KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Authors: Arnout Roebben, Toon van Waterschoot, and Marc Moonen Title: A Comparative Analysis of Generalised Echo and Interference Cancelling and Extended Multichannel Wiener Filtering for Combined Noise Reduction and Acoustic Echo Cancellation Conference: Proc. 2025 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP) Pages: 1-5 doi: 10.1109/ICASSP49660.2025.10890146
Archived version	☐ Submitted version ■ Accepted (Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher) ☐ Published ☐ Published (Open Access)
Acknowledgements	In this manuscript, the following funding is acknowledged: ☐ EU H22020 ☐ EU Horizon Europe ■ FWO ■ KULeuven Internal Funds ☐ VLAIO ☐ AI ☐ Other
Published version	https://ieeexplore.ieee.org/document/10890146
Author contact	arnout.roebben@esat.kuleuven.be



A Comparative Analysis of Generalised Echo and Interference Cancelling and Extended Multichannel Wiener Filtering for Combined Noise Reduction and Acoustic Echo Cancellation

Arnout Roebben, Toon van Waterschoot, and Marc Moonen
Department of Electrical Engineering (ESAT)
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
KU Leuven
Leuven, Belgium
{arnout.roebben, toon.vanwaterschoot, marc.moonen}@kuleuven.be

***Abstract*—Two algorithms for combined acoustic echo cancellation (AEC) and noise reduction (NR) are analysed, namely the generalised echo and interference canceller (GEIC) and the extended multichannel Wiener filter (MWF_{ext}). Previously, these algorithms have been examined for linear echo paths, and assuming access to voice activity detectors (VADs) that separately detect desired speech and echo activity. However, algorithms implementing VADs may introduce detection errors. Therefore, in this paper, the previous analyses are extended by 1) modelling general nonlinear echo paths by means of the generalised Bussgang decomposition, and 2) modelling VAD error effects in each specific algorithm, thereby also allowing to model specific VAD assumptions. It is found and verified with simulations that, generally, the MWF_{ext} achieves a higher NR performance, while the GEIC achieves a more robust AEC performance.**

***Index Terms*—Acoustic echo cancellation (AEC), Noise reduction (NR), Extended multichannel Wiener filter (MWF_{ext}), Generalised echo and interference canceller (GEIC)**

I. Introduction

In various speech recording applications, next to desired speech, unwanted noise and acoustic echo are recorded. The noise originates from within the room (the near-end room) while the echo originates from recordings in another room (the far-end room) that are played by loudspeakers in the near-end room [1]. To reduce the interferers, an algorithm for combined acoustic echo cancellation (AEC) and noise reduction (NR) is required [1]–[4]. Next to algorithms that cascade separate AEC and NR algorithms [5], [6], integrated algorithms can be derived from a single cost function, such as the generalised echo and interference canceller (GEIC) [2] and the extended multichannel Wiener filter (MWF_{ext}) [3], [4].

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).

©2025 IEEE

The GEIC has been analysed in [7], [8], and the MWF_{ext} in [3]. These studies, however, consider only linear echo paths, be it that in [7] undermodelling is also considered. In addition, the algorithms assume access to voice activity detectors (VADs) that separately detect desired speech and echo activity. However, algorithms implementing VADs may introduce detection errors [9], [10].

In this paper, the previous analyses are extended by 1) modelling general nonlinear echo paths by means of the generalised Bussgang decomposition [11], and 2) modelling VAD error effects in each specific algorithm. This analysis framework also allows to study specific VAD assumptions, e.g., considering a permanent doubletalk scenario where the desired speech and echo are simultaneously active [8]. It is found that, generally, the MWF_{ext} achieves a higher NR performance, while the GEIC achieves a more robust AEC performance. Coding examples are available [12].

II. Signal model

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

An M-microphone/L-loudspeaker scenario is considered, where $m_i(z)$ represents the microphone signal in microphone $i = \{1, ..., M\}$, and $l_j(z)$ represents the loudspeaker signal in loudspeaker $j = \{1, ..., L\}$. The signals $m_i(z)$ can then be collected in a microphone signal vector $\mathbf{m}(z) \in \mathbb{C}^{M\times 1}$ as

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

which consists of a desired speech signal vector $\mathbf{s}(z)$, noise signal vector $\mathbf{n}(z)$ and echo signal vector $\mathbf{e}(z)$ as

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

Herein, $\mathbf{s}(z)$, $\mathbf{n}(z)$ and $\mathbf{e}(z)$ are assumed mutually uncorrelated. The echo signal vector $\mathbf{e}(z) = F(\mathbf{l}(z))$ originates from the loudspeaker signal vector $\mathbf{l}(z) \in \mathbb{C}^{L\times 1}$

