ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413836

ROBUST STFT DOMAIN MULTI-CHANNEL ACOUSTIC ECHO CANCELLATION WITH ADAPTIVE DECORRELATION OF THE REFERENCE SIGNALS

Saeed Bagheri, Daniele Giacobello

Sonos Inc., Santa Barbara, CA, USA

ABSTRACT

In this paper, we present an algorithm for multi-channel acoustic echo cancellation for a high-fidelity audio reproduction system equipped with a microphone array for voice control. Since a key requirement for this type of systems is leaving the reference signals that drive the multiple loudspeakers unaltered, we propose an adaptive decorrelation procedure in the time domain before feeding these signals to the echo cancellers. This helps mitigate the so-called non-uniqueness problem arising from their high correlation. This approach offers several advantages: it is extendable to a varying number of channels, requires low computational complexity, and preserves the original reference content. The echo cancellation approach applied to the decorrelated reference signals is then based on STFT-domain robust adaptive methods that do not require double-talk detection. The combination of these two techniques makes our approach particularly desirable for multi-channel echo cancellation problems with very low signal-to-echo ratio.

***Index Terms*—** Multi-channel Acoustic Echo Cancellation (MCAEC), decorrelation, short-time Fourier transform (STFT)

1. INTRODUCTION

An acoustic echo cancellation (AEC) system is generally required to perform voice control on music playback devices, i.e., *smart speakers*, where the coupling of closely spaced loudspeakers and microphones can create very challenging signal-to-echo ratios [1–4].

In the case where a multi-channel speaker setup is deployed, either from multiple spatially distributed loudspeakers (e.g., a 5.1 surround system) or one device equipped with a number of loudspeakers (e.g., a soundbar), the loudspeaker-driving signals (i.e., the reference signals) are typically highly correlated. For P channels, implementing P independent adaptive filters in parallel, based on single-channel AEC, suffers from the so called *non-uniqueness* problem [5]. While stereo AEC has already been discussed extensively (see, e.g., [6] and references therein), contributions on multi-channel AEC (MCAEC) solutions have been less common in the literature, and mostly targeted towards hands-free voice communication [7–9], with only a handful of documented industrial-strength solutions, notably, the MCAEC for the Microsoft Kinect for Xbox [10].

There are two type of solutions to cope with the non-uniqueness problem in MCAEC. The first type, adds distortions to the loudspeaker signals to decorrelate the channels [5, 11–13]. While more recent solutions have applied perceptually motivated criteria in order to reduce audible distortions, e.g., [9, 14], the results are still considered unacceptable for the type of high-fidelity (Hi-Fi) loudspeaker systems we are considering. Furthermore, they might also interfere with the sound beamforming operations often present in this type of system [15], often sensitive to slight changes in the reference path [16]. A second type of solution for MCAEC is more suited for our application scenario and is based on applying *decorrelation filters* to the loudspeaker signals in order to make the convergence faster. The idea is to adjust the adaptive filters by decorrelating the reference signals [8, 17, 18], making these algorithms resilient to convergence issue and the non-uniqueness problem. However, these methods require very high computational and memory resources.

In the single channel AEC case, frequency-domain adaptive filtering (FDAF) [19] was proposed to reduce computational complexity and improve the convergence rate of traditional time-domain LMS algorithms. The multi-delay filter (MDF) [20] was then proposed to reduce the processing delay by segmenting the FDAF into smaller blocks. An interesting incarnation of MDF filtering is to perform system identification directly in the (aliased) STFT domain [21, 22]. This approach requires only one discrete Fourier transform (DFT), and one inverse DFT for the analysis and the synthesis, respectively, of each signal making particularly suited for multistage speech enhancement, usually done in the STFT domain [23].

The use of MCAEC to remove the music feedback at the microphones raises the problem of freezing the adaptive filters when near-end speech (i.e., the voice command) is present to avoid their divergence. A double-talk detection (DTD) system is generally employed to this purpose [24]. However, the variability of the output sound pressure level for music, makes traditional DTD methods based on correlative measures particularly tricky to tune and implement [24]. Robust adaptation algorithms are better suited for this type of problem [25–27]. A robust AEC (RAEC) allows for continuous and stable adaptation of the cancellation filters without applying DTD. The RAEC uses an error recovery nonlinearity (ERN) that enhances the filter estimation error prior to the adaptation.

In this paper, we propose a novel algorithm for MCAEC that applies decorrelation of the reference channels adaptively in the time-domain. The AEC operating on the decorrelated channels is applied directly in the STFT domain, similarly to [21], We incorporate ideas from the robust AEC algorithm to allow for robust update of the adaptive filters when near-end speech is present, similarly to [28]. The combination of these methods offers several advantages for the MCAEC problem in smart speakers: works at very low signal-to-music ratios, accommodates a variable number of reference channels, and attempts to limit distortion on the voice command.

