IEEE Open Journal of Signal Processing

Received 22 July 2023; revised 26 October 2023; accepted 1 November 2023. Date of publication 8 November 2023; date of current version 5 December 2023. The review of this article was arranged by Associate Editor R. Serizel.

Digital Object Identifier 10.1109/OJSP.2023.3331228

Multichannel Acoustic Echo Cancellation With Beamforming in Dynamic Environments

YUVAL KONFORTI, ISRAEL COHEN, AND BARUCH BERDUGO

Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering, Technion–Israel Institute of Technology, Haifa 3200003, Israel

CORRESPONDING AUTHOR: YUVAL KONFORTI (email: yuvalkon94@gmail.com).

This work was supported in part by Israel Science Foundation under Grant 1449/23 and in part by Pazy Research Foundation.

ABSTRACT Acoustic echo cancellers are integrated into various speech communication devices, such as hands-free conferencing systems and speakerphones. Microphone arrays can be employed to enhance the performance of such systems, though they assume a static environment when transitioning to double-talk, and rely on double-talk detection. This work introduces a multichannel echo canceller implemented by a microphone array beamformer that can adapt to a changing environment where the locations of both the far-end and near-end sources change during double-talk, with no double-talk detector. This is done by utilizing multiple recent frames in the short-time Fourier transform (STFT) domain. We show how can the acoustic paths be accurately estimated given the recent time frames of the far-end and microphone signals. Also, our beamformer aims to reduce background noise. Simulations are conducted in a reverberant room with nonlinear loudspeaker distortion and realistic low signal-to-echo ratio (SER) resembling a speakerphone. The experiments demonstrate the advantages of the proposed approach compared to normalized least-mean-squares (NLMS) based approaches.

INDEX TERMS Acoustic echo cancellation, adaptive filtering, array signal processing, beamforming.

I. INTRODUCTION

Acoustic echo control is a substantial part of any hands-free teleconferencing system [1], [2], [3], [4]. These systems have gained popularity in recent years. Hence, acoustic echo cancellation (AEC) and acoustic echo suppression (AES) emerge as topics of great importance. The echo cancellation problem is eliminating the undesired echo from the acoustic coupling between a loudspeaker and a microphone. This echo component should be removed so that talkers on each side of the conversation will not hear themselves as feedback. Numerous AEC methods were proposed following the work of Sondhi [5]. Some methods also consider the effects of background noise received by the microphone and loudspeaker nonlinear distortion induced by physical properties of the electrodynamic loudspeaker model [6].

Typically, echo cancellers are implemented by an adaptive filter operating on the reference loudspeaker signal that aims to portray the acoustic echo path. Then, the transmitted signal can be found by subtracting this echo component from the received signal by the microphone. The undesired echo (far-end) signal and the desired (near-end) signal may be active at the same time, making the problem a challenging one in the low signal-to-echo ratio (SER) case. Such periods are referred to as double-talk and can be found by double-talk detectors and voice activity detectors (VADs). The adaptive filter must be adapted during periods when there is no double-talk, with techniques such as least-mean-squares (LMS), normalized LMS (NLMS), proportionate NLMS (PNLMS), and others [1]. While advanced deep-learning-based multimodal detectors are available [7], some AEC applications do not contain additional modalities or lack the required data for network training. Therefore, it is inevitable that during some undetected double-talk periods, the filter adapts inaccurately. Some works have also explored implementing the echo canceller with a deep neural network (DNN) [8], [9], [10], [11].

One method to improve echo cancellation performance is to utilize a microphone array. Adding more microphones enables the use of spatial filtering. Furthermore, the adaptive filter coefficients can be adapted using information obtained by all microphone signals. Such a multi-sensor spatial filter is called a beamformer [12]. Beamformers can be used to amplify speech from some directions while attenuating speech

from other directions in noisy environments [13], [14], [15]. Strategies for combining echo cancellers and beamformers were presented by Kellermann [16], and later on, more methods were explored by [17], [18], [19], [20], [21]. Typically, the whole process is divided into two stages, echo canceling and beamforming, and careful attention should be paid to what stage comes first. If beamforming is used, one must also consider the location of the desired source for it to be preserved. In the case of a dynamic system, this requires dealing with moving sources, which can be challenging [22], [23]. If significant reverberation takes place on the way to the reference microphone, a dereverberation stage might be necessary subsequent to echo cancellation [24].

Another method to better echo control is to utilize multiple frames in the short-time Fourier transform (STFT) domain. This way, an adaptive filter operates on multiple frames of a single microphone. The idea of utilizing multiple frames was introduced by Benesty and Huang [25], [26] and was later used in [27], [28], [29], in the context of noise reduction. Naturally, these works were extended to the problem of echo cancellation in [11], [30], [31], [32], [33], [34]. The common property of these approaches is that a filter is adapted according to the inter-frame correlations. The difference lies in how the inter-frame correlations are estimated. In [31], [32], an initial guess of the desired signal is provided by finding an LMS solution, [30] use the statistics of the far-end signal, [33], [34] use an estimate obtained by NLMS, and in [11] a neural network is used. It should be noted that utilizing the inter-frame correlations can, at best, preserve only the expected component of the near-end signal, which is not necessarily the near-end signal itself. Furthermore, these methods assume that the far-end and near-end talkers are statistically uncorrelated, which is not valid in real scenarios.

The works in [35], [36], [37] consider the multi-microphone speech separation and noise reduction problems, which are inherently different from the AEC problem. In these scenarios, no available far-end signal produces an echo. The availability of the echo-producing loudspeaker signal is critical to diminishing the received echo component. Thus, for the problem of AEC, combining the two methods is beneficial.

Most importantly, typical echo cancellers rely on the assumption that during double-talk periods, the adapted filter from previous time segments can still provide a good estimate of the echo path. This assumption may be problematic in the case of a dynamic environment where the acoustic paths change during double-talk. In such environments, the experimental results show that performance is severely degraded once the acoustic path is changed for up to several seconds [19], [34].

In this paper, we introduce an algorithm for AEC that uses multiple microphones and multiple frames. A beamformer is created, where the coefficients are determined using the reference loudspeaker signal and the array's received signals. Our method requires no double-talk detector and can also adapt the beamformer coefficients during double-talk. This enables the beamformer to work well in dynamic environments where changes in the environment and the acoustic paths may take place during double-talk. Interestingly, the possibly varying environment is taken into account as an inherent part of the adaptation process. Furthermore, the proposed method does not require any knowledge of the array geometry. Background noise is also reduced, and nonlinear loudspeaker distortion is considered. The proposed method is simulated in a realistic low SER case in a reverberant room. Our simulations show that the far-end component is diminished considerably and that more of the near-end signal is preserved compared to existing methods.

