the MWF_{ext} suffers from loudspeaker and echo non-stationarities, while the AEC in the AEC-NR is less affected, thereby attaining a lower SER^{I} improvement than the AEC-NR. This effect is primarily visible at $\text{SER}^{\text{in}} = -15\,\text{dB}$ as the echo then dominates the noise.

The NR-AEC realises the same decreased SNR^{I} improvement and high SD^{I} with respect to the AEC-NR as the MWF. Indeed, the NR of the NR-AEC in its modified form (28) is identical to the MWF. Adding an AEC after the NR increases the SER^{I} compared to the MWF, but as the AEC neither alters the near-end room noise nor the desired speech, the AEC does not alter the SNR^{I} or the SD^{I}. Despite the addition of the AEC, the SER^{I} improvement of the NR-AEC is also inferior to the AEC-NR, as the NR in the NR-AEC already distorts the desired speech, and as the AEC needs to track the NR next to the echo paths. Due to the limited SNR^{I} improvement and high SD^{I}, the instrumental measure improvement of the NR-AEC is also limited, as mostly apparent at $\text{SER}^{\text{in}} = -15\,\text{dB}$ as the decreased performance of the NR is then mostly apparent.

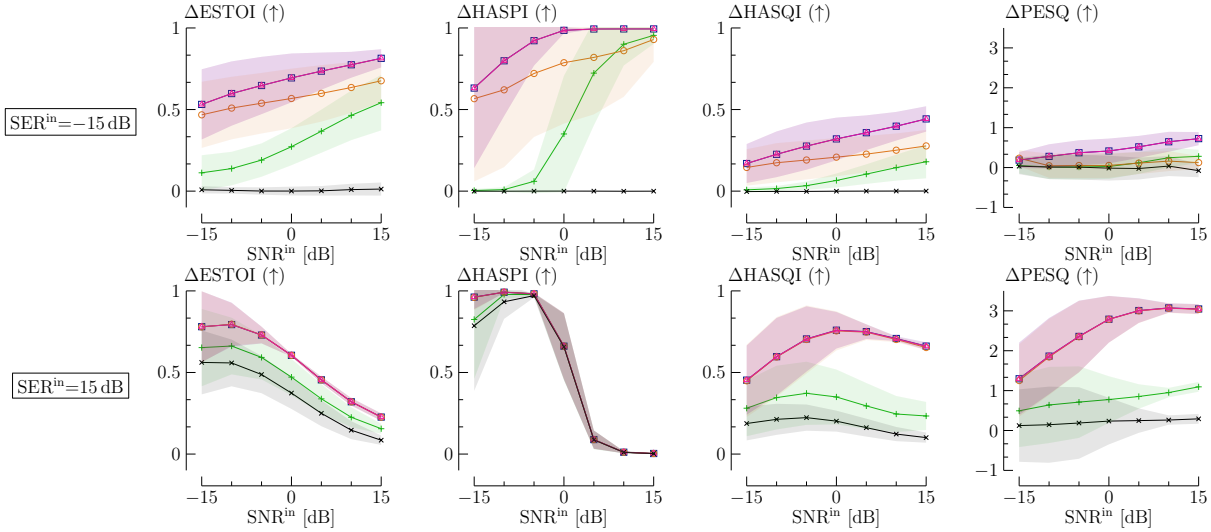
The NR_{ext}-AEC-PF also incorporates a filter to suppress the near-end room noise preceding the AEC as for the NR-AEC, namely the NR_{ext}. However, as opposed to the MWF, the NR_{ext} next to suppressing the near-end room noise also suppresses the far-end room noise component in the echo, and therefore scales with the number of loudspeakers, resulting in less SD^{I} and an increased SNR^{I} with respect to the MWF. The SNR^{I} improvement of the NR_{ext}-AEC-PF additionally decreases less quickly with an increasing SNR^{in} than the MWF. In fact, the NR_{ext}-AEC-PF attains similar performance to the AEC-NR. While, the AEC operates under reduced noise due to the addition of the NR_{ext}, the AEC, although theoretically independent of the NR_{ext}, is in practice affected by the NR_{ext} due to the GEVD and imperfect correlation matrix estimation. This effect is, nevertheless, much more limited than for the NR-AEC due to this theoretical independence between the NR_{ext} and the AEC as also argued in [22]. Both effects of reduced noise and NR_{ext} dependence, however, seem to balance each other out, such that the NR_{ext}-AEC-PF and the AEC-NR achieve a similar performance gain. Alternatively,