This paper is organized as follows. The STFT-domain MCAEC with adaptive filters is reviewed in Section 2. In Section 3, we present our proposed solution to adaptively decorrelate the reference channels. The STFT-domain single-channel RAEC is then discussed in Section 4. Finally, experimental evaluation and the conclusions are discussed in Section 5 and 6, respectively.

Notation: The transpose, and conjugate transpose of a matrix $\mathbf{X}$ are denoted by $\mathbf{X}^T$, and $\mathbf{X}^H$, respectively. $\mathbf{I}_N$ is the identity matrix of size N. The operators $\mathbb{E}\{\cdot\}$, $\otimes$, $\circ$ denote expectation, Kronecker product, Hadamard (element-wise) product, respectively. The operator $\text{diag}(\cdot)$ converts a vector into a diagonal matrix.

2. PROBLEM DEFINITION

We consider a P-channel acoustic echo canceller operating in the STFT domain [21]. Let $y[n]$ be the near-end microphone signal expressed as $y[n] = d[n] + v[n]$, which consists of the near-end speech and/or noise $v[n]$ mixed with the acoustic echo $d[n] = \sum_{p=1}^{P} h_p[n] * x_p[n]$, where $h_p[n]$ is the impulse response of the system for channel p, and $x_p[n]$ is the far-end reference signal of channel p. Let $\mathbf{x}_p^t[\ell] = [x_p[\ell R], \ldots, x_p[\ell R + N - 1]]^T$ be the ℓ-th frame of the p-th reference signal vector in time-domain where N is the length of STFT and R is its hop-size. The STFT of the reference signals is obtained by applying DFT as $\mathbf{x}_p[\ell] = \mathbf{F}\mathbf{W}_A\mathbf{x}_p^t[\ell]$ where $\mathbf{F}$ is the $N \times N$ DFT matrix and $\mathbf{W}_A$ is a diagonal matrix with analysis window vector on its main diagonal. In the STFT domain, the acoustic echo signal is represented as [22]

$$\mathbf{d}[\ell] = \sum_{p=1}^{P} \sum_{i=0}^{M-1} \mathbf{H}_{i,p}[\ell]\, \mathbf{x}_p[\ell - i]\,, \tag{1}$$

where $\mathbf{d}[\ell] = [D_0[\ell], \ldots, D_{N-1}[\ell]]^T$ is the DFT of the echo signal in frame ℓ, and M is the filter length in the multi-delay STFT domain adaptive filter implementation [20]. The $N \times N$ matrix $\mathbf{H}_{i,p}$ denotes the i-th acoustic impulse response matrix for channel p. In the special case that the impulse response matrices are all diagonal, (1) reduces to the multiplicative transfer function approximation [29].

The goal of a robust MCAEC algorithm is to estimate the channel matrices $\mathbf{H}_{i,p}$. The estimated echo is expressed as $\hat{\mathbf{d}}[\ell] = \sum_{p=1}^{P} \sum_{i=0}^{M-1} \widehat{\mathbf{H}}_{i,p}[\ell - 1]\mathbf{x}_p[\ell - i]$ where $\widehat{\mathbf{H}}_{i,p}$ denotes the estimated adaptive filter. The error signal vector in the STFT domain is defined as $\mathbf{e}[\ell] = \mathbf{y}[\ell] - \hat{\mathbf{d}}[\ell]$ which is decomposed as $\mathbf{e}[\ell] = \mathbf{v}[\ell] + \mathbf{b}[\ell]$, where $\mathbf{v}[\ell]$ and $\mathbf{b}[\ell] \triangleq \mathbf{d}[\ell] - \hat{\mathbf{d}}[\ell]$ are the noise vector and the noise-free error signal vector, respectively. In the presence of near-end speech/noise, the error signal vector $\mathbf{e}[\ell]$ deviates from the true, noise-free, residual echo signal vector $\mathbf{b}[\ell]$, we will show how we deal with this issue in Section 4.

3. DECORRELATION OF LOUDSPEAKERS SIGNALS

The main idea for the decorrelation of the reference channels is based on the following Lemma.