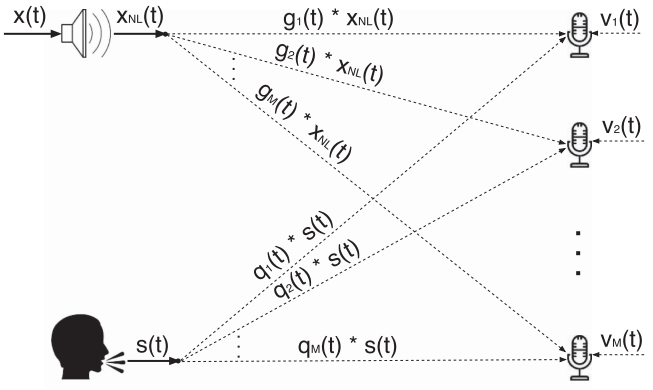

FIGURE 1. Illustration of the environment. *M* microphones capture the far-end speech signal, the near-end speech signal, and background noise. The far-end source is depicted by the loudspeaker and the near-end source is depicted by the speaker.

The rest of the paper is organized as follows. In Section II, we outline the signal model and define performance measures. We show the beamformer design scheme in Section III. Then, in Section IV, we show how multiple frames can be used to estimate the needed parameters to construct the beamformer accurately. Section V is dedicated to a simulative study. Finally, we summarize this work in Section VI.

II. PROBLEM FORMULATION

A. SIGNAL MODEL

Consider a system with M microphones, a far-end speech source, and a near-end speech source as depicted in Fig. 1. The loudspeaker receives a given signal $x(t)$ and emits the far-end signal $x_{\mathrm{NL}}(t)$ with nonlinear distortion induced by the loudspeaker. The talker emits a near-end signal $s(t)$. Additionally, background noise $v_m(t)$ is assumed to be zero-mean, stationary, and independent and identically distributed (IID) between all microphones. Using the signal model for acoustic echo, the m-th microphone signal $d_m(t)$ is given by

$$d_m(t) = g_m(t) * x_{\mathrm{NL}}(t) + q_m(t) * s(t) + v_m(t). \quad (1)$$

Here $g_m(t)$ is the impulse response from the loudspeaker to the m-th microphone, $q_m(t)$ is the impulse response from the talker to the m-th microphone, and $*$ represents the convolutional operator. This can also be written as

$$d_m(t) = y_m(t) + u_m(t) + v_m(t) \quad (2)$$

where

$$y_m(t) = g_m(t) * x_{\mathrm{NL}}(t) \tag{3}$$

is the received far-end speech by the m-th microphone, and

$$u_m(t) = q_m(t) * s(t) \tag{4}$$

is the received near-end speech by the m-th microphone. We arbitrarily define the reference microphone as the first, i.e., by $m = 1$.

Applying the STFT on (2), we get

$$D_m(k,n) = Y_m(k,n) + U_m(k,n) + V_m(k,n) \tag{5}$$

where $D_m(k,n)$, $Y_m(k,n)$, $U_m(k,n)$, and $V_m(k,n)$ are the STFTs of $d_m(t)$, $y_m(t)$ $u_m(t)$, and $v_m(t)$, respectively, at frequency bin k and time frame n. We will also use the approximations

$$Y_m(k,n) \approx G_m(k) X_{\mathrm{NL}}(k,n) \tag{6}$$

$$U_m(k,n) \approx Q_m(k) S(k,n) \tag{7}$$

where $G_m(k)$ is the transfer function (TF) from the loudspeaker to the m-th microphone, $Q_m(k)$ is the TF from the talker to the m-th microphone, and $X_{\mathrm{NL}}(k,n)$ and $S(k,n)$ are the STFTs of $x_{\mathrm{NL}}(t)$ and $s(t)$, respectively. These approximations hold when the lengths of the filters $g_m(t)$ and $q_m(t)$ are significantly shorter than the STFT window length [38].

Finally, we define the array steering vectors with the relative TFs (RTFs):

$$\mathbf{g}(k) = \left[1, \frac{G_2(k)}{G_1(k)}, \ldots, \frac{G_M(k)}{G_1(k)}\right]^T \tag{8}$$

$$\mathbf{q}(k) = \left[1, \frac{Q_2(k)}{Q_1(k)}, \ldots, \frac{Q_M(k)}{Q_1(k)}\right]^T \tag{9}$$

where $\mathbf{g}(k)$ is the steering vector toward the far-end source, $\mathbf{q}(k)$ is the steering vector toward the near-end source, and the superscript ^{T} represents the transpose operator.

Our objective is to find the near-end signal received by the reference microphone $U_1(k,n)$, given the far-end signal $X(k,n)$ and all microphone signals $D_m(k,n)$. This signal can then be sent to the far-end room. We propose to provide an estimate by applying a beamformer:

$$\hat{U}(k,n) = \sum_{m=1}^{M} H_m^*(k,n) D_m(k,n) \tag{10}$$

where $\hat{U}(k,n)$ is an estimate of the desired signal $U_1(k,n)$, $H_m(k,n)$ are the beamformer coefficients at bin k and frame n on the m-th sensor, and the superscript * marks the complex conjugate operator. This can also be written in vector form:

$$\hat{U}(k,n) = \mathbf{h}^H(k,n)\mathbf{d}(k,n) \tag{11}$$

where

$$\mathbf{d}(k,n) = \mathbf{y}(k,n) + \mathbf{u}(k,n) + \mathbf{v}(k,n), \tag{12}$$

$$\mathbf{y}(k,n) = [Y_1(k,n), Y_2(k,n), \ldots, Y_M(k,n)]^T, \tag{13}$$

$$\mathbf{u}(k,n) = [U_1(k,n), U_2(k,n), \ldots, U_M(k,n)]^T, \tag{14}$$

$$\mathbf{v}(k,n) = [V_1(k,n), V_2(k,n), \ldots, V_M(k,n)]^T, \tag{15}$$

$$\mathbf{h}(k,n) = [H_1(k,n), H_2(k,n), \ldots, H_M(k,n)]^T, \tag{16}$$

and the superscript ^{H} marks the transpose conjugate operator.

B. PERFORMANCE MEASURES

To understand how the beamformer in (11) impacts the echo component, we define the residual echo signal by

$$Y_{\mathrm{re}}(k,n) = \mathbf{h}^H(k,n)\mathbf{y}(k,n). \tag{17}$$

A good measure of echo attenuation is the echo-return loss enhancement (ERLE):

$$\xi(t) = \frac{\mathrm{LPF}\left\{y_1^2(t)\right\}}{\mathrm{LPF}\left\{y_{\mathrm{re}}^2(t)\right\}} \tag{18}$$

where $y_{\mathrm{re}}(t)$ is the inverse STFT (ISTFT) of $Y_{\mathrm{re}}(k,n)$, and LPF$\{\cdot\}$ describes a low pass filter. As the residual echo is diminished, the ERLE grows. Thus, the ERLE should be as large as possible.