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802



(a) Intelligibility-weighted signal-to-echo ratio improvement $(\Delta\text{SER}^{\text{I}})$, intelligibility-weighted signal-to-noise ratio improvement $(\Delta\text{SNR}^{\text{I}})$, and intelligibility weighted speech distortion (SD^{I})



(b) Extended short-time objective intelligibility improvement (ΔESTOI), hearing-aid speech perception index version 2 improvement (ΔHASPI), hearing-aid speech quality index version 2 improvement (ΔHASQI), and perceptual evaluation of speech quality improvement (ΔPESQ)

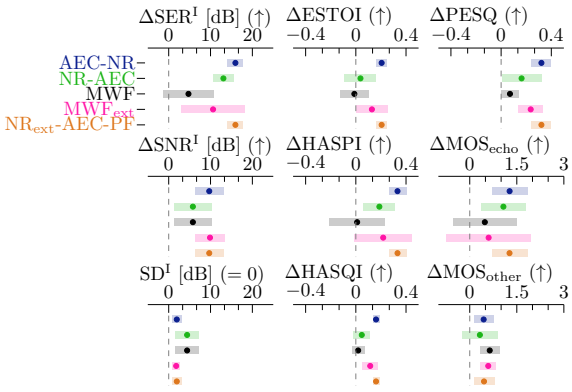
Fig. 4. Performance comparison between the integrated algorithms for Setup-1. The mean performance is represented by the solid line and the standard deviation by the shading. Due to non-stationarities and imperfect correlation matrix estimation the theoretically equivalent integrated algorithms differ in their practical performances. To this end, the AEC-NR and NR_{ext}-AEC-PF generally attain the best performance.

the NR_{ext}-AEC-PF can also be interpreted as an AEC-NR preceded by an NR_{ext}, such that preceding an AEC-NR by an NR_{ext} does not seem to further increase performance.

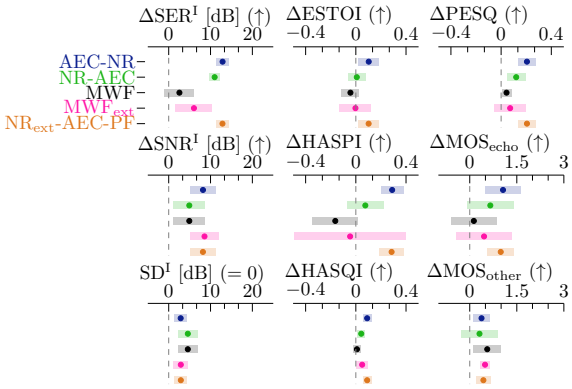
2) Setup-2: Fig. 5 compares the performance of the integrated algorithms for measured, rather than simulated, meeting room impulse responses, for (mild) loudspeaker non-linearities, and for office noise from the DEMAND database. The numerical results for each of the scenarios is also available in Supplementary material Section XII. In Fig. 5(a) ideal VADs are still assumed available. The relative differences between the algorithms is similar as in Fig. 4, and again the AEC-NR and NR_{ext}-AEC-PF generally attain the highest performance. So although the additive map assumption of (4), necessary to prove optimality of the NR_{ext}-AEC-PF as detailed in Section III-C5, is no longer satisfied due to the addition of non-linear processing to model loudspeaker non-linearities, the NR_{ext}-AEC-PF is still shown to be effective. Compared to Fig. 4, the NR-AEC is also closer to the AEC-NR due to there being only one rather than two noise sources and there being two microphones. Nevertheless, the AEC filters in the NR-AEC should still model the prior NR stage, such that the AEC-NR attains superior performance. Finally, the absolute performance of the algorithms is lower in Setup-2 than in Setup-1 due to the STFT window size being smaller than the length of the impulse responses, which degades performance of the frequency-domain implementation [73]. Indeed, when the echo path impulse response becomes smaller than the window size, this echo path cannot be modelled exactly by a