Lemma 1 *Assume that the reference channels are stationary discrete-time random processes. Applying an orthogonalization transformation to the reference channels in the time-domain can be utilized to transform the problem into an equivalent set of independent and parallel adaptive filters in the frequency-domain.*

Proof Define the vector $\mathbf{x}_t[n] \triangleq [x_1[n], \ldots, x_P[n]]^T \in \mathbb{R}^P$ which contains the time-domain samples of the P random processes corresponding to P reference channels at time n. We can define the cross-correlation matrix as $\mathbf{R}_{xx} \triangleq \mathbb{E}\{\mathbf{x}_t[n]\mathbf{x}_t^T[n]\}$, which is time-invariant, and $\mathbf{R}_{xx} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ as its singular-value decomposition (SVD). Assume that $\text{rank}(\mathbf{R}_{xx}) = K$, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_K, 0, \ldots, 0)$ where $\sigma_1 \geq \ldots \geq \sigma_K$. Define the transformation matrix $\mathbf{U}_{[K]}$ comprising the first K columns of $\mathbf{U}$. Following the notation introduced in Section 2, in frame ℓ, we define the $N \times P$ time-domain reference signal matrix as $\mathbf{X}^t[\ell] \triangleq [\mathbf{x}_1^t[\ell], \ldots, \mathbf{x}_P^t[\ell]]$ and its corresponding STFT domain signal matrix is defined as $\mathbf{X}[\ell] \triangleq [\mathbf{x}_1[\ell], \ldots, \mathbf{x}_P[\ell]] = \mathbf{F}\mathbf{W}_A\mathbf{X}^t[\ell]$. Applying the transformation matrix $\mathbf{U}$ in time-domain is equivalent to a matrix multiplication $\overline{\mathbf{X}}^t[\ell] = \mathbf{X}^t[\ell]\mathbf{U}$, which leads to $\overline{\mathbf{X}}[\ell] \triangleq \mathbf{F}\mathbf{W}_A\overline{\mathbf{X}}^t[\ell] = \mathbf{F}\mathbf{W}_A\mathbf{X}^t[\ell]\mathbf{U} = \mathbf{X}[\ell]\mathbf{U}$ in the STFT domain. Assuming that N is larger than the length of the acoustic impulse response of interest, we can just approximate the impulse response matrices by diagonal matrices. In this case, the $NP \times 1$ multi-channel adaptive filter vector can be defined as $\mathbf{h}[\ell] \triangleq [\mathbf{h}_1^T[\ell], \ldots, \mathbf{h}_P^T[\ell]]^T$ where individual channels are concatenated in a vector. Then, the echo signal can be expressed as $\mathbf{d}[\ell] = \mathbf{X}_{\text{diag}}[\ell]\mathbf{h}[\ell]$ where $\mathbf{X}_{\text{diag}}[\ell] \triangleq [\text{diag}(\mathbf{x}_1[\ell]), \ldots, \text{diag}(\mathbf{x}_P[\ell])]$ is a $N \times NP$ matrix. Using $\mathbf{X}[\ell] = \overline{\mathbf{X}}[\ell]\mathbf{U}^T$, we can write $\mathbf{X}_{\text{diag}}[\ell] = \overline{\mathbf{X}}_{\text{diag}}[\ell](\mathbf{U}^T \otimes \mathbf{I}_N)$. Using this expression, the echo signal can be equivalently written as

$$\mathbf{d}[\ell] = \mathbf{X}_{\text{diag}}[\ell]\mathbf{h}[\ell] = \overline{\mathbf{X}}_{\text{diag}}[\ell](\mathbf{U}^T \otimes \mathbf{I}_N)\mathbf{h}[\ell]. \tag{2}$$

This leads to the definition of the equivalent multi-channel adaptive filter vector as $\overline{\mathbf{h}}[\ell] \triangleq (\mathbf{U}^T \otimes \mathbf{I}_N)\mathbf{h}[\ell]$. This implies that applying a unitary transformation to the reference signals corresponds to a set of new acoustic impulse response matrices which are a linear combination of the "true" impulse responses. The frequency-domain normal equations to estimate $\overline{\mathbf{h}}[\ell]$ can be formed as

$$\mathbb{E}\left\{\overline{\mathbf{X}}_{\text{diag}}^H[\ell]\overline{\mathbf{X}}_{\text{diag}}[\ell]\right\}\overline{\mathbf{h}}[\ell] = \mathbb{E}\left\{\overline{\mathbf{X}}_{\text{diag}}^H[\ell]\mathbf{d}[\ell]\right\}. \tag{3}$$

Leveraging the rank of $\mathbf{R}_{xx}$, the new normal equations can be split into independent and parallel systems of equations. First, we have the following simplification

$$\begin{aligned}\mathbb{E}\left\{\overline{\mathbf{X}}_{\text{diag}}^H[\ell]\overline{\mathbf{X}}_{\text{diag}}[\ell]\right\} &= (\mathbf{U}^T \otimes \mathbf{I}_N)\mathbb{E}\left\{\mathbf{X}_{\text{diag}}^H[\ell]\mathbf{X}_{\text{diag}}[\ell]\right\}(\mathbf{U} \otimes \mathbf{I}_N)\\ &= (\mathbf{U}^T \otimes \mathbf{I}_N)(\mathbf{R}_{xx} \otimes (\mathbf{W}_A^H\mathbf{W}_A))(\mathbf{U} \otimes \mathbf{I}_N)\\ &= \mathbf{U}^T\mathbf{R}_{xx}\mathbf{U} \otimes (\mathbf{W}_A^H\mathbf{W}_A) = \mathbf{\Sigma} \otimes (\mathbf{W}_A^H\mathbf{W}_A).\end{aligned} \tag{4}$$