To examine how the filter in (11) impacts the desired signal component, we define the filtered near-end signal by

$$U_{\mathrm{f}}(k,n) = \mathbf{h}^H(k,n)\mathbf{u}(k,n). \tag{19}$$

Then, the near-end signal distortion can be assessed by the distortion index (DI):

$$\nu(t) = \frac{\mathrm{LPF}\left\{[u_1(t) - u_{\mathrm{f}}(t)]^2\right\}}{\mathrm{LPF}\left\{u_1^2(t)\right\}} \tag{20}$$

where $u_{\mathrm{f}}(t)$ is the ISTFT of $U_{\mathrm{f}}(k,n)$. As the filter distorts the signal, the DI grows. Thus, a small DI is desired.

Additionally, the Perceptual Evaluation of Speech Quality (PESQ) measure [39] can be used to evaluate performance when comparing $\hat{u}(t)$ to $u_1(t)$, where $\hat{u}(t)$ is the ISTFT of $\hat{U}(k,n)$. With PESQ, the residual noise at the filter output

$$V_{\mathrm{rn}}(k,n) = \mathbf{h}^H(k,n)\mathbf{v}(k,n) \tag{21}$$

is also taken into account.

III. BEAMFORMER DESIGN

Considering the above performance measures, the beamformer coefficients can be determined by various methods. This section presents a method that eliminates the echo component, maintains a distortionless response for the desired component, and reduces noise.

Substituting (6) into (13), we get

$$\mathbf{y}(k,n) = X_{\mathrm{NL}}(k,n)\,[G_1(k), G_2(k), \ldots, G_M(k)]^T, \tag{22}$$

then, substituting (22) into (17)

$$Y_{\mathrm{re}}(k,n) = X_{\mathrm{NL}}(k,n)\mathbf{h}^H(k,n)\,[G_1(k), G_2(k), \ldots, G_M(k)]^T. \tag{23}$$

Finally, substituting (8) into (23), we get:

$$Y_{\mathrm{re}}(k,n) = G_1(k) X_{\mathrm{NL}}(k,n)\mathbf{h}^H(k,n)\mathbf{g}(k). \tag{24}$$

Therefore, to eliminate the echo component, we impose the constraint:

$$\mathcal{C}_1\left[\mathbf{h}(k,n)\right]:\mathbf{h}^H(k,n)\mathbf{g}(k)=0. \tag{25}$$

Notice that this constraint eliminates the overall echo component, including the nonlinear distortion induced by the loudspeaker. Similarly, by substituting (7) into (14), we get

$$\mathbf{u}(k,n)=S(k,n)\left[Q_1(k),Q_2(k),\ldots,Q_M(k)\right]^T, \tag{26}$$

then substituting (26) into (19)

$$\begin{aligned}&U_{\mathrm{f}}(k,n)\\&=S(k,n)\mathbf{h}^H(k,n)\left[Q_1(k),Q_2(k),\ldots,Q_M(k)\right]^T.\end{aligned} \tag{27}$$

Finally, substituting (9) into (27), we get:

$$U_{\mathrm{f}}(k,n)=Q_1(k)S(k,n)\mathbf{h}^H(k,n)\mathbf{q}(k). \tag{28}$$

Therefore, to preserve the desired component, we impose the constraint:

$$\mathcal{C}_2\left[\mathbf{h}(k,n)\right]:\mathbf{h}^H(k,n)\mathbf{q}(k)=1. \tag{29}$$

In this way, the near-end signal received by the reference microphone is preserved.

Now, for noise reduction, we consider the residual noise component $V_{\mathrm{rn}}(k,n)$ and minimize

$$\begin{aligned}E\left[|V_{\mathrm{rn}}(k,n)|^2\right]&=\mathbf{h}^H(k,n)E\left[\mathbf{v}(k,n)\mathbf{v}^H(k,n)\right]\mathbf{h}(k,n)\\&=\sigma_v^2(k)\,||\mathbf{h}(k,n)||^2\end{aligned} \tag{30}$$

where E is the expectation operator, and $\sigma_v^2(k)=E[|V_m(k,n)|^2]$ is the noise variance. Overall considering (25), (29), and (30), we can formulate the following problem:

$$\begin{aligned}\mathbf{h}_{\mathrm{opt}}(k,n)=\arg\min_{\mathbf{h}(k,n)}&||\mathbf{h}(k,n)||^2\\ \text{s.t.}\quad&\mathcal{C}_1\left[\mathbf{h}(k,n)\right]\\&\mathcal{C}_2\left[\mathbf{h}(k,n)\right]\end{aligned} \tag{31}$$

where $\mathbf{h}_{\mathrm{opt}}(k,n)$ are the optimal coefficients that eliminate the echo component, preserve the desired component, and minimize the noise component. This is the linear-constraint-minimum-variance (LCMV) beamformer, which is given by

$$\mathbf{h}_{\mathrm{opt}}(k,n)=\mathbf{C}(k)\left[\mathbf{C}^H(k)\mathbf{C}(k)\right]^{-1}\mathbf{i}_c \tag{32}$$

where

$$\mathbf{C}(k)=\left[\mathbf{q}(k),\mathbf{g}(k)\right] \tag{33}$$

and

$$\mathbf{i}_c=[1,0]^T. \tag{34}$$

Theoretically, this beamformer eliminates the echo component and preserves the desired component, producing ideal ERLE and DI. However, it can be constructed only when the steering vectors $\mathbf{g}(k)$ and $\mathbf{q}(k)$ are known. In practice, these may change when the loudspeaker or talker moves or when there are changes in the acoustic paths. Furthermore, they must be estimated accurately. Performance may severely degrade due to steering vector inaccuracies. In the following section, we present a method to estimate both $\mathbf{g}(k)$ and $\mathbf{q}(k)$. These estimates are then used to construct the beamformer in (32), as indicated by the scheme given in Fig. 2.

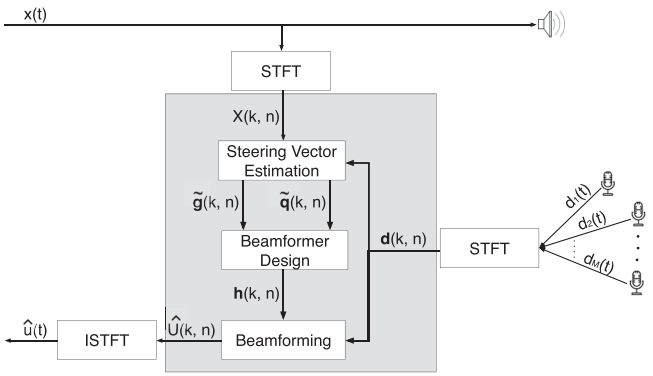

FIGURE 2. Proposed scheme for acoustic echo cancellation. The beamformer in (32) is applied on the *M* microphones with (11). The steering vector estimates determine the beamformer coefficients.

In many echo cancellation algorithms, an additional NLMS-based echo canceller is utilized after the beamformer to reduce residual echo further. While this may improve performance when a fixed beamformer is considered, it can degrade performance when an adaptive beamformer is considered [3]. In the adaptive case, the target filter of the NLMS may change in time, making the convergence of the NLMS filter unstable. In the case of a dynamic environment where the steering vectors change in time, an adaptive beamformer should be utilized. Therefore, such a stage is not utilized in the proposed design.