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802



(a) Ideal VADs



(b) Practical (NN-based) VADs

Fig. 5. Performance comparison for Setup-2. The NR_{ext}-AEC-PF is still effective despite the additive map assumption no longer holding true due to loudspeaker non-linearities, and performs similar to the AEC-NR. The mean performance only drops slightly when practical VADs are used.

per-frequency bin estimate [73]. Further increasing the STFT window size, or switching to a time-domain implementation, as discussed in Section II and Section IV, would improve the performance. In Fig. 5(b), the same scenarios are considered as in Fig. 5(a) but with practical NN-based VADs. As the performance drops only slightly with respect to the ideal VAD case, the VAD availability seems justifiable in practice.

In conclusion, the AEC-NR and NR_{ext}-AEC-PF generally attain top performance as the AEC (theoretically) only needs to model the echo paths, and as the NR operates with reduced echo due to the prior application of the AEC (and NR_{ext}).

VIII. Conclusion

In this paper, an integrated approach is proposed for the combined acoustic echo cancellation (AEC) and noise reduction (NR) problem in the general multi-microphone/multi-loudspeaker setup with possible linear dependencies between the loudspeaker and microphone signals. This integrated approach is achieved by selecting a single signal model of either the microphone signal vector or the extended signal vector by stacking microphone and loudspeaker signals, formulating a single mean squared error cost function, and using a common solution strategy. Employing the microphone signal model, an MWF is derived. Similarly, employing the extended signal, an extended MWF (MWF_{ext}) is derived, for which several theoretically equivalent expressions are found that turn out to be interpretable as specific cascade algorithms. More specifically, the MWF_{ext} is shown to be theoretically equivalent to the AEC preceding the NR (AEC-NR), the NR preceding the AEC (NR-AEC), and the extended NR (NR_{ext}) preceding the AEC and a post-filter (PF) (NR_{ext}-AEC-PF). Under rank-deficiency conditions the MWF_{ext} is non-unique, such that this theoretical equivalence amounts to the expressions being specific, not necessarily minimum-norm solutions, for this MWF_{ext}.

Nevertheless, although the MWF_{ext}, AEC-NR, NR-AEC and NR_{ext}-AEC-PF are theoretically equivalent, their practical performances differ due to non-stationarities and imperfect correlation matrix estimation, leading to the AEC-NR and NR_{ext}-AEC-PF generally attaining best overall performance.

Acknowledgements

We thank Prof. J. Kates for providing the MATLAB implementations for the HASPI and HASQI measures.

References

[1] E. Hänsler and G. Schmidt, *Topics in acoustic echo and noise control*. Berlin, Heidelberg: Springer-Verlag, 2006.

[2] N. Fleischhacker, "The tikzpeople package," 2017. [Online]. Available: https://ctan.org/pkg/tikzpeople

[3] F. Albu and H. Kwan, "Fast block exact gauss-seidel pseudo affine projection algorithm," *Electronics lett.*, vol. 40, no. 22, pp. 1451–1453, Oct. 2004.

[4] F. Yang, M. Wu, and J. Yang, "Stereophonic acoustic echo suppression based on Wiener filter in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 227–230, Apr. 2012.

[5] C. Zhang, J. Liu, H. Li, and X. Zhang, "Neural multi-channel and multi-microphone acoustic echo cancellation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1–12, Jun. 2023.

[6] A. Fazel, M. El-Khamy, and J. Lee, "CAD-AEC: Context-aware deep acoustic echo cancellation," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 6919–6923.

[7] E. Seidel, G. Enzner, P. Mowlaee, and T. Fingscheidt, "Neural Kalman filters for acoustic echo cancellation: Comparison of deep neural network-based extensions [special issue on model-based and data-driven audio signal processing]," *IEEE Signal Process. Mag.*, vol. 41, no. 6, pp. 24–38, Nov. 2024.