Using the fact that $\text{rank}(\mathbf{R}_{xx}) = K$ in (4), reduces the normal equations into K parallel and independent systems as follows

$$\sigma_p(\mathbf{W}_A^H\mathbf{W}_A)\overline{\mathbf{h}}_p[\ell] = \mathbb{E}\left\{\text{diag}(\overline{\mathbf{x}}_p^H[\ell])\mathbf{d}[\ell]\right\},\ \ p = 1, \ldots, K.$$

These individual systems do not suffer from the non-uniqueness problem. The corresponding echo signal using this simplified model is expressed as $\overline{\mathbf{d}}[\ell] = \overline{\mathbf{X}}_{\text{diag},[K]}[\ell]\overline{\mathbf{h}}_{[K]}[\ell]$ where $\overline{\mathbf{X}}_{\text{diag},[K]}[\ell] \triangleq [\text{diag}(\overline{\mathbf{x}}_1[\ell]), \ldots, \text{diag}(\overline{\mathbf{x}}_K[\ell])]$ and $\overline{\mathbf{h}}_{[K]}[\ell] \triangleq [\overline{\mathbf{h}}_1^T[\ell], \ldots, \overline{\mathbf{h}}_K^T[\ell]]^T$. Given that $\text{rank}(\mathbf{R}_{xx}) = K$, it is straightforward to show that $\mathbb{E}\{\|\mathbf{d}[\ell] - \overline{\mathbf{d}}[\ell]\|_2^2\} = 0$, which implies that the modeling MSE using the proposed K independent systems is zero. ■

We have made two key assumptions to establish the above result. First, we assumed that the reference channels are stationary random processes and the corresponding covariance matrix is time-invariant. Second, the length of the window is assumed to be large enough so that the impulse response matrices can be approximated by diagonal matrices. While, in practice, these assumptions do not hold, we can use Lemma 1 as a guideline to design a practical algorithm for adaptive decorrelation of the reference channels. In particular, the reference channels statistics are slowly time-varying. The statistics can change based on the loudspeakers input signals characteristics (audio source type, and genre) and the content dependent sound spatialization. A practical algorithm to find K and $\mathbf{U}_{[K]}$ online is proposed and summarized in the following.

1. *Initialization*: Use the first L frames to estimate the sample covariance matrix as $\widehat{\mathbf{R}}_{xx}[L] = \frac{1}{LR}\sum_{n=0}^{LR-1}\mathbf{x}_t[n]\mathbf{x}_t^T[n]$. Then, perform SVD to obtain $\widehat{\mathbf{R}}_{xx}[L] = \mathbf{U}_L\mathbf{\Sigma}_L\mathbf{U}_L^T$. The value of K is obtained as the number of singular values that satisfy $\sigma_i \geq \delta\sigma_1$ for some small value δ. $\mathbf{U}_{[K]}$ is defined as the first K columns

of $\mathbf{U}_L$. Also define $\widetilde{\mathbf{R}}_{xx} \triangleq \widehat{\mathbf{R}}_{xx}[L]$ as the reference (current estimate) to track the time-variation.

2. At frames $\ell > L$, using a smoothing coefficient α_R, update
$$\widehat{\mathbf{R}}_{xx}[\ell] = \alpha_R \widehat{\mathbf{R}}_{xx}[\ell-1] + \frac{1-\alpha_R}{R} \sum_{n=\ell R}^{\ell R+R-1} \mathbf{x}_t[n]\mathbf{x}_t^T[n].$$
3. At frame $\ell > L$, calculate the matrix cosine similarity (MCS) metric as a measure of distance between two covariance matrices
$$\eta[\ell] = \frac{|\mathrm{tr}\{\widetilde{\mathbf{R}}_{xx}^H \widehat{\mathbf{R}}_{xx}[\ell]\}|}{\sqrt{\mathrm{tr}\{\widetilde{\mathbf{R}}_{xx}^H \widetilde{\mathbf{R}}_{xx}\}\mathrm{tr}\{\widehat{\mathbf{R}}_{xx}^H[\ell] \widehat{\mathbf{R}}_{xx}[\ell]\}}}.$$
4. At frame $\ell > L$, if $\eta[\ell] \leq \eta_{\text{th}}$, where η_{th} is a defined tolerance threshold, perform SVD to obtain $\widehat{\mathbf{R}}_{xx}[\ell] = \mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{U}_\ell^T$ and update $\widetilde{\mathbf{R}}_{xx} \leftarrow \widehat{\mathbf{R}}_{xx}[\ell]$. Update the value of K and $\mathbf{U}_{[K]}$ similar to Step 1. In this step, re-initialize the state variables in the MCAEC algorithm since the transformation matrix has been updated.
5. Obtain the transformed channels: $\overline{\mathbf{X}}_{[K]}^t[\ell] = \mathbf{X}^t[\ell]\mathbf{U}_{[K]}$.