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

with a map $F(.) : \mathbb{C}^{L\times 1} \rightarrow \mathbb{C}^{M\times 1}$ representing the echo path from loudspeakers to microphones. Further, an extended microphone signal vector $\tilde{\mathbf{m}}(z) \in \mathbb{C}^{(M+L)\times 1}$ is defined by stacking $\mathbf{m}(z)$ and $\mathbf{l}(z)$ as

$$\begin{aligned} \tilde{\mathbf{m}}(z) &= \begin{bmatrix} \mathbf{m}(z)^\top & \mathbf{l}(z)^\top \end{bmatrix}^\top && \text{(4a)} \\ &= \tilde{\mathbf{s}}(z) + \tilde{\mathbf{n}}(z) + \tilde{\mathbf{e}}(z) && \text{(4b)} \\ &= \begin{bmatrix} \mathbf{s}(z) \\ \mathbf{0}_{L\times 1} \end{bmatrix} + \begin{bmatrix} \mathbf{n}(z) \\ \mathbf{0}_{L\times 1} \end{bmatrix} + \begin{bmatrix} \mathbf{e}(z) \\ \mathbf{l}(z) \end{bmatrix}. && \text{(4c)} \end{aligned}$$

For a single desired speech source, $\mathbf{s}(z)$ can be represented as $\mathbf{s}(z) = \mathbf{h}(z)s_r(z)$ with $\mathbf{h}(z) = \begin{bmatrix} h_1(z)/h_r(z) & ... & h_M(z)/h_r(z) \end{bmatrix}^\top \in \mathbb{C}^{M\times 1}$ collecting the relative transfer functions $\frac{h_i(z)}{h_r(z)}$ with $h_i(z)$ the transfer function from source $s(z)$ to microphone i, and $s_r(z) = h_r(z)s(z)$ the desired speech in reference microphone r. Similarly, $\tilde{\mathbf{s}}(z) = \tilde{\mathbf{h}}(z)s_r(z)$ with $\tilde{\mathbf{h}}(z) = \begin{bmatrix} \mathbf{h}(z)^\top & \mathbf{0}_{L\times 1}^\top \end{bmatrix}^\top$. From here on, z-indices are omitted for conciseness.

While the noise is assumed to be always-on, the desired speech and echo are on-off with near-end and far-end speaker pauses. This desired speech and echo activity are independent, resulting in four regimes, during which the algorithms estimate the required signal statistics [1]–[4]. Thus, separate voice activity detectors (VADs) [9], [10] are required for the desired speech (VAD_s) and echo (VAD_e) to distinguish desired speech activity ($\text{VAD}_s = 1$) and inactivity ($\text{VAD}_s = 0$), and echo activity ($\text{VAD}_e = 1$) and inactivity ($\text{VAD}_e = 0$).

With $\mathbb{E}_k\{.\}$ the expected value in periods where $\mathbf{k} \in \{\mathbf{s}, \mathbf{e}\}$ is active (i.e., $\text{VAD}_k = 1$), and assuming (short-term) stationarity and ergodicity during the activity (e.g., modelling a batch of data), the desired speech and echo correlation matrix can be computed as $\mathbb{E}_s\{\mathbf{s}\mathbf{s}^H\} = R_{ss} \in \mathbb{C}^{M\times M}$, and $\mathbb{E}_e\{\mathbf{e}\mathbf{e}^H\} = R_{ee}$ respectively. As the noise is always-on, $\mathbb{E}_s\{\mathbf{n}\mathbf{n}^H\} = \mathbb{E}_e\{\mathbf{n}\mathbf{n}^H\} = \mathbb{E}\{\mathbf{n}\mathbf{n}^H\} = R_{nn}$. Similarly, the extended desired speech, echo and noise correlation matrix are respectively defined as

$$\mathbb{E}_s\{\tilde{s}\tilde{s}^H\} = R_{\tilde{s}\tilde{s}} = \begin{bmatrix} R_{ss} & 0_{M\times L} \\ 0_{L\times M} & 0_{L\times L} \end{bmatrix} \tag{5a}$$

$$\mathbb{E}_e\{\tilde{e}\tilde{e}^H\} = R_{\tilde{e}\tilde{e}} = \begin{bmatrix} R_{ee} & R_{el} \\ R_{le} & R_{ll} \end{bmatrix} \tag{5b}$$

$$\mathbb{E}_s\{\tilde{n}\tilde{n}^H\} = \mathbb{E}_e\{\tilde{n}\tilde{n}^H\} = R_{\tilde{n}\tilde{n}} = \begin{bmatrix} R_{nn} & 0_{M\times L} \\ 0_{L\times M} & 0_{L\times L} \end{bmatrix}, \tag{5c}$$

with $\mathbb{E}_e\{\mathbf{l}\mathbf{l}^H\} = R_{ll} \in \mathbb{C}^{L\times L}$ and $\mathbb{E}_e\{\mathbf{e}\mathbf{l}^H\} = R_{el} \in \mathbb{C}^{M\times L}$ the loudspeaker and echo-loudspeaker correlation matrix.

However, VAD algorithms may introduce detection errors [9], [10], e.g., erroneously detecting desired speech as echo since both signals are speech signals with similar characteristics. Therefore, the (extended) microphone correlation matrices below will be re-defined by considering VAD errors. Furthermore, this framework will allow to study the effect of specific VAD assumptions, e.g., the effect of a (possibly error-free) VAD on the microphone signals only. This VAD is then only able to distinguish between noise-only periods and periods of desired speech or echo activity, or both, as both signals are speech signals with similar characteristics. Similarly, a scenario with permanent doubletalk can be studied, where the desired speech and echo are simultaneously active, such that VAD_s and VAD_e coincide [8].