IV. STEERING VECTOR ESTIMATION

In this section, we provide an estimate for the steering vectors $\mathbf{g}(k)$ and $\mathbf{q}(k)$ used for the beamformer design in Section III. To this end, we assume no movements in the room in the last L frames, i.e., the loudspeaker, talker, and microphones did not move in the last L frames.

Neglecting nonlinear echo and noise, we can utilize (5), (6), and (7) on all the $l\in[1,\ldots,L]$ last frames and get

$$\begin{aligned}D_m\left(k,n-l+1\right)&=G_m(k)X\left(k,n-l+1\right)\\&\quad+Q_m(k)S\left(k,n-l+1\right)\end{aligned} \tag{35}$$

Then, by taking $m=m_1$ and $m=m_2$ for any two sensors m_1 and m_2, we find the ratio

$$\begin{aligned}\frac{Q_{m_1}(k)}{Q_{m_2}(k)}=&\\&\frac{D_{m_1}\left(k,n-l+1\right)-G_{m_1}(k)X\left(k,n-l+1\right)}{D_{m_2}\left(k,n-l+1\right)-G_{m_2}(k)X\left(k,n-l+1\right)}.\end{aligned} \tag{36}$$

Notice that the left-hand side of (36) is independent of l. Thus, we can state that the right-hand side for any $l=l_1$ and $l=l_2$

Algorithm 1.

Inputs: $M, L,$
$X(k, n-l+1), D_m(k, n-l+1) \quad 1 \le l \le L \quad 1 \le m \le M$
Outputs: $\tilde{G}_m(k,n) \quad 1 \le m \le M$
Create list M_{list} of $\binom{M}{2}$ non-repetitive pairs (m_1, m_2)
Create list L_{list} of $\binom{L}{2}$ non-repetitive pairs (l_1, l_2)
$\mathbf{A}(k,n) \leftarrow \mathbf{0}_{\binom{M}{2}\binom{L}{2}\times M}$
for $i = 1, 2, \ldots, \binom{M}{2}$ **do**
  $(m_1, m_2) \leftarrow M_{\text{list}}(i)$
  for $j = 1, 2, \ldots, \binom{L}{2}$ **do**
    $(l_1, l_2) \leftarrow L_{\text{list}}(j)$
    $\mathbf{A}_{[(i-1)\binom{L}{2}+j, m_1]}(k,n) \leftarrow X(k, n-l_1+1)D_{m_2}(k, n-l_2+1) - X(k, n-l_2+1)D_{m_2}(k, n-l_1+1)$
    $\mathbf{A}_{[(i-1)\binom{L}{2}+j, m_2]}(k,n) \leftarrow X(k, n-l_2+1)D_{m_1}(k, n-l_1+1) - X(k, n-l_1+1)D_{m_1}(k, n-l_2+1)$
    $\mathbf{b}_{[(i-1)\binom{L}{2}+j]}(k,n) \leftarrow D_{m_1}(k, n-l_1+1)D_{m_2}(k, n-l_2+1) - D_{m_1}(k, n-l_2+1)D_{m_2}(k, n-l_1+1)$
  end for
end for
$[\tilde{G}_1(k,n), \tilde{G}_2(k,n), \ldots, \tilde{G}_M(k,n)]^T \leftarrow \mathbf{A}^{\dagger}(k,n)\mathbf{b}(k,n)$

are equal

$$\frac{\frac{D_{m_1}(k,n-l_1+1)-G_{m_1}(k)X(k,n-l_1+1)}{D_{m_2}(k,n-l_1+1)-G_{m_2}(k)X(k,n-l_1+1)}}{\frac{D_{m_1}(k,n-l_2+1)-G_{m_1}(k)X(k,n-l_2+1)}{D_{m_2}(k,n-l_2+1)-G_{m_2}(k)X(k,n-l_2+1)}} = \tag{37}$$

Multiplying (37) with the common denominator and simplifying both sides of the equation, the quadratic components of $G_{m_1}(k)G_{m_2}(k)$ are reduced. This yields a linear equation with respect to $G_{m_1}(k)$ and $G_{m_2}(k)$:

$$\begin{aligned}
&G_{m_1}(k)\left[X\left(k, n-l_1+1\right)D_{m_2}\left(k, n-l_2+1\right)\right.\\
&\quad\left.- X\left(k, n-l_2+1\right)D_{m_2}\left(k, n-l_1+1\right)\right]\\
&\quad+ G_{m_2}(k)\left[X\left(k, n-l_2+1\right)D_{m_1}\left(k, n-l_1+1\right)\right.\\
&\quad\left.- X\left(k, n-l_1+1\right)D_{m_1}\left(k, n-l_2+1\right)\right]\\
&= D_{m_1}\left(k, n-l_1+1\right)D_{m_2}\left(k, n-l_2+1\right)\\
&\quad- D_{m_1}\left(k, n-l_2+1\right)D_{m_2}\left(k, n-l_1+1\right)
\end{aligned} \tag{38}$$

Overall, for any pick of $1 \le m_1, m_2 \le M$ and $1 \le l_1, l_2 \le L$, we get an equation as in (38). This set of equations can be used to find $G_m(k)$ for any $1 \le m \le M$, since $X(k, n-l+1)$ and $D_m(k, n-l+1)$ are given. Some of these equations are trivial. For any case where $l_1 = l_2$ or $m_1 = m_2$, the equation is reduced to a degenerate one. Also, notice that when substituting l_1 by l_2 and vice versa, or m_1 by m_2 and vice versa, we get essentially the same equation. Thus, the informative equations are the ones taking (38) for any $l_1 \neq l_2$ and $m_1 \neq m_2$, where the pairs (l_1, l_2) and (m_1, m_2) are not reiterated.

Since every equation corresponds to a different pick of (m_1, m_2) and (l_1, l_2), we arrive at $\binom{M}{2}\binom{L}{2}$ linearly independent equations. The solution to this system can be found only if the number of variables is lower than the number of equations, i.e.,

$$M \le \binom{M}{2}\binom{L}{2}, \tag{39}$$

which when simplified, can be written as:

$$L\left(L-1\right)\left(M-1\right) \ge 4. \tag{40}$$

It is clear from (40) that only a small L and M are needed, i.e., $L \ge 3$ or $M \ge 3$, for us to solve the system. In general, there may be more equations than variables, meaning not all equations are needed. However, due to noise and loudspeaker non-linearity, more equations may help us find a better estimate for $G_m(k)$.