In the proposed method, we only need to estimate and track a $P \times P$ covariance matrix and calculate its SVD when the MCS metric is below a pre-defined threshold. This guarantees very low computational complexity and memory requirements. Other sub-space tracking techniques might be used, or other metrics instead of MCS can be used to track the time-variation. By exploiting this method of dimensionality reduction, only K adaptive filters are needed which reduces the computational complexity of the algorithm in a practical and inexpensive embedded implementation when P is large.

4. ROBUST ADAPTIVE SINGLE-CHANNEL AEC

In this section, we discuss the main aspects of the robust AEC system where in the multi-channel implementation, the transformed reference channels are used as the inputs to the RAEC block.

Error Recovery Non-linearity: The ERN tries to recover the true residual echo signal from the error signal prior to the adaptive filter update by applying a non-linear clipping function [25]. The RAEC system utilizing ERN allows for robust update of the adaptive filter coefficients even when strong near-end interference is present. Several non-linear clipping functions are investigated in [25] based on different distribution models of the residual echo and near-end signal. The statistical model where the residual echo signal and the near-end signal are assumed to be Gaussian distributed and Laplace distributed, respectively, has been shown to provide the best performance [25–27]. The non-linear clipping function corresponding to this signal model is expressed as

$$\phi(E_m[\ell]) = \begin{cases} \frac{\sqrt{P_{e,m}[\ell]}}{|E_m[\ell]|} E_m[\ell], & |E_m[\ell]| \geq \sqrt{P_{e,m}[\ell]}, \\ E_m[\ell], & \text{otherwise}, \end{cases} \tag{5}$$

where $P_{e,m}[\ell]$ denotes the power spectral density (PSD) of the error signal and is defined as

$$P_{e,m}[\ell] \triangleq \mathbb{E}\{|E_m[\ell]|^2\} \approx \alpha P_{e,m}[\ell-1] + (1-\alpha)|E_m[\ell]|^2, \tag{6}$$

where α is a smoothing coefficient. In vector form, the estimate of the true error signal after applying ERN is defined as $\phi(\mathbf{e}[\ell]) \triangleq [\phi(E_0[\ell]), \ldots, \phi(E_{N-1}[\ell])]^T$.

Adaptive Time-Frequency Dependent Step-Size: The regularization parameter plays an important role in adaptive algorithms to stabilize the filter update. When near-end noise/speech is present, the step-size should be small in order to avoid divergence. When the acoustic impulse response matrices change and as a result the error signal increases, the step-size should increase for increased convergence rate. The RAEC utilizes a noise-robust adaptive step-size [30] which is generalized to the frequency domain in [31]. This adaptive step-size is extended to the STFT-domain crossband filters for single-channel in [28] as (for m-th frequency-band)

$$\mu_{p,m,l}[\ell] = \mu \frac{1}{P_{\bar{x}_p,l}[\ell]} \times \frac{1}{1+\gamma\delta_{p,m,l}[\ell]}, \tag{7}$$

where $\delta_{p,m,l}[\ell] \triangleq P_{e,m}^2[\ell]/P_{\bar{x}_p,l}^2[\ell]$ is the cross-frequency dependent regularization term, γ is a tuning parameter, and $P_{\bar{x}_p,m}[\ell]$ is the PSD of the p-th transformed reference channel estimated as

$$P_{\bar{x}_p,m}[\ell] \triangleq \mathbb{E}\{|\overline{X}_{p,m}[\ell]|^2\} \approx \alpha P_{\bar{x}_p,m}[\ell-1] + (1-\alpha)|\overline{X}_{p,m}[\ell]|^2.$$

