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 extensively discussed (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 $X$ are denoted by $X^T$, and $X^H$, respectively. $I_N$ is the identity matrix of size $N$. The operators $\mathbb{E}\{\cdot\}$, $\odot$, $\circ$ 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 $x_p[t] = [x_p[1]|x_p[2]|...|x_p[N]]^T$ be the $t$-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 $X_p[\ell] = \mathcal{F}W_A x_p[\ell]$, where $W_A$ is a diagonal matrix with analysis window vector on its main diagonal. In the STFT domain, the acoustic echo signal is represented as [22]

$$d[\ell] = \sum_{p=1}^{P} \sum_{i=0}^{M-1} H_{i,p}[\ell] x_p[\ell - i],$$

where $d[\ell] = [D_0[\ell], ..., D_{N-1}[\ell]]^T$ is the DFT of the echo signal in frame $\ell$, and $M$ is the filter length in the multi-delay STFT domain adaptive filter implementation [20]. The $N \times N$ matrix $H_{i,p}$ 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 $H_{i,p}$. The estimated echo is expressed as $d[\ell] = \sum_{\ell=0}^{\ell-1} H_{i,p}[\ell-1] x_p[\ell-1]$ where $H_{i,p}$ denotes the estimated adaptive filter. The error signal vector in the STFT domain is defined as $e[\ell] = y[\ell] - d[\ell]$ which is decomposed as $e[\ell] = v[\ell] + b[\ell]$, where $v[\ell]$ and $b[\ell] \triangleq d[\ell] - d[\ell]$ are the noise vector and the noise-free error signal vector, respectively. In the presence of near-end speech/noise, the error signal vector $e[\ell]$ deviates from the true, noise-free residual echo signal vector $b[\ell]$, 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 $x_p[n] \triangleq [x_1[n], ..., x_P[n]]^T \in \mathbb{R}^P$ 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 $R_{x,x} = \mathbb{E} \{x_1[n] x_2[n]^T\}$, which is time-invariant, and $R_{x,x} = \Sigma U U^T$ as its singular-value decomposition (SVD). Assume that rank($R_{x,x}$) = $K$, and $\Sigma = \text{diag}(\sigma_1, ..., \sigma_K, 0, ..., 0)$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K$. Define the transformation matrix $U_{[K]}$ comprising the first $K$ columns of $U$. Following the notation introduced in Section 2, in frame $\ell$, we define the $N \times P$ time-domain reference signal matrix as $X[\ell] \triangleq [x_1[\ell]|x_2[\ell]|...|x_P[\ell]]$ and its corresponding STFT domain signal matrix is defined as $X[\ell] \triangleq [X_1[\ell]|X_2[\ell]|...|X_P[\ell]] = \mathcal{F}W_A X[\ell]$. Applying the transformation matrix $U$ in time-domain is equivalent to a matrix multiplication $X[\ell] = X'[\ell] U$, which leads to $X[\ell] = \mathcal{F}W_A X'[\ell] = U_{[K]} U = X'[\ell] U$ 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 $h[\ell] \triangleq [h_1[\ell]|...|h_P[\ell]]^T$ where individual channels are concatenated in a vector. Then, the echo signal can be expressed as $d[\ell] = X_{\text{diag}}[\ell] h[\ell]$ where $X_{\text{diag}}[\ell] \triangleq \{\text{diag}(X_1[\ell]), ..., \text{diag}(X_P[\ell])\}$ is an $N \times NP$ matrix. Using $X[\ell] = X'[\ell] U$, we can write $X_{\text{diag}}[\ell] = X_{\text{diag}}[\ell] (U^T \otimes I_N)$. Under this assumption, the echo signal can be equivalently written as

$$d[\ell] = X_{\text{diag}}[\ell] h[\ell] = X_{\text{diag}}[\ell] (U^T \otimes I_N) h[\ell].$$

(2)

