Robust STFT Domain Multi-Channel Acoustic Echo Cancellation with Adaptive Decorrelation of the Reference Signals

Saeed Bagheri and Daniele Giacobello

SONOS

2076

ICASSP 2021 TORONTO Canada June 6-11, 2021
Metro Toronto Convention Centre

Challenges and Objectives

- Sonos voice enabled smart multi-channel soundbars

Sonos Beam (5 loudspeakers) — Sonos Arc (11 loudspeakers)

- **Challenges**:
  - Number of loudspeakers, and their configurations vary by product
  - Product dependent performance requirements and CPU utilization budget
  - Low speech-to-echo scenarios in music playback
- **Objectives**:
  - A robust and scalable multi-channel acoustic echo cancellation method
  - Easy to deploy on different devices, and different loudspeaker configurations
  - Fast prototyping, testing, and deployment
- Two types of solutions to cope with the **non-uniqueness problem** [Sondhi et al., 1995]
  1) Add distortions to the loudspeaker signals
     - Examples: Add independent random noise, add perceptually inaudible signals to one of the channels using nonlinear processing, add a time-varying one-sample delay, resample the signals with a rate very close to 1, etc.
  2) Applying *decorrelation filters* to the loudspeaker signals
     - Multi-channel adaptive filtering: extended RLS algorithm, extended LMS, Kalman filters, Affine projection algorithms
- What makes our our problem different
  - High-fidelity (Hi-Fi) loudspeaker systems
    - Distortion-based solutions are considered unacceptable
    - The added distortion interferes with the sound beamforming operations
  - CPU and memory budget
    - Decorrelation filters require high computational and memory resources

MCAEC Problem Formulation

- **Observation Model**: Acoustic echo signal in the STFT domain [Avargel and Cohen, 2007]

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

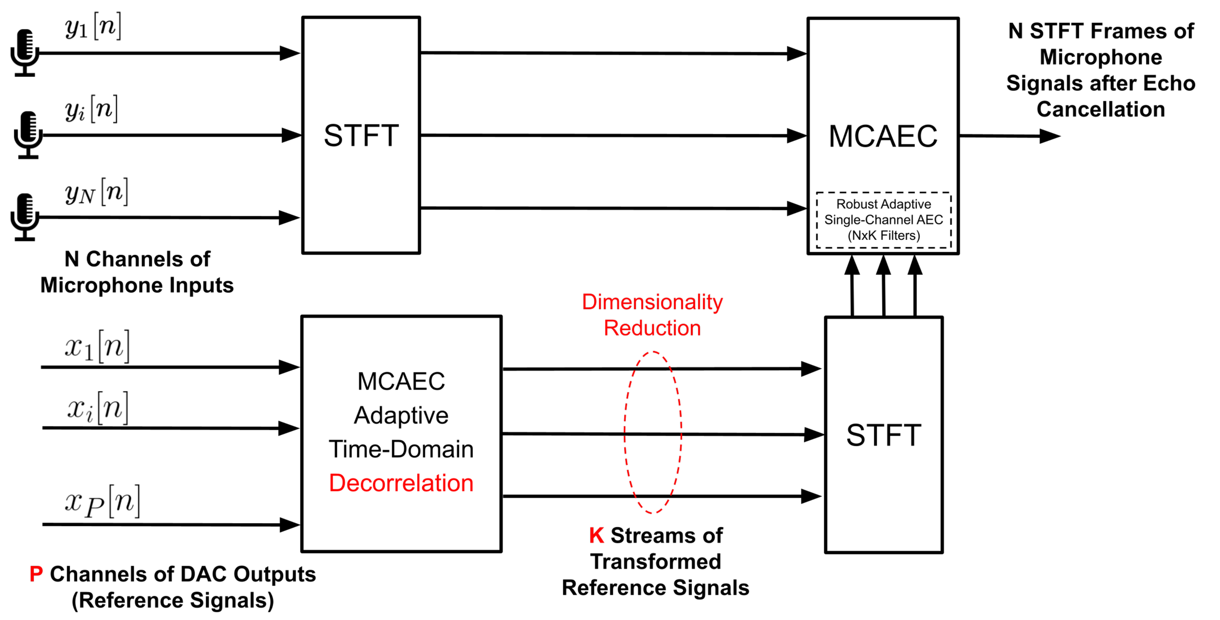
M: filter length in the multi-delay adaptive filter implementation [Soo and Pang, 1990]

- **Objective**: Estimate the channel matrices $\mathbf{H}_{i,p}$ and form the estimated echo

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

Cancel echo from microphone input: $\mathbf{e}[\ell] = \mathbf{y}[\ell] - \hat{\mathbf{d}}[\ell] = \mathbf{v}[\ell] + (\mathbf{d}[\ell] - \hat{\mathbf{d}}[\ell])$

Our Implementation



Decorrelation Algorithm

Key Idea: An orthogonalization transformation in the time-domain transforms the problem into an equivalent set of independent and parallel adaptive filters in the frequency-domain.