[8] E. Seidel, P. Mowlaee, and T. Fingscheidt, "Convergence and performance analysis of classical, hybrid, and deep acoustic echo control," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 1–15, May 2024.

[9] J. Benesty, J. Chen, Y. A. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech enhancement*, ser. Signals and Commun. Technol., J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Heidelberg: Springer, 2005, pp. 9–41.

[10] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[11] S. Gannot, "Speech processing utilizing the Kalman filter," *IEEE Instrum. Meas. Mag.*, vol. 15, no. 3, pp. 10–14, Jun. 2012.

[12] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state-of-the-art*. Switzerland: Springer, 2013.

[13] W. Jiang, C. Sun, F. Chen, Y. Leng, and Q. Guo, "A novel skip connection mechanism based on channel-wise cross transformer for speech enhancement," *Multimedia Tools and Appl.*, vol. 83, no. 12, pp. 34 849–34 866, Sep. 2023.

[14] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *Proc. 2021 IEEE Int. Conf. Acoust., Speech, Signal Proces. (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6089–6093.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

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

[16] 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 *Proc. 26th European Signal Process. Conf. (EUSIPCO)*. Rome, Italy: EURASIP, Sep. 2018, pp. 6–10.

[17] M. Luis Valero and E. A. P. Habets, "Low-complexity multi-microphone acoustic echo control in the short-time Fourier transform domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 595–609, Mar. 2019.

[18] F. Albu and H. Kwan, "Combined echo and noise cancellation based on gauss-seidel pseudo affine projection algorithm," in *2004 IEEE Int. Symp. Circuits Syst.*, vol. 3. Vancouver, BC, Canada: IEEE, May 2004, pp. III–505.

[19] R. Martin, S. Gustafsson, and M. Moser, "Acoustic echo cancellation for microphone arrays using switched coefficient vectors," in *Proc. 1997 Int. Workshop Acoustic Echo Noise Control (IWAENC)*. London, United Kingdom: IEEE, Sep. 1997, pp. I–85–I–88.

[20] S. Doclo, M. Moonen, and E. de Clippel, "Combined acoustic echo and noise reduction using GSVD-based optimal filtering," in *Proc. 2000 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2. Istanbul, Turkey: IEEE, Jun. 2000, pp. II1061–II1064.

[21] M. Schrammen, A. Bohlender, S. Kühl, and P. Jax, "Change prediction for low complexity combined beamforming and acoustic echo cancellation," in *Proc. 27th European Signal Process. Conf. (EUSIPCO)*. A Coruna, Spain: EURASIP, Sep. 2019, pp. 1–5.

[22] A. Roebben, T. van Waterschoot, and M. Moonen, "Cascaded noise reduction and acoustic echo cancellation based on an extended noise reduction," in *Proc. 32th European Signal Process. Conf. (EUSIPCO)*. Lyon, France: EURASIP, Aug. 2024, pp. 11–15.

[23] W. Jeannes, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: A survey," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 808–820, Nov. 2001.

[24] W. Herbordt, H. Buchner, and W. Kellermann, "An acoustic human-machine front-end for multimedia applications," *EURASIP J. Advances Signal Process.*, vol. 2003, no. 1, pp. 1–11, Dec. 2003.

[25] T. Burton and R. Goubran, "A new structure for combining echo cancellation and beamforming in changing acoustical environments," in *Proc. 2007 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Honolulu, HI, USA: IEEE, Apr. 2007, pp. I–77–I–80.

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

[27] S. Gustafsson, P. Jax, A. Kamphausen, and P. Vary, "A postfilter for echo and noise reduction avoiding the problem of musical tones," in *Proc. 1999 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2. Phoenix, AZ, USA: IEEE, Mar. 1999, pp. 873–876.

[28] Y. Konforti, I. Cohen, and B. Berdugo, "Multichannel acoustic echo cancellation with beamforming in dynamic environments," *IEEE Open J. Signal Process.*, vol. 4, pp. 479–488, Nov. 2023.