This leads to the definition of the equivalent multi-channel adaptive filter vector as $\hat{h}[\ell] \triangleq (U^T \otimes I_N) h[\ell]$. 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 $\hat{h}[\ell]$ can be formed as

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

Leveraging the rank of $R_{x,x}$, the new normal equations can be split into independent and parallel systems of equations. First, we have the following simplification

$$E\left\{ \mathcal{X}_{\text{diag}}[\ell]^H \mathcal{X}_{\text{diag}}[\ell] \right\} = (U^T \otimes I_N) E\left\{ \mathcal{X}_{\text{diag}}[\ell]^H \hat{X}_{\text{diag}}[\ell] \right\} (U \otimes I_N)$$

$$= (U^T \otimes I_N) (R_{x,x} \otimes (W_A^H W_A)) (U \otimes I_N)$$

$$= U^T R_{x,x} U \otimes (W_A^H W_A) = \Sigma \otimes (W_A^H W_A). \tag{4}$$

Using the fact that rank($R_{x,x}$) = $K$ in (4), reduces the normal equations into $K$ parallel and independent systems as follows

$$\sigma_p(W_A^H W_A) \hat{h}_p = E\left\{ \text{diag}(\mathcal{X}_p[\ell]^H) d[\ell] \right\}, \ p = 1, \ldots , K. \tag{5}$$

These individual systems do not suffer from the non-uniqueness problem. The corresponding echo signal using this simplified model is expressed as $\hat{d}[\ell] = \mathcal{X}_{\text{diag}}[\ell] \hat{h}[\ell] \triangleq [\text{diag}(X_1[\ell]), ..., \text{diag}(X_K[\ell])]$ and $\hat{h}[K] \triangleq [\hat{h}_1[\ell]|...|\hat{h}_K[\ell]]^T$. Given that rank($R_{x,x}$) = $K$, it is straightforward to show that $E\{|d[\ell] - \hat{d}[\ell]|^2\} = 0$, 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 $U_{[K]}$ online is proposed and summarized in the next section.

1. **Initialization:** Use the first $L$ frames to estimate the sample covariance matrix as $\tilde{R}_{x,x}[L] = \frac{1}{L} \sum_{\ell=0}^{L-1} x_1[n] x_2[n]^T$. Then, perform SVD to obtain $\tilde{R}_{x,x}[L] = \tilde{U}_L \tilde{\Sigma}_L \tilde{U}_L^T$. The value of $K$ is obtained as the number of singular values that satisfy $\sigma_i \geq \delta \sigma_1$ for some small value $\delta$. $U_{[K]}$ is defined as the first $K$ columns.
the signal model is expressed as distributed, respectively, has been shown to provide the best perfor-

mation utilizing ERN allows for robust update of the adaptive filter update by applying a non-linear clipping function \[25\]. The RAEC residual echo signal from the error signal prior to the adaptive filter Error Recovery Non-linearity: where in the multi-channel implementation, the transformed ref-

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 methods 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 sig-

Non-linear clipping function corresponding to this signal model is expressed as

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

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

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

where \( \alpha \) is a smoothing coefficient. In vector form, the estimate of the true error signal after applying ERN is defined as \( \phi(e[l]) \triangleq [\phi(E_1[l]), \ldots, \phi(E_{N-1}[l])]' \).

Adaptive Time-Frequency Dependent Step-Size: The regularization parameter plays an important role in adaptive algorithms to stabili-

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

\[
\mu_{p,m,n}[l] = \mu \frac{1}{P_{p,m}[l]} \times \frac{1}{1 + \gamma_\delta p,m,n[l]}, \tag{7}
\]

where \( \delta_{p,m,n}[l] \triangleq \frac{P_{e,m}[l]}{P_{p,m}[l]} \) is the cross-frequency dependent regularization term, \( \gamma \) is a tuning parameter, and \( P_{p,m}[l] \) is the PSD of the \( p \)-th transformed reference channel estimated as

\[
P_{p,m}[l] \triangleq \mathbb{E}(|X_{p,m}[l]|^2) \approx \alpha P_{p,m}[l-1] + (1-\alpha)|X_{p,m}[l]|^2.
\]