Now, we express this system in matrix form and solve it, as described in Algorithm 1. The matrix $\mathbf{A}(k,n)$ and vector $\mathbf{b}(k,n)$ are constructed so that they define our system. Notice that in the general case, where there are more equations than variables, the system is unsolvable due to conflicting equations. Nevertheless, these conflicts stem from the appearance of noise and loudspeaker non-linearity, which introduce relatively small perturbations. Thus, we can find the least-squares estimator for the system to mitigate the effect of these perturbations. The least-squares estimator can be found by $\mathbf{A}^{\dagger}(k,n)\mathbf{b}(k,n)$, where the superscript † marks the pseudo-inverse operator and

$$\mathbf{A}^{\dagger}(k,n) \triangleq \left[\mathbf{A}^H(k,n)\mathbf{A}(k,n)\right]^{-1}\mathbf{A}^H(k,n). \tag{41}$$

Once we have found all $\tilde{G}_m(k,n)$ the steering vector towards the loudspeaker can be found by utilizing (8)

$$\tilde{\mathbf{g}}(k,n) = \left[1, \frac{\tilde{G}_2(k)}{\tilde{G}_1(k)}, \ldots, \frac{\tilde{G}_M(k)}{\tilde{G}_1(k)}\right]^T \tag{42}$$

and the steering vector toward the talker can be found by substituting (9) and (36)

$$\tilde{\mathbf{q}}(k,n) = \left[1, \frac{D_2(k,n)-\tilde{G}_2(k,n)X(k,n)}{D_1(k,n)-\tilde{G}_1(k,n)X(k,n)}, \ldots , \frac{D_M(k,n)-\tilde{G}_M(k,n)X(k,n)}{D_1(k,n)-\tilde{G}_1(k,n)X(k,n)}\right]^T. \tag{43}$$

The larger M and L are, the more information is used when estimating the steering vectors. Specifically, as L grows, the system obtains more equations while the number of variables remains, producing more accurate results. In this case, however, the acoustic paths are assumed to be static for long periods, which may be problematic in real scenarios. The larger M is the more equations and variables in the system, so while spatial sampling is increased by adding microphones, the steering vector estimation task is more challenging.

Analyzing the computational complexity of Algorithm 1, there are 2 nonzero elements in a row of $\mathbf{A}(k,n)$ that require 2 multiplications each, and all elements of $\mathbf{b}(k,n)$ require 2 multiplications as well. Therefore, the construction of $\mathbf{A}(k,n)$ and $\mathbf{b}(k,n)$ is of complexity

$$\mathcal{O}\left\{\binom{M}{2}\binom{L}{2}\right\} = \mathcal{O}\left\{M^2L^2\right\}. \tag{44}$$

Then, we must find $\mathbf{A}^{\dagger}(k,n)$. From (41), this contains a matrix multiplication, an inverse operation, and another matrix multiplication with overall complexity

$$\mathcal{O}\left\{M\left[\binom{M}{2}\binom{L}{2}\right]^2 + M^3 + M^2\binom{M}{2}\binom{L}{2}\right\} = \mathcal{O}\left\{M^5L^4\right\} \tag{45}$$

Finally, solving the system requires

$$\mathcal{O}\left\{M\binom{M}{2}\binom{L}{2}\right\} = \mathcal{O}\left\{M^3L^2\right\} \tag{46}$$

multiplications. Thus, the overall complexity of Algorithm 1 is $\mathcal{O}\{M^5L^4\}$. Due to this, M and L must be carefully selected. While performance should be maximized, utilizing large values of M and L may affect runtime. In Section V, we show that relatively small values are sufficient for high performance. Also, notice that $\mathbf{A}(k,n)$ is a sparse matrix, thereby reducing runtime and hardware resources.

Note that no double-talk detector is used throughout the proposed scheme in Fig. 2. Our proposed algorithm cancels acoustic echo regardless of the scenario, be it double-talk or single-talk of any of the speakers. This is because the RTFs can be estimated during double-talk, as opposed to RTF estimation methods [14], [40], [41], [42], [43], [44] that assume the existence of only the corresponding source. Specifically, cross-relation system identification methods [43], [44] also exploit the relations between sensor pairs but assume that only one source exists. Furthermore, in a dynamic environment where the talker or loudspeaker moves or a different talker speaks in the near-end room, our algorithm adjusts to the new paths after L time frames. During this time, double-talk may also take place.

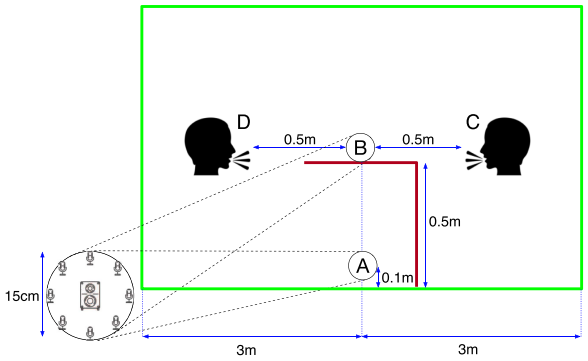

FIGURE 3. Simulated room. The speakerphone, consisting of a microphone array in a UCA geometry and a loudspeaker, can be at $\mathcal{A}$ or $\mathcal{B}$. The active talker can be at $\mathcal{C}$ or $\mathcal{D}$.

V. SIMULATIONS

In this section, we evaluate the proposed beamformer. The ERLE (18), DI (20), and PESQ measurements are studied. Also, the signals received by the microphones and the beamformer output are presented to give better visual perception.

We evaluate the proposed method in a simulated room of dimensions 6 m × 6 m × 4.5 m. To simulate a realistic low SER, we use a speakerphone consisting of a microphone array structured in a uniform circular array (UCA) geometry of radius 7.5 cm and a loudspeaker at the center of the array. Significant acoustic coupling takes place in this configuration. This array geometry was chosen due to its ability to produce high-directivity spatial filters toward any location [45]. Overall, four location configurations depicted by Fig. 3 are considered in our experiment, where each configuration defines a time segment of length 10 s. The speakerphone may be located at $\mathcal{A}$ - the room center on the floor in coordinates [3, 3, 0.1], or at $\mathcal{B}$ - the room center installed on a surface in coordinates [3, 3, 0.5]. The active talker may be located at $\mathcal{C}$ - 0.5 m to the right of $\mathcal{B}$, or at $\mathcal{D}$ - 0.5 m to the left of $\mathcal{B}$. This can fit a scenario where two distinct speakers converse, and the loudspeaker is moved between the floor and the surface. To understand how both loudspeaker and near-end talker locations affect performance, the following four configurations are simulated in this working order:

1) Speakerphone at $\mathcal{A}$, talker at $\mathcal{C}$.
2) Speakerphone at $\mathcal{A}$, talker at $\mathcal{D}$.
3) Speakerphone at $\mathcal{B}$, talker at $\mathcal{C}$.
4) Speakerphone at $\mathcal{B}$, talker at $\mathcal{D}$.