As will be seen in Section III-B, the GEIC requires access to the regime $\{\text{VAD}_s = 0 \wedge \text{VAD}_e = 1\}$, corresponding to a correlation matrix $R_{\tilde{n}\tilde{n}} + R_{\tilde{e}\tilde{e}}$. However, with VAD errors, this correlation matrix will be replaced by $R_{\tilde{m}\tilde{m}}^{\{\alpha_s,\alpha_e\}}$ with

$$R_{\tilde{m}\tilde{m}}^{\{\alpha_s,\alpha_e\}} = \alpha_s R_{\tilde{s}\tilde{s}} + R_{\tilde{n}\tilde{n}} + \alpha_e R_{\tilde{e}\tilde{e}}. \tag{6}$$

Herein, $\alpha_s, \alpha_e \in [0, 1]$ are scaling factors arising when calculating the average correlation matrix. Indeed, due to VAD errors, in some periods the echo (desired speech) will be inactive (active), thereby reducing (increasing) its contribution in the average correlation matrix. Without VAD errors, $\alpha_s = 0$ and $\alpha_e = 1$. Considering an error-free VAD in the permanent doubletalk scenario yields $\alpha_s = \alpha_e = 1$.

Similarly, it will be seen in Section III-C that the MWF_{ext} requires access to the regimes $\{\text{VAD}_s = 1 \wedge \text{VAD}_e = 1\}$ and $\{\text{VAD}_s = 0 \wedge \text{VAD}_e = 1\}$, which correspond to the correlation matrices $R_{\tilde{s}\tilde{s}} + R_{\tilde{n}\tilde{n}} + R_{\tilde{e}\tilde{e}}$ and $R_{\tilde{n}\tilde{n}} + R_{\tilde{e}\tilde{e}}$ respectively. However, with VAD errors, these correlation matrices will be replaced by $R_{\tilde{m}\tilde{m}}^{\{\beta_s,\beta_e\}}$ and $R_{\tilde{m}\tilde{m}}^{\{\gamma_s,\gamma_e\}}$ respectively with

$$R_{\tilde{m}\tilde{m}}^{\{\beta_s,\beta_e\}} = \beta_s R_{\tilde{s}\tilde{s}} + R_{\tilde{n}\tilde{n}} + \beta_e R_{\tilde{e}\tilde{e}} \tag{7a}$$

$$R_{\tilde{m}\tilde{m}}^{\{\gamma_s,\gamma_e\}} = \gamma_s R_{\tilde{s}\tilde{s}} + R_{\tilde{n}\tilde{n}} + \gamma_e R_{\tilde{e}\tilde{e}}. \tag{7b}$$

Herein, $\beta_s, \beta_e, \gamma_s, \gamma_e \in [0, 1]$ are again scaling factors. Without VAD errors, $\beta_s = \beta_e = \gamma_e = 1$ and $\gamma_s = 0$. Considering an (error-free) VAD only operating on the microphone signals, which is then only able to distinghuish between noise-only periods and periods of desired speech or echo activity, or both, $\gamma_s = \gamma_e = 0$, and considering an (error-free) VAD in the permanent doubletalk scenario, $\beta_s = \beta_e = 1$ and $\gamma_s = \gamma_e = 0$.

R_{ss} is assumed rank 1, modelling a single desired speaker, and R_{nn} and R_{ll} are assumed full rank.

III. Filter design

A. Generalised Bussgang decomposition

To analyse a general nonlinear echo path map $F(.)$, the generalised Bussgang decomposition is applied to decompose the echo signal vector $\mathbf{e}$ into two uncorrelated components, $\mathbf{e}_{\text{lin}} \in \mathbb{C}^{M\times 1}$ and $\mathbf{e}_{\text{res}} \in \mathbb{C}^{M\times 1}$ [11]. Herein, $\mathbf{e}_{\text{lin}} = F_{\text{lin}}^H \mathbf{l}$ represents the linear echo component, with $F_{\text{lin}} = R_{ll}^{-1} R_{le}$ the minimum mean squared error optimal estimate of the echo path [13], and $\mathbf{e}_{\text{res}} = \mathbf{e} - F_{\text{lin}}^H \mathbf{l}$ represents the residual echo component after subtracting $\mathbf{e}_{\text{lin}}$ from $\mathbf{e}$. Indeed, $\mathbf{e}_{\text{lin}}$ and $\mathbf{e}_{\text{res}}$ are uncorrelated as [11]

$$\mathbb{E}\left\{\mathbf{e}_{\text{res}}\mathbf{l}^H\right\} = 0_{M\times L} \quad \wedge \quad \mathbb{E}\left\{\mathbf{e}_{\text{res}}\mathbf{e}_{\text{lin}}^H\right\} = 0_{M\times M}. \tag{8}$$