- **Objective**: Find a decorrelation matrix $\mathbf{U}_{[K]}$ of size $P \times K$
- **Initialization** (first L frames)
  - Estimate the sample covariance matrix and perform SVD

$$\widehat{\mathbf{R}}_{xx} \triangleq \widehat{\mathbf{R}}_{xx}[L] = \frac{1}{LR}\sum_{n=0}^{LR-1} \mathbf{x}_t[n]\mathbf{x}_t^T[n] = \mathbf{U}_L\boldsymbol{\Sigma}_L\mathbf{U}_L^T$$

  - $K \longleftarrow$ number of singular values that satisfy $\sigma_i/\sigma_1 \geq \delta$
  - $\mathbf{U}_{[K]} \longleftarrow$ first K columns of $\mathbf{U}_L$
- **Adaptive Time-Tracking Steps** (at frame $\ell > L$)
  - Update covariance matrix: using smoothing factor α_R

$$\widehat{\mathbf{R}}_{xx}[\ell] = \alpha_R\widehat{\mathbf{R}}_{xx}[\ell-1] + \frac{1-\alpha_R}{R}\sum_{n=\ell R}^{\ell R+R-1} \mathbf{x}_t[n]\mathbf{x}_t^T[n]$$

  - Calculate matrix cosine similarity (MCS) metric between the *stored* and *current* estimates
  - If MCS $\leq \eta_{\text{th}}$:
    - Perform SVD to obtain $\widehat{\mathbf{R}}_{xx}[\ell] = \mathbf{U}_\ell\boldsymbol{\Sigma}_\ell\mathbf{U}_\ell^T$ and update $\widehat{\mathbf{R}}_{xx} \longleftarrow \widehat{\mathbf{R}}_{xx}[\ell]$
    - Update K and $\mathbf{U}_{[K]}$

Robust Adaptive Single Channel AEC

NLMS Adaptive Filter

- Adaptation rule for $i = 0, \dots, M-1$ and $p = 1, \dots, K$.

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

- $\overline{\mathbf{x}}_p[\ell]$: transformed reference signal
- $\phi(\mathbf{e}[\ell])$: estimate of the true error signal after applying ERN [Wada and Juang, 2012]
- $\mathbf{M}_p[\ell]$: noise-robust adaptive step-size matrix
- The *a posteriori* estimated echo $\longrightarrow \hat{\mathbf{d}}_{\text{post}}[\ell] = \sum_{p=1}^{K}\sum_{i=0}^{M-1} \widehat{\overline{\mathbf{H}}}_{i,p}[\ell]\,\overline{\mathbf{x}}_p[\ell-i]$.

Error Recovery Non-linearity (ERN) $\longrightarrow \phi(\mathbf{e}[\ell])$

- **Goal**: Robust update in the presence of strong near-end interference
- **Method**: Non-linear clipping functions are proposed based on distribution models of the residual echo and near-end signal [Wada and Juang, 2012]
  - Residual echo: Gaussian distributed, near-end signal: Laplace distributed

$$\phi(E_m[\ell]) = \begin{cases} \left(\sqrt{P_{e,m}[\ell]}/|E_m[\ell]|\right)E_m[\ell], & |E_m[\ell]| \geq \sqrt{P_{e,m}[\ell]}, \\ E_m[\ell], & \text{otherwise.} \end{cases}$$

  - $P_{e,m}[\ell] \rightarrow$ the power spectral density (PSD) of the error signal
  - PSDs are estimated by exponential smoothing with factor α