Using (7), we define the noise-robust adaptive step-size matrix as $(\mathbf{M}_p[\ell])_{m+1,l+1} = \mu_{p,m,l}[\ell]$ which is used in the expression for the adaptive filter update matrix. The adaptive step-size in (7) includes a regularization term similar to the step-size of the normalized LMS (NLMS) and a scaling term between 0 and 1. The second term automatically scales down the step-size and improves the robustness when near-end noise/speech is present. To improve the overall MCAEC performance, we need to enhance the algorithm's reaction time in the presence of strong near-end noise/speech. Thus, we propose to use a time-frequency dependent tuning parameter and replace γ with $\gamma_0\gamma_{p,m,l}[\ell]$ where γ_0 is a fixed tuning parameter and

$$\gamma_{p,m,l}[\ell] \triangleq \mathbb{E}\{\delta_{p,m,l}^{-1}[\ell]\} \approx \alpha_\gamma \gamma_{p,m,l}[\ell-1] + (1-\alpha_\gamma)\frac{P_{\bar{x}_p,l}^2[\ell]}{P_{e,m}^2[\ell]},$$

where α_γ is a smoothing factor close to 1. In the proposed method, we estimate the expected value of $\delta_{p,m,l}^{-1}[\ell]$, so the scaling term in (7) finds the deviations from its long-term time-average. This method applies different tuning parameters to different frequency-bands based on the reference channel and error signal contents. The parameter α_γ provides a trade-off between adaptation stability and tracking of the impulse response changes.

Finally, the estimated adaptive filters are updated as

$$\widehat{\overline{\mathbf{H}}}_{i,p}[\ell] = \widehat{\overline{\mathbf{H}}}_{i,p}[\ell-1] + \mathbf{M}_p[\ell] \circ (\phi(\mathbf{e}[\ell])\,\overline{\mathbf{x}}_p^H[\ell-i]), \tag{8}$$

for $i = 0, \ldots, M-1$ and $p = 1, \ldots, K$.

The *a posteriori* estimated echo can be expressed as $\hat{\mathbf{d}}_{\text{post}}[\ell] = \sum_{p=1}^{K}\sum_{i=0}^{M-1} \widehat{\overline{\mathbf{H}}}_{i,p}[\ell]\overline{\mathbf{x}}_p[\ell-i]$. The corresponding error signal vector is $\mathbf{e}_{\text{post}}[\ell] = \mathbf{y}[\ell] - \hat{\mathbf{d}}_{\text{post}}[\ell]$.

5. NUMERICAL EXPERIMENTS

The numerical experiments were performed using the model of microphone array and loudspeaker array of the *Sonos Beam*, with $P = 5$ loudspeakers. The room dimension was [6 6 3]m and the microphone embedded on the Sonos Beam was located at [3 1 1]m. The microphone signal ($y[n]$) was generated by convolving the loudspeakers signals with the corresponding acoustic room impulse responses (RIR) ($d[n] = \sum_{p=1}^{P} h_p[n] * x_p[n]$) and then adding near-end speech and noise signal $v(t)$: $y[n] = d[n] + v[n]$. The RIRs were generated using the image source model. The loudspeaker signals ($x_p[n]$, $p = 1, \ldots, 5$) were picked randomly from an internal database of multi-channel loudspeaker signals for the *Sonos Beam* at different volume levels. Rather than using simulations, this was done to capture the internal signal processing and arraying applied to the reference signals (i.e., upmixing/downmixing to 5 channels).

The noise and echo levels were controlled by varying the signal-to-noise ratio (SNR $= \mathbb{E}\{s^2(t)\}/\mathbb{E}\{n^2(t)\}$) and the signal-to-echo ratio (SER $= \mathbb{E}\{s^2(t)\}/\mathbb{E}\{d^2(t)\}$) values. The near-end speech and noise signal were decomposed as $v(t) = s(t) + n(t)$ where $s(t)$ denotes the target speech signal, and $n(t)$ denotes the spatially and temporally white Gaussian noise (AWGN).

The clean speech signals were taken randomly from the TIMIT database. In the simulation, the target talker's mouth was located randomly in the room with respect to the center of the microphone array. The speech sound pressure level (SPL) in the room is picked from a normal distribution with mean $\mu_s = 67$ dBA SPL, and standard deviation $\sigma_s = 9$ dB. The target distance from the microphone array is uniformly distributed in $[1, 4]$m, its azimuth and elevation are randomly selected to be in the interval $[0°, 180°]$ and $[45°, 135°]$, respectively. We considered fixed SNR $= 25$ dB while SER was varied from -35 to -5 dB, and the reverberation time $T_{60} \in \{300, 600\}$ ms. For each combination of SER and T_{60}, we randomly picked 1024 samples from the simulated microphone signals for our evaluations. The sampling frequency was 16 kHz, and the frame length of $N = 512$ samples and $R = 256$ were used in the STFT implementation using Hann windows. The parameters used in the algorithm implementation were: $M = 10$, $\mu = 0.04$, $\alpha = 0.9$, $\alpha_\gamma = 0.999$, and $\eta_{\text{th}} = 0.85$. Only 1 crossband filter is used in the simulation. In our simulation setup, we used fixed values for K to better demonstrate its impact on the performance.