The speech signals $x(t)$ and $s(t)$ were taken from the TIMIT database [46] and $x_{\mathrm{NL}}(t)$ was generated with the model by Thompson [47], [48]. The signals $y_m(t)$ and $u_m(t)$ were found by convolving $x_{\mathrm{NL}}(t)$ and $s(t)$ with the room impulse responses (RIRs) corresponding to the loudspeaker and talker locations, respectively, and the m-th microphone location. The

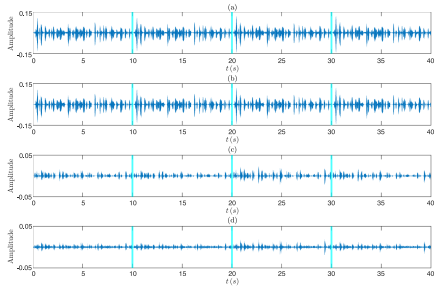

FIGURE 4. Received signals by the first microphone and the beamformer output as a function of time. (a) the total received signal in the reference microphone $d_1(t)$, (b) the echo component signal in the reference microphone $y_1(t)$, (c) the desired component signal in the reference microphone $u_1(t)$, and (d) the beamformer output signal $\hat{u}(t)$. The cyan lines mark the different time segments. $M = L = 4$. Note the difference in scale between (a), (b) and (c), (d).

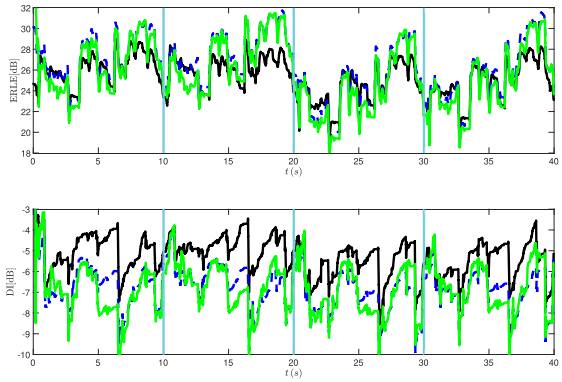

FIGURE 5. ERLE and DI as a function of time for various SNRs. The black, blue-dotted, and green lines mark SNR = 10 dB, 20 dB, and 30 dB, respectively. The cyan lines mark the different time segments. $M = L = 4$.

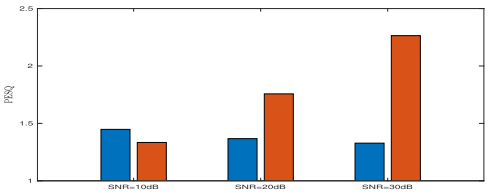

FIGURE 6. PESQ over the entire simulation for various SNRs. The blue and red bars mark the PESQ before and after filtering, respectively. $M = L = 4$.

RIRs were found with the RIR generator by Habets [49]. Speech is sampled at 16 kHz. Unless stated otherwise, white Gaussian noise is added subsequently to all microphones with SNR = 30 dB, and a reverberation time of $T_{60} = 0.3$ s is simulated. The SERs measured in the reference microphone during the four configurations are −17.89 dB, −18.7 dB, −15.27 dB, and −16.89 dB, respectively. For the STFT, a Kaiser window with $\beta = 5$ is used with a length of 512 samples (32 ms) and 75% overlap.

The rest of this section is organized as follows. First, we demonstrate our echo-canceling ability visually with the observed and resulting signals. Then, we investigate how environmental reverberation and noise impact performance. Later, we evaluate how the algorithm parameters M and L impact performance. Finally, we compare the proposed beamformer with the NLMS-based methods in [33], [34] that utilize multiple frames.

A. ECHO CANCELLATION

Fig. 4 shows the received signals by the reference microphone and the beamformer output as a function of time. Notice the scale difference in Fig. 4 between (a), (b) and (c), (d). It is evident the echo component is the main contributor to the received signal due to the small distance between the loudspeaker and the microphone, as is typical in speakerphones. Also, one can see how the near-end component is preserved despite the significant echo component.

B. PERFORMANCE AS FUNCTION OF NOISE AND REVERBERATION

Let us start by studying how environmental noise impacts performance. It has been shown that design immunity to white noise increases robustness to microphone mismatch errors [50]. Therefore, analyzing noise robustness can also be viewed from the perspective of microphone mismatch robustness. The ERLE and DI as a function of time are presented in Fig. 5 for various SNRs. It appears there is some improvement in both ERLE and DI when comparing SNR = 20 dB to SNR = 10 dB, and that the performance of SNR = 30 dB and SNR = 20 dB is comparable. The PESQ for various SNRs over the entire simulation is in Fig. 6. The PESQ before filtering (using $u_1(r)$ and $d_1(t)$) is also presented for reference. Notice that since the SNR varies, the PESQ achieved before filtering varies as well. In contrast to the ERLE and DI, a clear improvement in the PESQ can be observed comparing SNR = 30 dB and SNR = 20 dB. This is because PESQ also takes into account the residual noise at the filter output, which is diminished as SNR increases.

Now, we continue by analyzing how reverberation impacts performance. The ERLE and DI as a function of time are presented in Fig. 7 for various values of T_{60}. A steady decline in both ERLE and DI can be observed as reverberation time increases. The PESQ for various values of T_{60} over the entire simulation is in Fig. 8. Overall, as more reverberations occur, performance is degraded. This can be attributed to the longer impulse responses in the room. In this case, the approximations in (6) and (7) are less accurate, thereby degrading performance.

C. PERFORMANCE AS A FUNCTION OF MICROPHONES AND FRAMES

The ERLE and DI as a function of time are presented in Fig. 9 for various M. As M grows, a slight improvement can be seen in the ERLE, which is most significant when M grows

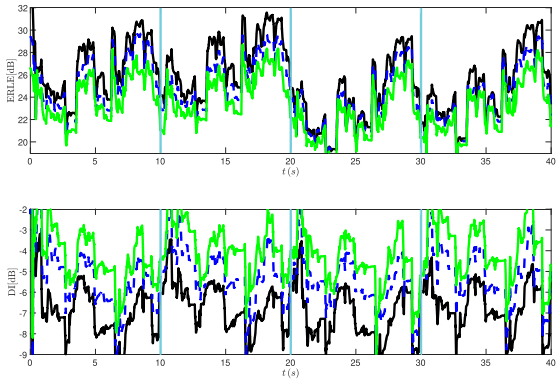

FIGURE 7. ERLE and DI as a function of time for various values of T_{60}. The black, blue-dotted, and green lines mark $T_{60} = 0.3$ s, 0.4 s, and 0.5 s, respectively. The cyan lines mark the different time segments. $M = L = 4$.

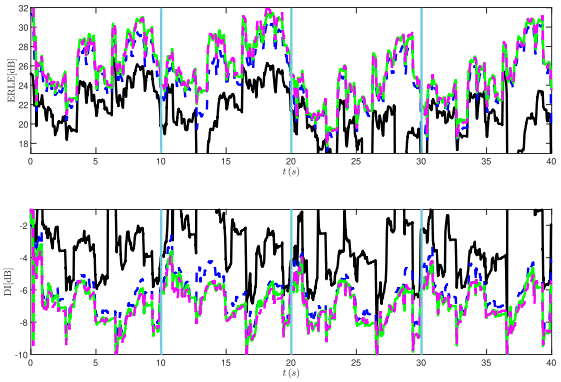

FIGURE 9. ERLE and DI as a function of time for various M. The black, blue-dotted, green, and magenta-dotted lines mark $M = 2$, 3, 4, and 5, respectively. The cyan lines mark the different time segments. $L = 4$.