Noise-robust Adaptive Step-size $\longrightarrow \mathbf{M}_p[\ell]$

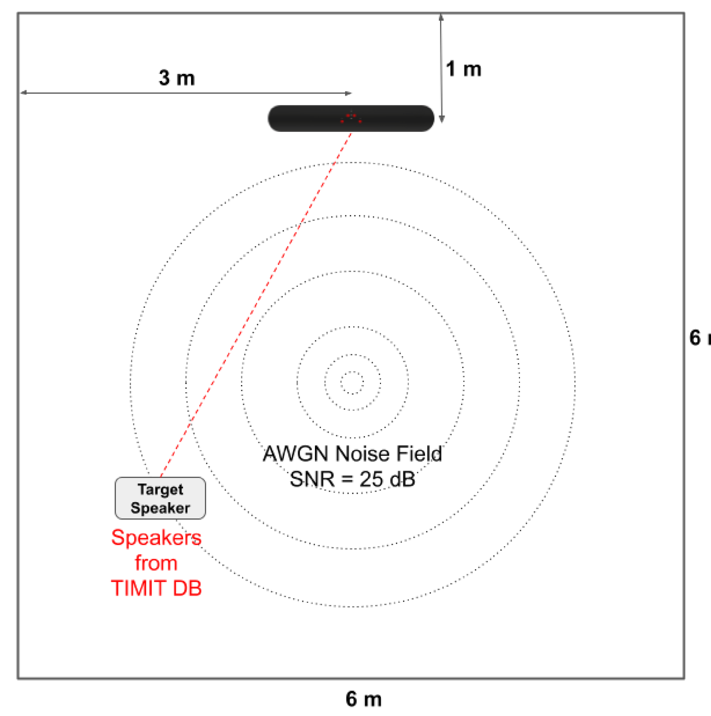
- **Goal**: Small step-size when near-end noise/speech is present
  Increase step-size when the acoustic impulse response matrices change
  - **Method**: Adaptive step-size in the STFT-domain crossband filters [Wung et al., 2014]

$$\left(\mathbf{M}_p[\ell]\right)_{m+1,l+1} = \mu \times \frac{1}{P_{\bar{x}_p,l}[\ell]} \times \frac{1}{1+\gamma\,\delta_{p,m,l}[\ell]}$$

  - $\mu \rightarrow$ adaptation parameter between 0 and 1
  - $P_{\bar{x}_p,m}[\ell] \rightarrow$ PSD of the transformed reference signal
  - $\delta_{p,m,l}[\ell] \rightarrow$ error PSD to reference PSD ratio : $P_{e,m}^2[\ell]/P_{\bar{x}_p,l}^2[\ell]$
  - $\gamma \rightarrow$ tunable regularization parameter
    - Time-frequency dependent tuning parameter: $\gamma \rightarrow \gamma_0\,\gamma_{p,m,l}[\ell]$
    - $\gamma_{p,m,l}[\ell] \triangleq \mathbb{E}\{\delta_{p,m,l}^{-1}[\ell]\} \approx \alpha_\gamma\,\gamma_{p,m,l}[\ell-1] + (1-\alpha_\gamma)\,\delta_{p,m,l}^{-1}[\ell]$

Numerical Experiments

Room Model



Simulation Setup

Sampling frequency	16 KHz
Loudspeaker array	Sonos Beam
# of Loudspeakers	5
Frame length	512
Frame overlap	50%
Window function	Hann
T_{60}	300, 600 ms
# of crossband filters	1
RIR generation	Image source method
Loudspeaker data-set	Internal multi-channel DB
Speech SPL	$\mathcal{N}(67,\ 9)$ dB
SER	$\{-35, -5\}$ dB
Talker distance	$\mathcal{U}(1m,\ 4m)$
Talker azimuth	$\mathcal{U}(0^\circ,\ 180^\circ)$
Talker elevation	$\mathcal{U}(45^\circ,\ 135^\circ)$

Parameters used to implement the proposed algorithm

$M = 10$	$\mu = 0.04$	$\alpha = 0.9$
$\alpha_\gamma = 0.999$	$\eta_{\text{th}} = 0.85$	