[29] W. Herbordt, W. Kellermann, and S. Nakamura, "Joint optimization of LCMV beamforming and acoustic echo cancellation," in *Proc. 12th European Signal Process. Conf. (EUSIPCO)*. Vienna, Austria: EURASIP, Sep. 2004, pp. 2003–2006.

[30] M. H. Maruo, J. C. M. Bermudez, and L. S. Resende, "On the optimal solutions of beamformer assisted acoustic echo cancellers," in *Proc. 2011 IEEE Statistical Signal Process. Workshop (SSP)*. Nice: IEEE, Jun. 2011, pp. 641–644.

[31] K. Nathwani, "Joint acoustic echo and noise cancellation using spectral domain Kalman filtering in double-talk scenario," in *Proc. 2018 Int. Workshop Acoustic Echo Noise Control (IWAENC)*. Tokyo, Japan: IEEE, Sep. 2018, pp. 1–330.

[32] E. A. P. Habets, I. Cohen, and S. Gannot, "MMSE log-spectral amplitude estimator for multiple interferences," in *Proc. 2006 Int. Workshop Acoustic Echo Noise Control (IWAENC)*. Paris, France: IEEE, Sep. 2006, pp. 1–4.

[33] Y.-S. Park and J.-H. Chang, "Integrated acoustic echo and background noise suppression technique based on soft decision," *EURASIP J. Advances in Signal Process.*, vol. 2012, no. 1, p. 11, Jan. 2012.

[34] E. P. Jayakumar, P. V. M. Shifas, and P. S. Sathidevi, "Integrated acoustic echo and noise suppression in modulation domain," *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 611–621, Sep. 2016.

[35] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. Interspeech 2019*. Graz, Austria: ISCA, Jun. 2019, pp. 15–19.

[36] E. Indenbom, N.-C. Ristea, A. Saabas, T. Parnamaa, J. Guzvin, and R. Cutler, "DeepVQE: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation," in *Proc. Interspeech 2023*. Dublin,Ireland: ISCA, Aug. 2023, pp. 3819–3823.

[37] E. Seidel, P. Mowlaee, and T. Fingscheidt, "Efficient deep acoustic echo suppression with condition-aware training," in *Proc. 2023 IEEE Workshop Appl. Signal Process. Audio and Acoust. (WASPAA)*. New Paltz, New York, U.S.A.: IEEE, Oct. 2023, pp. 1–5.

[38] S. Braun and M. L. Valero, "Task splitting for dnn-based acoustic echo and noise removal," in *Proc. 2022 Int. Workshop Acoustic Echo Noise Control (IWAENC)*. IEEE, Sep. 2022, pp. 1–5.

[39] E. Seidel, P. Mowlaee, and T. Fingscheidt, "Efficient high-performance Bark-scale neural network for residual echo and noise suppression," in *Proc. 2024 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 1386–1390.

[40] S. S. Shetu, N. Kumar Desiraju, J. M. Martinez Aponte, E. A. P. Habets, and E. Mabande, "A hybrid approach for low-complexity joint acoustic echo and noise reduction," in *Proc. 2024 Int. Workshop Acoustic Echo Noise Control (IWAENC)*. Kos Island, Greece: IEEE, Sep. 2024, pp. 349–353.

[41] S. S. Shetu, N. K. Desiraju, W. Mack, and E. A. P. Habets, "Align-ULCNet: Towards low-complexity and robust acoustic echo and noise reduction," in *Proc. 33rd European Signal Process. Conf. (EUSIPCO)*. Palermo, Italy: IEEE, Oct. 2024.

[42] J. Franzen and T. Fingscheidt, "Deep residual echo suppression and noise reduction: A multi-input fcrn approach in a hybrid speech enhancement system," in *Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Singapore,Singapore: IEEE, May 2022, pp. 666–670.

[43] T. Haubner, M. M. Halimeh, A. Brendel, and W. Kellermann, "A synergistic Kalman- and deep postfiltering approach to acoustic echo cancellation," in *Proc. 29th European Signal Process. Conf. (EUSIPCO)*. Dublin, Ireland: EURASIP, Aug. 2021, pp. 990–994.