Therefore, $R_{\tilde{e}\tilde{e}}$ can be split into two uncorrelated components:

$$R_{\tilde{e}\tilde{e}} = R_{\tilde{e}_{\text{lin}}\tilde{e}_{\text{lin}}} + R_{\tilde{e}_{\text{res}}\tilde{e}_{\text{res}}} \tag{9a}$$

$$= \begin{bmatrix} F_{\text{lin}}^H R_{ll} F_{\text{lin}} & F_{\text{lin}}^H R_{ll} \\ R_{ll} F_{\text{lin}} & R_{ll} \end{bmatrix} + \begin{bmatrix} R_{e_{\text{res}}e_{\text{res}}} & 0_{M\times L} \\ 0_{L\times M} & 0_{L\times L} \end{bmatrix}. \tag{9b}$$

B. Generalised echo and interference canceller (GEIC)

The GEIC [2] aims at designing a filter $\tilde{\mathbf{w}}_{\text{GEIC}} \in \mathbb{C}^{(M+L)\times 1}$ that minimises the filter output power while preserving the desired speech in the reference microphone

$$\begin{aligned} \tilde{\mathbf{w}}_{\text{GEIC}} = \underset{\tilde{\mathbf{w}}}{\text{argmin}} \quad & \tilde{\mathbf{w}}^H R_{\tilde{m}\tilde{m}}^{\{\alpha_s,\alpha_e\}} \tilde{\mathbf{w}} \\ \text{s.t.} \quad & \tilde{\mathbf{w}}^H \tilde{\mathbf{h}} = 1, \end{aligned} \tag{10}$$

of which the solution equals [2]:

$$\tilde{\mathbf{w}}_{\text{GEIC}} = \begin{bmatrix} \mathbf{w}_c - B\mathbf{w}_a \\ -\mathbf{a} \end{bmatrix}. \tag{11}$$

Herein, a quiescent beamformer $\mathbf{w}_c = \frac{\mathbf{h}}{\mathbf{h}^H\mathbf{h}} \in \mathbb{C}^{M\times 1}$ is defined, as well as a blocking matrix $B \in \mathbb{C}^{M\times(M-1)}$ with $\mathbf{h}^H B = \mathbf{0}_{(M-1)\times 1}^\top$, and $\mathbf{w}_a \in \mathbb{C}^{(M-1)\times 1}$ and $\mathbf{a} \in \mathbb{C}^{L\times 1}$ are jointly optimised by means of an adaptive filter procedure, which, using [14, (2.3)], converges to [2], [7]

$$\begin{bmatrix} \mathbf{w}_a \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} S^{-1}(B^H R_{mm}^{\{\alpha_s,\alpha_e\}} \mathbf{w}_c - \alpha_e B^H R_{el} R_{ll}^{-1} R_{le} \mathbf{w}_c) \\ -R_{ll}^{-1} R_{le} B \mathbf{w}_a + R_{ll}^{-1} R_{le} \mathbf{w}_c \end{bmatrix}, \tag{12}$$

with $S = B^H R_{mm}^{\{\alpha_s,\alpha_e\}} B - \alpha_e B^H R_{el} R_{ll}^{-1} R_{le} B$. Referring to the generalised Bussgang decomposition (Section III-A), $\mathbf{a}$ aims at modelling the linear component of the overall echo path from the loudspeakers to the output of the GEIC, consisting of a combination of the optimal mean squared error estimate of the echo path $F_{\text{lin}} = R_{ll}^{-1} R_{le}$, with the parallel filters $\mathbf{w}_c$ and $-B\mathbf{w}_a$. Thus, $\mathbf{a}$ allows for a perfect cancellation of $\mathbf{e}_{\text{lin}}$ regardless of α_e, i.e., $\tilde{\mathbf{w}}_{\text{GEIC}}^H \tilde{\mathbf{e}}_{\text{lin}} = \tilde{\mathbf{w}}_{\text{GEIC}}^H \begin{bmatrix} \mathbf{l}^H F_{\text{lin}} & \mathbf{l}^H \end{bmatrix}^H = 0$. On the other hand, $\mathbf{w}_a$ aims at estimating the noise and residual echo after applying $\mathbf{w}_c$, from the noise and residual echo after applying B. Indeed, $\mathbf{w}_a$ (12) can be simplified as

$$\mathbf{w}_a = (B^H(R_{nn} + \alpha_e R_{e_{\text{res}}e_{\text{res}}})B)^{-1}(B^H(R_{nn} + \alpha_e R_{e_{\text{res}}e_{\text{res}}})\mathbf{w}_c), \tag{13}$$

such that the $\mathbf{w}_a$ reduces noise and residual echo. When the echo is linear, i.e., $\mathbf{e}_{\text{res}} = \mathbf{0}_{M\times 1}$, $\mathbf{a}$ removes the entire echo, and $\mathbf{w}_a$ solely reduces the noise, which is consistent with [8].