Test Scenarios Configurations

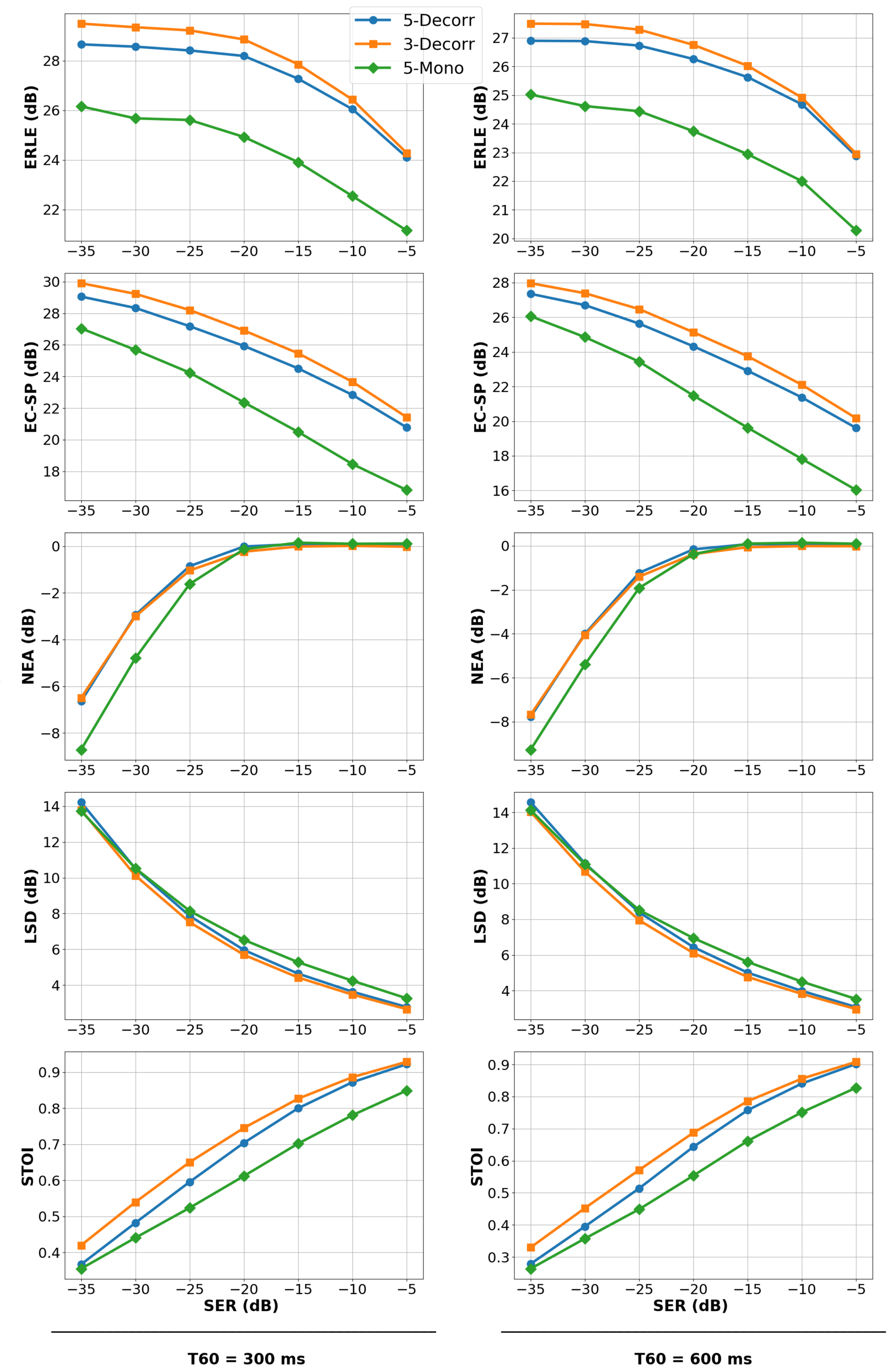
Test Name	Description	flops
"5-Mono"	5 Mono RAEC, no decorrelation, $\gamma = 10$	baseline
"5-Decorr"	proposed decorrelation technique, $\gamma_0 = 0.3$, fixed $K = 5$	baseline
"3-Decorr"	proposed decorrelation technique, $\gamma_0 = 0.3$, fixed $K = 3$	60% of baseline

Evaluation Metrics:

- ERLE: $\dfrac{\mathbb{E}\{e^2(t)\}}{\mathbb{E}\{y^2(t)\}}$
- EC-SP: $\dfrac{\mathbb{E}\{(e(t)-v(t))^2\}}{\mathbb{E}\{(y(t)-v(t))^2\}}$
- NEA: $\dfrac{\mathbb{E}\{v^2(t)\}}{\mathbb{E}\{e^2(t)\}}$
- Log-spectral distortion (LSD)
- Short-Time Objective Intelligibility (STOI)

Comments on Performance:

- Improvement in ERLE and EC-SP
- Same NEA and LSD values → used to tune the algorithm
- STOI → improvement in speech intelligibility with decorrelation
- Lower number of channels → faster convergence, improved robustness during double-talk



References

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*, 15(4):1305–1319, 2007.
M. M. Sondhi, D. R. Morgan, and J. L. Hall. Stereophonic acoustic echo cancellation – An overview of the fundamental problem. *IEEE Signal Processing Letters*, 2(8):148–151, 1995.
J. S. Soo and K. K. Pang. Multidelay block frequency domain adaptive filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(2):373–376, 1990.
T. S. Wada and B. H. Juang. Enhancement of residual echo for robust acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):175–189, 2012.
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*, 2014.