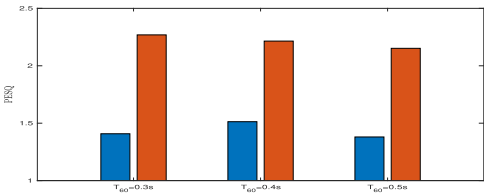

FIGURE 8. PESQ over the entire simulation for various values of T_{60}. The blue and red bars mark the PESQ before and after filtering, respectively. $M = L = 4$.

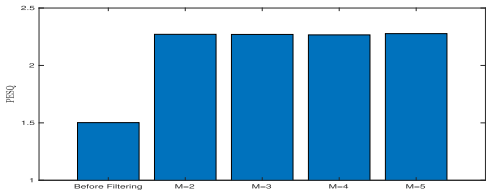

FIGURE 10. PESQ over the entire simulation for various M. $L = 4$.

from 2 to 3. The DI barely changes, except from $M = 2$ to $M = 3$. The PESQ for various values of M over the entire simulation is in Fig. 10. Here, all values of M are comparable. The only slight change in performance as a function of M can be attributed to the fact that increasing M has a dual impact on steering vector estimation. On the one hand, more equations are added to the system; on the other hand, more variables are added. This is translated to a performance limit.

Now, we examine how L impacts performance. The ERLE and DI as a function of time are presented in Fig. 11 for various L. We can clearly state that $L = 2$ frames are insufficient for the proposed approach. Utilizing just the two recent frames does not give an accurate steering vector estimation. This may be because insufficient equations are utilized in the system. Only 6 equations are used to estimate 4 variables. Therefore the appearance of noise and nonlinear echo significantly impacts performance. Furthermore, since the impulse responses are longer than the STFT window length, the TFs may vary between frames; therefore, utilizing only 2 frames degrades performance. Utilizing a large number of frames can help us successfully portray the TFs. This can also be observed when examining the PESQ for various L in Fig. 12. For $L \geq 3$, performance is only slightly improved as L increases.

To sum up, to achieve good echo cancellation with acceptable distortion, both M and L should be larger than 3, although a solution can be produced for lower values that guarantee (39).

D. METHOD COMPARISON

We now compare the proposed method with [33] and [34]. In both competing methods the NLMS filter was adapted during a previous segment that is identical to our first segment, only that the near-end talker is silent. The ERLE and DI as a function of time are presented for all methods in Fig. 13. Higher ERLE and lower DI are obtained with the proposed method, even more so after the speakerphone moves.

Notice that, neglecting reflections, the echo paths do not change throughout the experiment. Therefore, one can expect NLMS-based methods to work well even when the speakerphone moves during double-talk. However, this is not what happens in practice. Both ERLE and DI worsen for the competing methods once the speakerphone moves. This means that varying reflections significantly impact the echo path, degrading the accuracy produced by the NLMS algorithm. The PESQ for all methods, before and after the change in reflections, is in Fig. 14. The PESQ is degraded due to changing reflections as well in the competing methods. Indeed, the adapted NLMS filter is irrelevant once the echo path has changed. This explains the clear advantage of the proposed method in the last two segments. Considering the first two segments, the advantage may be explained by the ability of our method to contain background noise and nonlinear echo during double-talk, as the least-squares estimator in Algorithm 1 is designed to reduce the impact of these on the estimate.

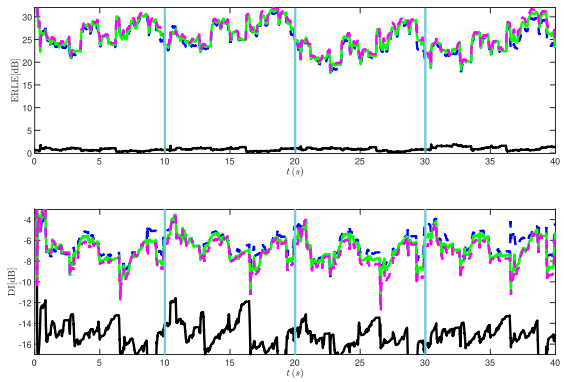

FIGURE 11. ERLE and DI as a function of time for various L. The black, blue-dotted, green, and magenta-dotted lines mark $L = 2$, 3, 4, and 5, respectively. The cyan lines mark the different time segments. $M = 4$.

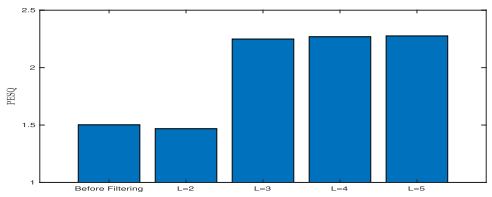

FIGURE 12. PESQ over the entire simulation for various values of L. $M = 4$.

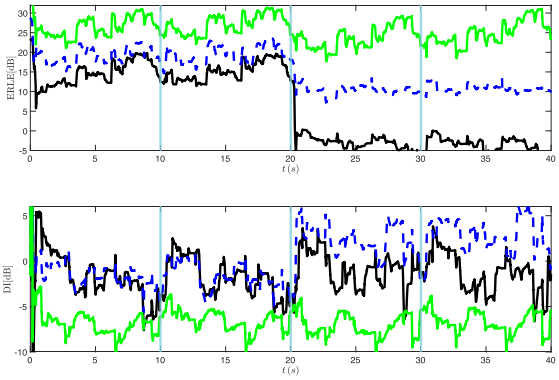

FIGURE 13. ERLE and DI as a function of time for all methods. The black, blue-dotted, and green lines mark the methods [33], [34], and the proposed method, respectively. The cyan lines mark the different time segments. $M = L = 4$.

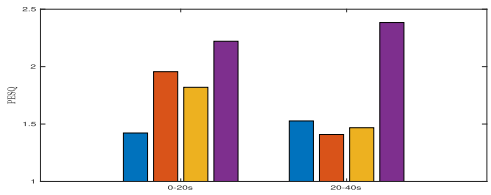

FIGURE 14. PESQ before echo reflections change (0-20 s) and after echo reflections change (20-40 s), for all competing methods. The blue, red, yellow, and purple bars mark the PESQ before filtering, using [33], using [34], and using the proposed method, respectively. $M = L = 4$.