An imperfect blocking matrix ($\mathbf{h}^H B \neq \mathbf{0}_{(M-1)\times 1}^\top$) leads to suppression of the desired speech, such that $\mathbf{w}_a$ should be updated ideally under desired speech inactivity and echo activity. Indeed, as α_s decreases due to fewer VAD errors, less weight is put on reducing the desired speech.

C. Extended multichannel Wiener filter (MWF$_{ext}$)

The MWF$_{\text{ext}}$ as proposed by [3], [4] aims at designing a filter $\tilde{\mathbf{w}}_{\text{MWF}_{\text{ext}}} \in \mathbb{C}^{(M+L)\times 1}$ to retain the desired speech in the reference microphone r, while reducing noise and echo:

$$\tilde{\mathbf{w}}_{\text{MWF}_{\text{ext}}} = \underset{\tilde{\mathbf{w}}}{\text{argmin}}\, \mathbb{E}\left\{ \left\| s_r - \tilde{\mathbf{w}}^H \tilde{\mathbf{m}} \right\|_2^2 \right\}, \tag{14}$$

which leads to:

$$\tilde{\mathbf{w}}_{\text{MWF}_{\text{ext}}} = R_{\tilde{m}\tilde{m}}^{-1} R_{\tilde{s}\tilde{s}} \tilde{\mathbf{t}}_r, \tag{15}$$

where $\tilde{\mathbf{t}}_r \in \mathbb{C}^{(M+L)\times 1}$ corresponds to a unit vector with all zeros except at position r. Without VAD errors, $R_{\tilde{s}\tilde{s}}$ can be computed by subtracting the correlation matrix recorded when $\{\text{VAD}_s = 0 \wedge \text{VAD}_e = 1\}$ from the correlation matrix recorded when $\{\text{VAD}_s = 1 \wedge \text{VAD}_e = 1\}$ [15]. However, with VAD errors, this subtraction of (7b) from (7a) yields $(\beta_s - \gamma_s)R_{\tilde{s}\tilde{s}} + (\beta_e - \gamma_e)R_{\tilde{e}_{\text{lin}}\tilde{e}_{\text{lin}}} + (\beta_e - \gamma_e)R_{\tilde{e}_{\text{res}}\tilde{e}_{\text{res}}}$ rather than $R_{\tilde{s}\tilde{s}}$.

It is shown that using a generalised eigenvalue decomposition (GEVD), remarkably, the linear echo component can be removed entirely despite it being observed together with the desired speech, while the residual echo cannot be removed entirely. Thereby the GEVD procedure of [3] proposed for linear echo paths is extended to general nonlinear echo paths. The GEVD of $\{R_{\tilde{m}\tilde{m}}^{\{\beta_s,\beta_e\}}, R_{\tilde{m}\tilde{m}}^{\{\gamma_s,\gamma_e\}}\}$ is defined as

$$R_{\tilde{m}\tilde{m}}^{\{\beta_s,\beta_e\}} = Q\text{diag}(\lambda_{\tilde{m}_1}^{\{\beta_s,\beta_e\}}, ..., \lambda_{\tilde{m}_{(M+L)}}^{\{\beta_s,\beta_e\}}) Q^H \tag{16a}$$

$$R_{\tilde{m}\tilde{m}}^{\{\gamma_s,\gamma_e\}} = Q\text{diag}(\lambda_{\tilde{m}_1}^{\{\gamma_s,\gamma_e\}}, ..., \lambda_{\tilde{m}_{(M+L)}}^{\{\gamma_s,\gamma_e\}}) Q^H, \tag{16b}$$

with $Q \in \mathbb{C}^{(M+L)\times(M+L)}$ collecting the generalised eigenvectors in its columns, $\lambda_{\tilde{m}_p}^{\{\beta_s,\beta_e\}}$ and $\lambda_{\tilde{m}_p}^{\{\gamma_s,\gamma_e\}}$ ($p = \{1, ..., M+L\}$) defining the generalised eigenvalues, and diag(.) representing the operation to create a diagonal matrix. $R_{\tilde{s}\tilde{s}}$, $R_{\tilde{n}\tilde{n}}$ and $R_{\tilde{e}_{\text{res}}\tilde{e}_{\text{res}}}$ contain a zero structure related to the loudspeaker signals, since $\mathbf{s}$, $\mathbf{n}$ and $\mathbf{e}_{\text{res}}$ are uncorrelated with $\mathbf{l}$. However, $R_{\tilde{e}_{\text{lin}}\tilde{e}_{\text{lin}}}$ lacks this zero structure due to the correlation of $\mathbf{e}^{lin}$ with $\mathbf{l}$. Consequently, Q reflects this structure:

$$Q = \left[\begin{array}{c|c} Q_1 & Q_2 \\ \hline 0_{L\times R} & \end{array} \right], \tag{17}$$