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

[45] G. Rombouts and M. Moonen, "An integrated approach to acoustic noise and echo cancellation," *Signal Process.*, vol. 85, no. 4, pp. 849–871, Apr. 2005.

[46] S. Ruiz, T. van Waterschoot, and M. Moonen, "Distributed combined acoustic echo cancellation and noise reduction in wireless acoustic sensor and actuator networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 534–547, Jan. 2022.

[47] O. Schwartz and S. Gannot, "Efficient joint bemforming and acoustic echo cancellation structure for conference call scenarios," in *Proc. Interspeech 2024*. Kos Island, Greece: ISCA, Sep. 2024, pp. 167–171.

[48] S. Ruiz, T. van Waterschoot, and M. Moonen, "Cascade Multi-Channel Noise Reduction and Acoustic Feedback Cancellation," in *Proc. 2022 IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 676–680.

[49] A. Roebben, "Github repository: Integrated minimum mean square error algorithms for combined acoustic echo cancellation and noise reduction," https://github.com/Arnout-Roebben/Integrated_AEC_NR, 2024.

[50] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in network and acoustic echo cancellation*, ser. Digital Signal Processing, A. Lacroix and A. Venetsanopoulos, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001.

[51] M. James, "The generalised inverse," *The Math. Gazette*, vol. 62, no. 420, pp. 109–114, Jun. 1978.

[52] R. Penrose, "A generalized inverse for matrices," *Math. Proc. Cambridge Philos. Soc.*, vol. 51, no. 3, pp. 406–413, Jul. 1955.

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

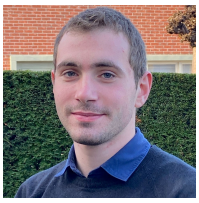
[54] S. Kuo, W. Gan, and P. Asthana, "Integrated noise reduction and acoustic echo cancellation in hands-free systems," in *Proc. 2005 Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*. Hong Kong, China: IEEE, Dec. 2005, pp. 805–808.

[55] S. Shimauchi and S. Makino, "Stereo projection echo canceller with true echo path estimation," in *Proc. 1995 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 5. Detroit, MI, USA: IEEE, May 1995, pp. 3059–3062.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Audio, Speech and Language Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TASLPRO.2025.3648802

[56] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the solution of algebraic eigenvalue problems*. Philadelphia, U.S.A.: SIAM, 2000.
[57] A. Sayed and T. Kailath, "A state-space approach to adaptive RLS filtering," *IEEE Signal Process. Mag.*, vol. 11, no. 3, pp. 18–60, Jul. 1994.
[58] G. H. Golub and C. F. Van Loan, *Matrix computations*, 4th ed., ser. Johns Hopkins studies in the mathematical sciences. Baltimore, Maryland, USA: The Johns Hopkins University Press, 2013.
[59] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 774–786, Apr. 2015.
[60] M. Nilsson, S. D. Ewert, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099, Feb. 1994.
[61] Auditec, "Auditory Tests (Revised), Compact Disc, Auditec," St. Louis, MO, 1997.
[62] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 2009 IEEE Int. Conf. Digital Signal Process. (DSP)*. Santorini, Greece: IEEE, Jul. 2009, pp. 1–5.
[63] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, vol. 19, no. 1, Acoust. Soc. Amer. Montreal, Canada: ISCA, Jun. 2013, p. 035081.
[64] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," Jun. 2021. [Online]. Available: http://arxiv.org/abs/2106.04624
[65] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *Acoust. Soc. Amer.*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.
[66] A. Spriet, "Adaptive filtering techniques for noise reduction and acoustic feedback cancellation in hearing aids," Ph.D. dissertation, KU Leuven, Leuven, Belgium, 2004.
[67] Acoustical Society of America, "ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index," Washington D.C., Washington, USA, 1997.
[68] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
[69] T. S. S. I.-T. International Telecommunication Union, "Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," Geneva, Switzerland, 2007.
[70] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI) Version 2," *Speech Communication*, vol. 131, pp. 35–46, Jul. 2021.
[71] J. Kates and K. Arehart, "The hearing-aid speech quality index (HASQI) Version 2," *J. Audio Eng. Soc.*, vol. 62, no. 3, pp. 99–117, Mar. 2014.
[72] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "AECMOS: A speech quality assessment metric for echo impairment," in *Proc. 2022 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 901–905.
[73] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.