The proposed algorithm performance in terms of echo cancellation is evaluated using echo return loss enhancement (ERLE) $\mathbb{E}\{e^2(t)\}/\mathbb{E}\{y^2(t)\}$ in the frames that speech is not present, and echo cancellation in speech presence (EC-SP) $\mathbb{E}\{(e(t) - v(t))^2\}/\mathbb{E}\{(y(t) - v(t))^2\}$ in the frames that speech is present. The impact on the speech signal was evaluated by calculating the undesired near-end signal attenuation (NEA) defined as $\mathbb{E}\{v^2(t)\}/\mathbb{E}\{e^2(t)\}$ and the log-spectral distortion (LSD). Positive values of NEA are a sign of speech cancellation which is extremely undesirable, while its highly negative values demonstrate the presence of considerable residual echo. Ideally, NEA should be close to 0 dB. For a fair comparison, we used NEA and LSD at -10 and -15 dB SER to tune γ_0 in different methods. We also report the Short-Time Objective Intelligibility (STOI) metric as an indicator of speech intelligibility. These metrics were then averaged over all the 1024 samples in our dataset. To demonstrate the impact of decorrelation in our implementation, we compared 3 configurations: 1) "5-Mono" was based on 5 Mono RAEC with no decorrelation, and using fixed $\gamma = 10.0$ (tuned independently) in (7), 2) "5-Decorr" was based on the proposed decorrelation technique with $\gamma_0 = 0.3$ and with fixed $K = 5$, and 3) "3-Decorr" was based on the proposed decorrelation technique with $\gamma_0 = 0.3$ and fixed $K = 3$, where dimensionality reduction has been also applied. Note that, "5-Mono" and "5-Decorr" have the same level of computational complexity (in terms of number of floating point operations), while 3-Decorr requires approximately 60% of operations of the other two methods.

The performance metrics are illustrated in Fig. 1 as a function of SER. Note that NEA and LSD values are very close in all 3 configurations since we used these metrics at higher SER values to tune the algorithm. In lower SER values, the NEA is improved when the decorrelation technique is applied. The echo cancellation metrics (ERLE and EC-SP) show clearly the benefit of decorrelation technique. Both 3-Decorr and 5-Decorr outperform 5-Mono. Note that the STOI metric shows improvement in speech intelligibility when the decorrelation technique is applied to the reference channels while the echo cancellation is also improved. This demonstrates the desirable performance improvement both in the echo cancellation metrics and in the speech distortion metric. It is notable that 3-Decorr slightly outperforms 5-Decorr which highlights the advantage of using less channels. Lower number of channels leads to faster convergence and improved robustness and stability during double-talk. Considering the lower computational complexity of 3-Decorr, the observed improvement demonstrates the benefits of the proposed decorrelation technique on the AEC performance.

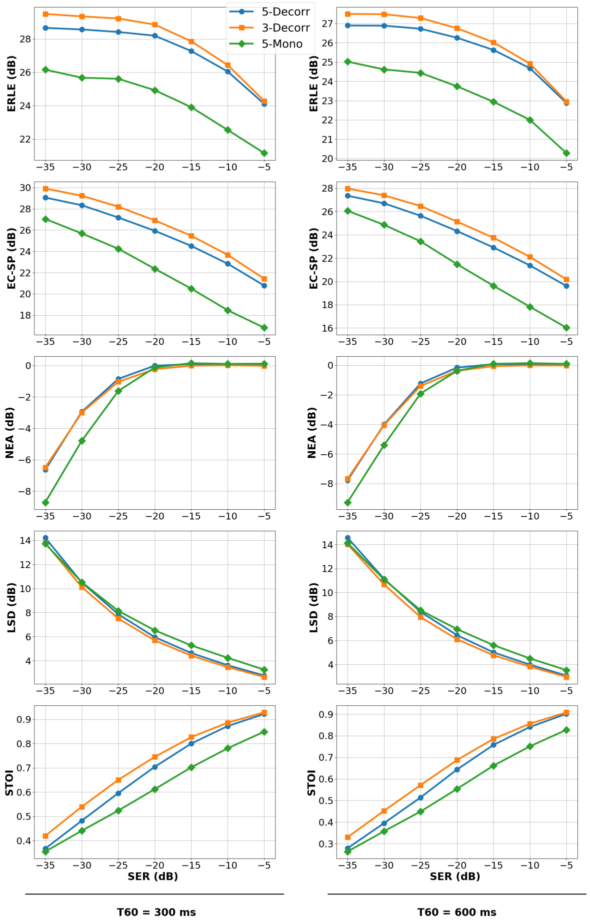

Fig. 1. Performance metrics for $T_{60} = 300$ ms (left) and $T_{60} = 600$ ms (right)

6. CONCLUSION

In order to mitigate the non-uniqueness problem associated with multi-channel AEC, we proposed a time-domain adaptive decorrelation approach for the reference channels. This approach is extendable to a varying number of reference channels and does not modify the loudspeaker signals, both key requirements in high-fidelity *smart* music reproduction system capable of supporting multiple audio formats like the *Sonos Beam* and the *Sonos Arc*, for which the algorithm was designed and evaluated. The combination of this approach with known robust AEC methods in the STFT domain, allows for excellent ERLE performance, while not significantly distorting or attenuating the near-end signal (i.e., the voice command).