with $R = \text{rank}((\beta_s - \gamma_s)R_{\tilde{s}\tilde{s}} + (\beta_e - \gamma_e)R_{\tilde{e}_{\text{res}}\tilde{e}_{\text{res}}})$. If $R = 1$, $\begin{bmatrix} Q_1^\top & \mathbf{0}_{L\times 1}^\top \end{bmatrix}^\top$ models $\tilde{\mathbf{s}}$. If $R = M$, $\begin{bmatrix} Q_1^\top & 0_{L\times R}^\top \end{bmatrix}^\top$ is related to $\tilde{\mathbf{s}}$, $\tilde{\mathbf{n}}$, and $\tilde{\mathbf{e}}_{\text{res}}$, and Q_2 is related to $\tilde{\mathbf{e}}_{\text{lin}}$. Explicitly setting the generalised eigenvalues related to Q_2, and thus related to $R_{\tilde{e}_{\text{lin}}\tilde{e}_{\text{lin}}}$, to 0, the component corresponding to $\tilde{\mathbf{e}}_{\text{lin}}$ can still be removed regardless of β_e and γ_e. The estimate for the desired speech correlation matrix $\hat{R}_{\tilde{s}\tilde{s}}$ is then computed as

$$\begin{aligned} \hat{R}_{\tilde{s}\tilde{s}} &= Q\text{diag}(\lambda_{\tilde{m}_1}^{\{\beta_s,\beta_e\}} - \lambda_{\tilde{m}_1}^{\{\gamma_s,\gamma_e\}}, ..., \lambda_{\tilde{m}_M}^{\{\beta_s,\beta_e\}} - \lambda_{\tilde{m}_M}^{\{\gamma_s,\gamma_e\}}, 0, ..., 0) Q^H \\ &= (\beta_s - \gamma_s)R_{\tilde{s}\tilde{s}} + (\beta_e - \gamma_e)R_{\tilde{e}_{\text{res}}\tilde{e}_{\text{res}}}. \end{aligned} \tag{18}$$

Plugging $\hat{R}_{\tilde{s}\tilde{s}}$ into (15) leads to the following realisable filter:

$$\hat{\tilde{\mathbf{w}}}_{\text{MWF}_{\text{ext}}} = R_{\tilde{m}\tilde{m}}^{-1}((\beta_s - \gamma_s)R_{\tilde{s}\tilde{s}} + (\beta_e - \gamma_e)R_{\tilde{e}_{\text{res}}\tilde{e}_{\text{res}}})\, \tilde{\mathbf{t}}_r, \tag{19}$$

thereby reducing the noise and the linear echo component, but not the residual echo. Indeed, a weighted average of the desired speech and residual echo is being reconstructed, and $\hat{R}_{\tilde{s}\tilde{s}}$ can even be negative definite due to the VAD errors. Explicitly setting the rank of $\hat{R}_{\tilde{s}\tilde{s}}$ to its theoretical value of 1 is realised by sorting the ratio of generalised eigenvalues,

$$\left\{ \lambda_{\tilde{m}_1}^{\{\beta_s,\beta_e\}} / \lambda_{\tilde{m}_1}^{\{\gamma_s,\gamma_e\}}, ..., \lambda_{\tilde{m}_M}^{\{\beta_s,\beta_e\}} / \lambda_{\tilde{m}_M}^{\{\gamma_s,\gamma_e\}} \right\}, \tag{20}$$

according to an increased ratio and only keeping this largest ratio [3], [16]. Although this further reduces the residual echo as

eigenvalue modes containing residual echo are removed, partial desired speech cancellation or residual echo reconstruction might occur. Indeed, the matrix pencil $\{R_{\tilde{m}\tilde{m}}^{\{\beta_s,\beta_e\}}, R_{\tilde{m}\tilde{m}}^{\{\gamma_s,\gamma_e\}}\}$ is jointly diagonalised, but $R_{\tilde{s}\tilde{s}}$ not necessarily.

When the spatial structure of $R_{\tilde{e}_{\text{lin}}\tilde{e}_{\text{lin}}}$ would differ between both sides of $\{R_{\tilde{m}\tilde{m}}^{\{\beta_s,\beta_e\}}, R_{\tilde{m}\tilde{m}}^{\{\gamma_s,\gamma_e\}}\}$, e.g., due to changing echo paths, the zero-structure in Q would however be destroyed.

D. Relation between GEIC and MWF$_{ext}$

When $R_{\tilde{s}\tilde{s}}$ is rank-1, (19) can be decomposed using the ground-truth $\tilde{\mathbf{h}}$ according to a similar proof as in [17] as

$$\hat{\tilde{\mathbf{w}}}_{\text{MWF}_{\text{ext}}} = \tilde{\mathbf{w}}_{\text{GEIC}}^{\{\beta_s,\beta_e\}} \frac{(\beta_s-\gamma_s)P_{y_{\tilde{s}}y_{\tilde{s}}}}{\beta_s P_{y_{\tilde{s}}y_{\tilde{s}}} + P_{y_{\tilde{n}}y_{\tilde{n}}} + \beta_e P_{y_{\tilde{e}_{\text{res}}}y_{\tilde{e}_{\text{res}}}}} + R_{\tilde{m}\tilde{m}}^{-1}\left((\beta_e-\gamma_e)R_{\tilde{e}_{\text{res}}\tilde{e}_{\text{res}}}\right)\tilde{\mathbf{t}}_r. \quad (21)$$