Using (7), we define the noise-robust adaptive step-size matrix as \( \{P_{p,m}[l]\}_{m+1,t+1} = \mu P_{p,m}[l] \) which is used in the expression for the adaptive filter update matrix. The adaptive step-size in (7) in-
cludes a regularization term similar to the step-size of the normal-
ized 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 \( \gamma \) with \( \gamma_\delta \gamma_{p,m,n}[l] \) where \( \gamma_\delta \) is a fixed tuning parameter and \( \gamma_{p,m,n}[l] \triangleq \mathbb{E}(\delta_{p,m,n}[l]) \approx \alpha \gamma_{p,m,n}[l-1] + (1-\alpha)P_{p,m}[l] \)

where \( \alpha_\gamma \) is a smoothing factor close to 1. In the proposed method, we estimate the expected value of \( \delta_{p,m,n}[l] \), 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 \( \alpha_\gamma \) provides a trade-off between adaptation stability and tracking of the impulse response changes.

Finally, the estimated adaptive filters are updated as

\[
\mathbf{H}_{i,p,m}[l] = \mathbf{H}_{i,p,m}[l-1] + \mathbf{M}_{p,m}[l] \circ (\phi(e[l]) \mathbf{x}_p[l-1]), \tag{8}
\]

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

The \textit{a posteriori} estimated echo can be expressed as \( \mathbf{e}_{\text{post}}[l] = \sum_{p=1}^{K} \sum_{i=0}^{M-1} \mathbf{H}_{i,p,m}[l] \mathbf{x}_p[l-i] \). The corresponding error signal vector is \( \mathbf{e}_{\text{post}}[l] = \mathbf{y}[l] - \mathbf{d}_{\text{post}}[l] \).

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 \([6 \times 6 \times 3]\)m and the microphone embedded on the Sonos Beam was located at \([3 \times 1 \times 1]\)m. The microphone signal \( y(n) \) was generated by convolving the loud-
speakers signals with the corresponding acoustic room impulse responses (RIR) \( d[n] = \sum_{p=1}^{K} \mathbf{x}_p[n] + x_p[n] \) 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 sig-

nals \([x_p[n], p = 1, \ldots, 5]\) 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 = \( E\{s^2(t)\}/E\{n^2(t)\}\)) and the signal-to-echo ratio (SER = \( E\{s^2(t)\}/E\{d^2(t)\}\)) 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 \( \sigma_s = 9 \) dBA. 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 \([0^\circ, 180^\circ]\) and \([45^\circ, 135^\circ]\), respectively. We considered fixed SNR = 25 dBA while SER was varied from \(-35\) to \(-5\) dBA, 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 \( N = 512 \) samples and \( R = 256 \) were used in the STFT implementation using Hann windows. The parameters used in the algorithm implementation were: \( M = 10, \mu = 0.04, \alpha = 0.9, \alpha_s = 0.999, \) and \( \eta_n = 0.85 \). 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) \( \mathcal{E}\{e^2(t)\}/\mathcal{E}\{v^2(t)\} \) in the frames that speech is not present, and echo cancellation in speech presence (EC-SP) \( \mathcal{E}\{(e(t) - v(t))^2\}/\mathcal{E}\{(y(t) - v(t))^2\} \) 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 \( \mathcal{E}\{v^2(t)\}/\mathcal{E}\{e^2(t)\} \) 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 \( \gamma \) 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-Decor" was based on the proposed decorrelation technique with \( \gamma_0 = 0.3 \) and with fixed \( K = 5 \), and 3) "3-Decor" 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-Decor" have the same level of computational complexity (in terms of number of floating point operations), while 3-Decor 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-Decor and 5-Decor 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.

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).

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