7. REFERENCES

[1] A. Chhetri, P. Hilmes, T. Kristjansson, W. Chu, M. Mansour, X. Li, and X. Zhang, "Multichannel audio front-end for far-field automatic speech recognition," in *European Signal Processing Conference (EUSIPCO)*, 2018.

[2] R. Pichevar, J. Wung, D. Giacobello, and J. Atkins, "Design and optimization of a speech recognition front-end for distant-talking control of a music playback device," *arXiv preprint arXiv:1405.1379*, 2014.

[3] Y. A. Huang, T. Shabestary, and A. Gruenstein, "Hotword cleaner: Dual-microphone adaptive noise cancellation with deferred filter coefficients for robust keyword spotting," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[4] D. Giacobello, "An online expectation-maximization algorithm for tracking acoustic sources in multi-microphone devices during music playback," in *European Signal Processing Conference (EUSIPCO)*, 2018.

[5] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation – An overview of the fundamental problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, 1995.

[6] T. Gänsler and J. Benesty, "Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview," *International Journal of Adaptive Control and Signal Processing*, vol. 14, no. 6, pp. 565–586, 2000.

[7] H. Buchner and W. Kellermann, "Acoustic echo cancellation for two and more reproduction channels," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2001.

[8] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication," *Signal Processing*, vol. 85, no. 3, pp. 549–570, 2005.

[9] H. Buchner, "Acoustic echo cancellation for multiple reproduction channels: from first principles to real-time solutions," in *ITG Conference on Voice Communication*, 2008.

[10] I. Tashev, *Sound capture and Processing: Practical Approaches*. John Wiley & Sons, 2009.

[11] A. Gilloire and V. Turbin, "Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[12] A. Sugiyama, Y. Mizuno, A. Hirano, and K. Nakayama, "A stereo echo canceller with simultaneous input-sliding and sliding-period control," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[13] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.

[14] J.-M. Valin, "Channel decorrelation for stereo acoustic echo cancellation in high-quality audio communication," *arXiv preprint arXiv:1603.03364*, 2016.

[15] T. Hooley, "Single box surround sound," *Acoustical science and technology*, vol. 27, no. 6, pp. 354–360, 2006.

[16] K. Wegler, F. Wendt, and R. Höldrich, "How level, delay, and spatial separation influence the echo threshold," *DAGA*, 2019.

[17] J. Benesty, P. Duhamel, and Y. Grenier, "Multi-channel adaptive filtering applied to multi-channel acoustic echo cancellation," in *European Signal Processing Conference (EUSIPCO)*, 1996.

[18] ——, "A multichannel affine projection algorithm with applications to multichannel acoustic echo cancellation," *IEEE Signal Processing Letters*, vol. 3, no. 2, pp. 35–37, 1996.

[19] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[20] J. S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[21] C. Avendano and G. Garcia, "STFT-based multi-channel acoustic interference suppressor," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[22] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[23] 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*. Morgan & Claypool Publishers, 2013.

[24] M. Schneider and E. A. P. Habets, "Comparison of multichannel doubletalk detectors for acoustic echo cancellation," in *European Signal Processing Conference (EUSIPCO)*, 2015.

[25] T. S. Wada and B. H. Juang, "Enhancement of residual echo for robust acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 175–189, 2012.

[26] ——, "Enhancement of residual echo for improved frequency-domain acoustic echo cancellation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007.

[27] ——, "Towards robust acoustic echo cancellation during double-talk and near-end background noise via enhancement of residual echo," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008.

[28] J. Wung, D. Giacobello, and J. Atkins, "Robust acoustic echo cancellation in the short-time fourier transform domain using adaptive crossband filters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[29] Y. Avargel and I. Cohen, "Adaptive system identification in the short-time fourier transform domain using cross-multiplicative transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 162–173, 2008.

[30] A. Hirano and A. Sugiyama, "A noise-robust stochastic gradient algorithm with an adaptive step-size suitable for mobile hands-free telephones," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995.

[31] T. S. Wada and B. H. Juang, "Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009.