Herein, $\tilde{\mathbf{w}}_{\text{GEIC}}^{\{\beta_s,\beta_e\}}$ is a GEIC computed using (7a), $P_{y_{\tilde{k}}y_{\tilde{k}}} = \tilde{\mathbf{w}}_{\text{GEIC}}^{\{\beta_s,\beta_e\}^H} R_{\tilde{k}\tilde{k}} \tilde{\mathbf{w}}_{\text{GEIC}}^{\{\beta_s,\beta_e\}}$ with $\tilde{\mathbf{k}} \in \{\tilde{\mathbf{s}}, \tilde{\mathbf{n}}, \tilde{\mathbf{e}}_{\text{res}}, \tilde{\mathbf{e}}_{\text{lin}}\}$, and $P_{y_{\tilde{e}_{\text{lin}}}y_{\tilde{e}_{\text{lin}}}} = 0$. Thus, the MWF$_{\text{ext}}$ can be interpreted as a cascade of a GEIC and a single-channel postfilter, parallel to a filter providing the minimum mean squared error optimal residual echo estimate.

Applying the rank-1 approximation to $\hat{R}_{\tilde{s}\tilde{s}}$ (Section III-C), i.e., $\hat{R}_{\tilde{s}\tilde{s}} = \left(\lambda_{\tilde{m}_1}^{\{\beta_s,\beta_e\}} - \lambda_{\tilde{m}_1}^{\{\gamma_s,\gamma_e\}}\right)\tilde{\mathbf{q}}\tilde{\mathbf{q}}^H$ with $\tilde{\mathbf{q}}$ the generalised eigenvector in Q corresponding to the largest ratio in (20), corresponds to replacing the ground-truth $\tilde{\mathbf{h}}$ by its estimate $\hat{\tilde{\mathbf{h}}}_{\text{GEVD}} = \tilde{\mathbf{q}}/\tilde{\mathbf{q}}(1) \in \mathbb{C}^{(M+L)\times 1}$, in which the last L elements are zero (Section III-C) [18]. Estimating the correlation matrix of the signal to suppress as $\hat{R}_{\tilde{i}\tilde{i}} = R_{\tilde{m}\tilde{m}}^{\{\beta_s,\beta_e\}} - \hat{R}_{\tilde{s}\tilde{s}}$, (19), can be decomposed using a similar proof as in [17] as,

$$\hat{\tilde{\mathbf{w}}}_{\text{MWF}_{\text{ext}}} = \tilde{\mathbf{w}}_{\text{GEIC,GEVD}}^{\{\beta_s,\beta_e\}} \frac{P_{y_{\tilde{s}}y_{\tilde{s}},\text{GEVD}}}{P_{y_{\tilde{s}}y_{\tilde{s}},\text{GEVD}} + P_{y_{\tilde{i}}y_{\tilde{i}},\text{GEVD}}}. \quad (22)$$

Herein, $\tilde{\mathbf{w}}_{\text{GEIC,GEVD}}^{\{\beta_s,\beta_e\}}$ is a GEIC computed using (7a) and $\hat{\tilde{\mathbf{h}}}_{\text{GEVD}}$, and $P_{y_{\tilde{k}}y_{\tilde{k}},\text{GEVD}} = \tilde{\mathbf{w}}_{\text{GEIC,GEVD}}^{\{\beta_s,\beta_e\}^H} \hat{R}_{\tilde{k}\tilde{k}} \tilde{\mathbf{w}}_{\text{GEIC,GEVD}}^{\{\beta_s,\beta_e\}}$ with $\tilde{\mathbf{k}} \in \{\tilde{\mathbf{s}}, \tilde{\mathbf{i}}\}$ determining the succeeding single-channel postfilter.

IV. Simulations

Five scenarios in a $5\,\text{m} \times 5\,\text{m} \times 3\,\text{m}$ room with $M = 2$ microphones at $\begin{bmatrix}2 & 1.9 & 1\end{bmatrix}\text{m}$ and $\begin{bmatrix}2 & 1.8 & 1\end{bmatrix}\text{m}$, $L = 2$ loudspeakers, 1 babble noise source [19], and 1 desired speech source are studied, with varying positions for the desired speech source, the loudspeakers and the noise source. The desired speech originates from sentences of the Hearing in Noise Test (HINT) database concatenated with 5 s of silence [20]. The loudspeaker signals contain a mixture of HINT sentences and white noise at a power ratio of 0 dB. All signals are 10 s long. Impulse responses of length 128 samples at a sampling rate of 16 kHz are generated using the randomised image method with reflection coefficient 0.15 and randomised distances of 0.13 m [21]. The input signal-to-noise ratio (SNR) and signal-to-echo ratio are set to 5 dB. The echo path is modelled either as a linear path characterised by its impulse response, or as a Hammerstein model, i.e., a static nonlinearity $(.)^3$ followed by a linear path. To illustrate the analysis framework, a permanent doubletalk scenario with error-free VAD is considered with $\alpha_s = \alpha_e = \beta_s = \beta_e = 1$ and $\gamma_s = \gamma_e = 0$.