Arnout Roebben obtained a Bachelor of Engineering, and a Master of Electrical Engineering from KU Leuven, Belgium, in 2020 and 2022, respectively. He has been a Ph.D. researcher at KU Leuven since September 2022. His research is a joint venture between the Department of Electrical Engineering and the Department of Neurosciences, and focuses on the design of integrated digital signal processing algorithms.



Toon van Waterschoot received MSc (2001) and PhD (2009) degrees in Electrical Engineering, both from KU Leuven, Belgium, where he is currently a Professor. His research interests are in signal processing, machine learning, and numerical optimization, applied to acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction. He has been the Scientific Coordinator for several major European research projects: the FP7-PEOPLE Marie Curie Initial Training Network "Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS, 2013-2016)", the H2020 ERC Consolidator Grant "The Spatial Dynamics of Room Acoustics (SONORA, 2018-2023)", and the H2020 MSCA European Training Network "Service-Oriented Ubiquitous Network-Driven Sound (SOUNDS, 2021-2025)".

He has been serving as an Associate Editor for the Journal of the Audio Engineering Society and for the EURASIP Journal on Audio, Music, and Speech Processing. He is Executive Director of the European Association for Signal Processing (EURASIP) and Founding Member of the EAA Technical Committee in Audio Signal Processing. He was the General Chair of the 60th AES International Conference in Leuven, Belgium (2016), and has been serving on the Organizing Committee of the European Conference on Computational Optimization (EUCCO 2016), the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017), and the 28th and 29th European Signal Processing Conferences (EUSIPCO 2020 and 2021). He is a member of EURASIP, IEEE, ASA, and AES.



Jan Wouters is a Full Professor at the Dept. of Neurosciences, KU Leuven, Leuven, Belgium, since 2005 and received the Master's and Ph.D. degrees in physics from the University of Leuven, KU Leuven, in 1982 and 1989, respectively. From 1989 to 1992, he was a Postdoctoral Research Fellow with the Belgian National Fund for Scientific Research, Institute of Nuclear Physics, Université de Louvain UCL and at NASA Goddard Space Flight Center, USA. His research interests include hearing sciences, audiology and auditory neural processing, signal processing for cochlear implants and hearing aids. He is on the Editorial Board of the International Journal of Audiology.

Jan Wouters served as President and Secretary-General of the European Federation of Audiological Societies (EFAS), as President of the Belgian Audiological Society (B-Audio), and as Executive Board member of the International Collegium for Rehabilitative Audiology (ICRA) and of the Dutch Acoustical Society (NAG). He is an elected member of the International Collegium for ORL (CORLAS). He is a Honorary Member of the Deutsche Gesellschaft für Audiologie DGA (2020) and received the EFAS Lifetime Achievement Award (2021).



Marc Moonen is a Full Professor with the Electrical Engineering Department, KU Leuven, Belgium. He is a Fellow of the IEEE (2007), a Fellow of EURASIP (European Association for Signal Processing, 2018) and the 2024 recipient of the IEEE Signal Processing Society Claude Shannon-Harry Nyquist Technical Achievement Award. He was President of EURASIP (2007-2008 and 2011-2012), Vice-President for Publications of the IEEE Signal Processing Society (2021-2023) and is currently President-Elect of the IEEE Signal Processing Society (2026-2027). He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003-2005), Area Editor for Feature Articles in the IEEE Signal Processing Magazine (2012-2014), and has been a member of the Editorial Board of Signal Processing, IEEE Transactions on Circuits and Systems-II, IEEE Signal Processing Magazine, EURASIP Journal on Wireless Communications and Networking, and EURASIP Journal on Advances in Signal Processing.

Authorized licensed use limited to: KU Leuven University Library. Downloaded on January 05,2026 at 10:20:24 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.