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

Saeed Bagheri Daniele Giacobello

SONOS

ICASSP2021
TORONTO
Canada
June 6-11, 2021
Metro Toronto Convention Centre

Introduction and Motivations

- Sonos voice enabled smart multi-channel soundbars

Sonos Beam

(5 loudspeakers)

Sonos Arc

(11 loudspeakers)

Introduction and Motivations

- Sonos voice enabled smart multi-channel soundbars

Sonos Beam

(5 loudspeakers)

Sonos Arc

(11 loudspeakers)

- **Challenges**
  - Number of loudspeakers and configurations varies by product
  - Industrial design, form factors, and HW modules are different
  - Performance requirements and CPU utilization budget is product dependent
  - Low speech-to-echo scenarios in music playback

Introduction and Motivations

- Sonos voice enabled smart multi-channel soundbars

Sonos Beam

(5 loudspeakers)

Sonos Arc

(11 loudspeakers)

- **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

Relevant Work - MCAEC

- *Non-uniqueness* problem [Sondhi et al., 1995].
- Stereo AEC: [Gänsler and Benesty, 2000] and references therein
- Solutions targeted towards hands-free voice communication [Buchner and Kellermann, 2001; Buchner et al., 2005; Buchner, 2008]
  - A notable industrial-strength solution: Microsoft Kinect for Xbox [Tashev, 2009]
- Two types of solutions to cope with the non-uniqueness problem
  1) Add distortions to the loudspeaker signals
     - Add independent random noise to each channel [Sondhi et al., 1995]
     - Add perceptually inaudible signals to one of the channels using nonlinear processing [Gilloire and Turbin, 1998]
     - Add a non-linearly processed source signal to the source signal itself [Benesty et al., 1998]
     - Add a time-varying one-sample delay to the channels [Sugiyama et al., 2010]
     - Resample the signals with a rate very close to one [Wada et al., 2011]
     - Perceptually motivated criteria to reduce audible distortions [Buchner, 2008; Valin, 2016]

Relevant Work

- Two types of solutions to cope with the non-uniqueness problem
  2) Applying *decorrelation filters* to the loudspeaker signals
     - Multi-channel adaptive filtering that jointly estimates the adaptive filters using extended RLS algorithm, extended LMS [Benesty et al., 1996a]
     - Kalman filters [Buchner et al., 2005]
     - Affine projection algorithms [Benesty et al., 1996b].
- What is different in our scenario?
  - High-fidelity (Hi-Fi) loudspeaker systems
    - Distortion-based solutions are considered unacceptable for the type of systems we are considering
    - The added distortion interferes with the sound beamforming operations [Hooley, 2006], often sensitive to slight changes in the reference path [Wegler et al., 2019]
  - CPU and memory budget
    - The decorrelation filters require very high computational and memory resources

Problem Definition

- The microphone signal

$$y[n] = d[n] + v[n]$$

$v[n]$: near-end speech and/or noise
$d[n]$: acoustic echo with P loudspeaker channels

$$d[n] = \sum_{p=1}^{P} h_p[n] * x_p[n]$$

Problem Definition

- The microphone signal

$$y[n] = d[n] + v[n]$$

$v[n]$: near-end speech and/or noise
$d[n]$: acoustic echo with P loudspeaker channels

$$d[n] = \sum_{p=1}^{P} h_p[n] * x_p[n]$$

- **Observation Model**: Acoustic echo signal in the STFT domain [Avendano and Garcia, 2001; Avargel and Cohen, 2007] (at ℓ-th frame)

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

M: filter length in the multi-delay adaptive filter implementation [Soo and Pang, 1990] $\rightarrow$ Reduces the processing delay

Problem Definition

- The microphone signal

$$y[n] = d[n] + v[n]$$

$v[n]$: near-end speech and/or noise
$d[n]$: acoustic echo with P loudspeaker channels