VI. CONCLUSION

An adaptive beamformer for AEC was developed, where the adaptation process considers recent frames of the reference loudspeaker signal and the received microphone signals. This enabled the beamformer to adapt appropriately in a dynamic environment during double-talk, with no double-talk detection. Furthermore, in theory, if there is no background noise and no nonlinear distortion from the loudspeaker, our method completely cancels the echo component while preserving the desired component, as the steering vectors can be accurately estimated from the recent frames. The steering vectors were estimated using a least-squares estimator approach designed to reduce the impact of those two factors specifically. Finally, experiments in a simulated room were conducted. Our experiments indicate that far-end component attenuation, near-end component distortion, and PESQ, all achieve higher performance when compared to NLMS-based methods. This improvement is mainly attributed to our method's ability to adjust to a changing echo path during double-talk, as it responds even to secondary echo path variations that stem from reflections. Future research may focus on more advanced strategies for steering vector estimation utilizing multiple frames. For example, incorporating a nonlinear loudspeaker distortion model in the steering vector estimation process may improve the steering vector estimates for beamformer design.

REFERENCES

[1] M. M. Sondhi, "Adaptive echo cancelation for voice signals," in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2008, ch. 45, pp. 903–928.
[2] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1985.
[3] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001, ch. 13, pp. 281–306.
[4] E. Hansler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Hoboken, NJ, USA: Wiley, 2004.
[5] M. M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol. 46, pp. 497–511, 1967.
[6] W. Klippel, "Loudspeaker nonlinearities – causes, parameters, symptoms," *Audio Eng. Soc. Conv.*, vol. 119, pp. 1–69, 2005.
[7] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 265–274, May 2019.
[8] H. Zhang and D. Wang, "A deep learning approach to multi-channel and multi-microphone acoustic echo cancellation," in *Proc. Interspeech*, 2021, pp. 1139–1143.
[9] T. Haubner and W. Kellermann, "Deep learning-based joint control of acoustic echo cancellation, beamforming and postfiltering," in *Proc. 30th Eur. Signal Process. Conf.*, 2022, pp. 752–756.
[10] H. Zhang, S. Kandadai, H. Rao, M. Kim, T. Pruthi, and T. Kristjansson, "Deep adaptive AEC: Hybrid of deep learning and adaptive acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 756–760.
[11] Y. Tsai, Y. Hsu, and M. R. Bai, "Acoustic echo suppression using a learning-based multi-frame minimum variance distortionless response (MFMVDR) filter," in *Proc. Int. Workshop Acoust. Signal Enhancement*,2022, pp. 1–5.
[12] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. Hoboken, NJ, USA: Wiley, 2018.

[13] R. Berkun, I. Cohen, and J. Benesty, "Combined beamformers for robust broadband regularized superdirective beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 877–886, May 2015.
[14] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2008, ch. 47, pp. 945–978.
[15] E. A. P. Habets, J. Benesty, S. Gannot, and I. Cohen, "The MVDR beamformer for speech enhancement," in *Speech Processing in Modern Communication: Challenges and Perspectives*. Berlin, Germany: Springer, 2010, ch. 9, pp. 225–254.
[16] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1997, pp. 219–222.
[17] G. Reuven, S. Gannot, and I. Cohen, "Joint acoustic echo cancellation and transfer function GSC in the frequency domain," in *Proc. IEEE 23rd Conv. Elect. Electron. Engineers Isr.*, 2004, pp. 412–415.
[18] G. Reuven, S. Gannot, and I. Cohen, "Multichannel acoustic echo cancellation and noise reduction in reverberant environments using the transfer-function GSC," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2007, pp. 81–84.
[19] T. Burton and R. Goubran, "A new structure for combining echo cancellation and beamforming in changing acoustical environments," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2007, pp. 77–80.
[20] G. Reuven, S. Gannot, and I. Cohen, "Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller," *Speech Commun.*, vol. 49, pp. 623–635, 2007.
[21] 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 Eur. Signal Process. Conf.*, 2018, pp. 6–10.
[22] Z. Yermeche, N. Grbic, and I. Claesso, "Blind subband beamforming with time-delay constraints for moving source speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2360–2372, Nov. 2007.
[23] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Online localization and tracking of multiple moving speakers in reverberant environments," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 88–103, Mar. 2019.
[24] E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1433–1451, Nov. 2008.
[25] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 273–276.
[26] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1256–1269, May 2012.
[27] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 201–205.
[28] D. Fischer and S. Doclo, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement*, 2018, pp. 41–45.
[29] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for single-microphone speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 8443–8447.
[30] H. Huang, J. Benesty, J. Chen, K. Helwani, and H. Buchner, "A study of the MVDR filter for acoustic echo suppression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 615–619.
[31] K. Helwani, H. Buchner, J. Benesty, and J. Chen, "A single-channel MVDR filter for acoustic echo suppression," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 351–354, Apr. 2013.
[32] K. Helwani, *Adaptive Identification of Acoustic Multichannel Systems Using Sparse Representations*. Berlin, Germany: Springer, 2015.
[33] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "A multiframe parametric Wiener filter for acoustic echo suppression," in *Proc. IEEE Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.
[34] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "Multiframe echo suppression based on orthogonal signal decompositions," in *Proc. Speech Commun. 12. ITG Symp.*, 2016, pp. 287–291.
[35] M. Tammen and S. Doclo, "Deep multi-frame MVDR filtering for binaural noise reduction," in *Proc. Int. Workshop Acoustic Signal Enhancement*, 2022, pp. 1–5.
[36] E. A. P. Habets, J. Benesty, and J. Chen, "Multi-microphone noise reduction using interchannel and interframe correlations," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2012, pp. 305–308.
[37] Z. Zhang et al., "Multi-channel multi-frame ADL-MVDR for target speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3526–3540, 2021.
[38] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, May 2007.
[39] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, 2001. [Online]. Available: https://www.itu.int/rec/T-REC-P.862
[40] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Speech Audio Process.*, vol. 12, no. 5, pp. 451–459, Sep. 2004.
[41] Y. Avargel and I. Cohen, "Adaptive system identification in the short-time fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 162–173, Jan. 2008.
[42] Y. Avargel and I. Cohen, "Modeling and identification of nonlinear systems in the short-time Fourier transform domain," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 291–304, Jan. 2010.
[43] H. Liu, G. Xu, and L. Tong, "A deterministic approach to blind identification of multi-channel FIR systems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1994, pp. 581–584.
[44] A. Aissa-El-Bey, M. Grebici, K. Abed-Meraim, and A. Belouchrani, "Blind system identification using cross-relation methods: Further results and developments," in *Proc. 7th Int. Symp. Signal Process. Appl.*, 2003, pp. 649–652.
[45] J. Benesty, J. Chen, and I. Cohen, *Design of Circular Differential Microphone Arrays*. Berlin, Germany: Springer, 2015.
[46] J. S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus," 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93s1
[47] S. Thompson, "Nonlinear modeling of a moving coil loudspeaker," *MATLAB Central File Exchange*, 2022. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/121263-nonlinear-modeling-of-a-moving-coil-loudspeaker
[48] S. Thompson, "Acoustical domain for simscape," *MATLAB Central File Exchange*, 2023. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/109029-acoustical-domain-for-simscape
[49] E. A. P. Habets, "Room impulse response generator," pp. 1–21, 2010. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator
[50] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 617–631, Feb. 2007.