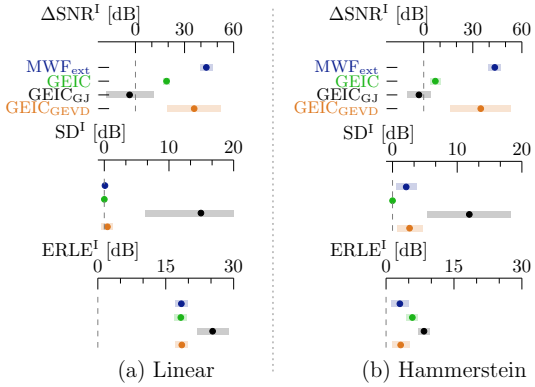

Fig. 1. Performance comparison, where dots show the mean performance across scenarios, and shading the standard deviation. The MWF$_{\text{ext}}$ achieves higher NR, while the GEIC appears more robust regarding the residual echo.

The filters are implemented in the short-time Fourier- (STFT) domain using a square-root Hann window and window shift of 512 and 256 samples respectively. Correlation matrices are adapted using exponential averaging with a weight for the previous estimate of 0.995. The rank of $\hat{R}_{\tilde{s}\tilde{s}}$ is set to 1 using the GEVD procedure. Three GEIC variations are studied: one with access to $\mathbf{h}$ (denoted by GEIC), one with access only to the angle of arrival, yielding a delay-and-sum beamformer and imperfect Griffiths-Jim blocking matrix (denoted by GEIC$_{\text{GJ}}$) [22], and one using $\hat{\tilde{\mathbf{h}}}_{\text{GEVD}}$ (denoted by GEIC$_{\text{GEVD}}$).

Performance is evaluated using the intelligibility-weighted SNR improvement $\Delta\text{SNR}^{\text{I}}$ (the higher the better), echo return loss enhancement ERLE^{I} (the higher the better), and speech distortion (SD) SD^{I} (the closer to zero the better) [23]–[25].

V. Results and Discussion

While the performance comparison depends on the rank approximation used in (19) and the choice of estimate for $\tilde{\mathbf{h}}$, the MWF$_{\text{ext}}$ generally achieves a higher NR (Fig. 1) than the GEIC for both echo path types due to the equivalence of (21) and (22), where the presence of the postfilter further reduces the noise with respect to the GEIC. This improved noise reduction comes at the expense of a slightly increased SD^{I} when the GEIC is implemented with a perfect blocking matrix. However, for imperfect blocking matrices, desired speech cancellation occurs, thereby also skewing the $\Delta\text{SNR}^{\text{I}}$.

Regarding the AEC performance, considering a linear echo path, both the GEIC and MWF$_{\text{ext}}$ achieve similar performance as both similarly suppress linear echo (Section III-B and III-C). However, considering the non-linear echo path, the GEIC appears more robust (Section III-C), although the GEIC assumes access to $\mathbf{h}$, which might be unavailable in practice.

VI. Conclusion

The generalised echo and interference canceller (GEIC) and the extended multichanncel Wiener filter (MWF$_{\text{ext}}$) for combined noise reduction (NR) and acoustic echo cancellation

(AEC) have been analysed by 1) modelling general nonlinear echo paths using the generalised Bussgang decomposition, and 2) modelling voice activity detection (VAD) error effects in each specific algorithm. Both algorithms suppress the linear echo component regardless of the VAD errors. Generally, the GEIC is found to be more robust towards the residual echo, while the MWF_{ext} achieves the highest NR performance.

REFERENCES

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

[2] 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, Sep. 2004, pp. 2003–2006.

[3] 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, 2022.

[4] 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.

[5] 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, Sep. 2019, pp. 1–5.

[6] 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, Sep. 2018, pp. 6–10.

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

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

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

[10] M. Hamidia and A. Amrouche, "A new robust double-talk detector based on the Stockwell transform for acoustic echo cancellation," *Digital Signal Process.*, vol. 60, pp. 99–112, Jan. 2017.

[11] O. T. Demir and E. Bjornson, "The Bussgang decomposition of nonlinear systems: Basic theory and MIMO extensions [Lecture notes]," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 131–136, Jan. 2021.

[12] A. Roebben, "Github repository: A Comparative analysis of generalised echo and interference cancelling and extended multichannel Wiener filtering for combined noise reduction and acoustic echo cancellation," https://https://github.com/Arnout-Roebben/AECNR_GEIC_vs_MWFext, 2024.

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

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

[15] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, ser. Signals and Communication Technology. Berlin, Heidelberg: Springer-Verlag, 2005.

[16] 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.

[17] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. John Wiley & Sons, Ltd, 2010, ch. 9, pp. 269–302.

[18] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 544–548.

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

[20] 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.

[21] 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.

[22] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[23] 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.

[24] S. Makino, J. Noebauer, Y. Haneda, and A. Nakagawa, "SSB subband echo canceller using low-order projection algorithm," in *Proc. 1996 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Atlanta, GA, USA, May 1996, pp. 945–948 vol. 2.

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