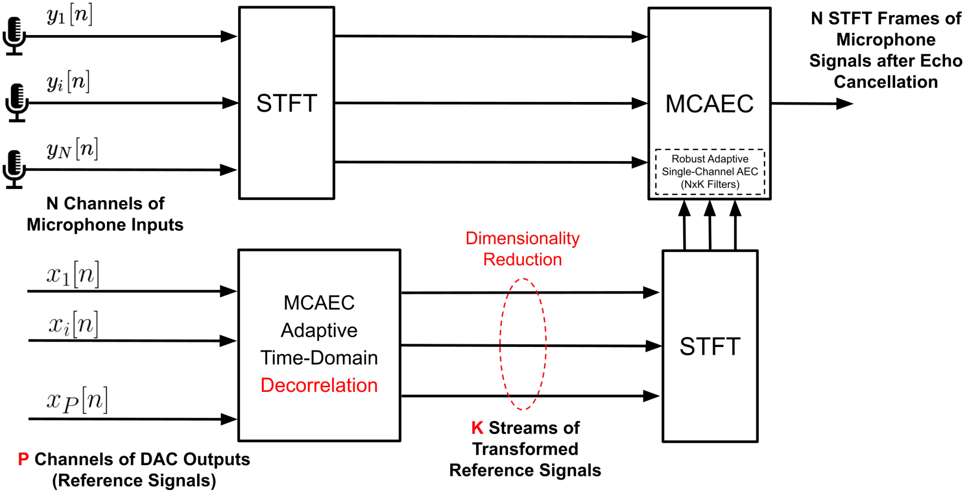
$$d[n] = \sum_{p=1}^{P} h_p[n] * x_p[n]$$

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

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

Echo Cancellation: $\mathbf{e}[\ell] = \mathbf{y}[\ell] - \hat{\mathbf{d}}[\ell] = \mathbf{v}[\ell] + (\mathbf{d}[\ell] - \hat{\mathbf{d}}[\ell])$

Our Implementation



Decorrelation Idea

Lemma

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.

- Goal: Find an orthogonalization transformation matrix
  - Based on the reference channels cross-correlation matrix
- The dimension of the problem can be reduced to K transformed channels
- Echo signal in the transformed space

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

Decorrelation Method

- *Objective*: Find a decorrelation matrix $\mathbf{U}_{[K]}$ of size $P \times K$
- *Initialization*: First L frames
  - Estimate the sample covariance matrix
  - Perform SVD on the sample covariance matrix
  - $K \leftarrow$ number of singular values that satisfy $\frac{\sigma_i}{\sigma_1} \geq \delta$ for some small value δ
  - $\mathbf{U}_{[K]} \leftarrow K$ singular-vectors
- *Adaptive Time-Tracking Steps*: At frame $\ell > L$
  - Update the covariance matrix (using exponential smoothing with smoothing factor α_R)
  - Calculate a measure of distance between **current** and **previous** covariance matrices $\rightarrow$ we use matrix cosine similarity (MCS) metric
  - If MCS $\leq \eta_{\text{th}} \Longrightarrow$ Update stored covariance matrix. Perform SVD to update K and $\mathbf{U}_{[K]}$

Robust Adaptive Single-Channel AEC

- **NLMS Adaptive Filter**

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

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

$\circ \rightarrow$ Hadamard (element-wise) product operation

- $\overline{\mathbf{x}}_p[\ell]$: **transformed** reference signal
- $\phi(\mathbf{e}[\ell])$: estimate of the true error signal after applying Error Recovery Non-linearity (ERN)
- $\mathbf{M}_p[\ell]$: noise-robust adaptive step-size matrix
- The *a posteriori* estimated echo

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

Error Recovery Non-linearity

- **Goal**: Robust update of the adaptive filter coefficients even in the presence of strong near-end interference
- **Method**: Recover the true residual echo from the error signal [Wada and Juang, 2012]
- Non-linear clipping functions are proposed based on distribution models of the residual echo and near-end signal [Wada and Juang, 2012]
  - Residual echo signal: Gaussian distributed; Near-end signal: Laplace distributed
  - The non-linear clipping function

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

$P_{e,m}[\ell] \rightarrow$ the power spectral density (PSD) of the error signal

  - PSDs are estimated by exponential smoothing with factor α

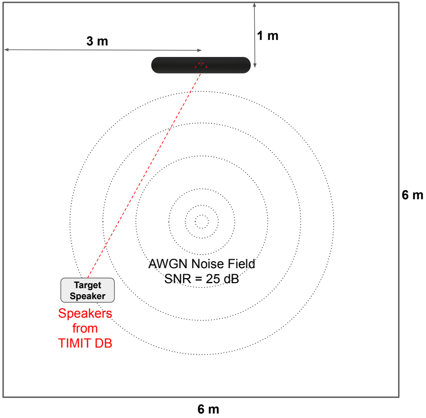
Noise-robust Adaptive Step-size

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

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

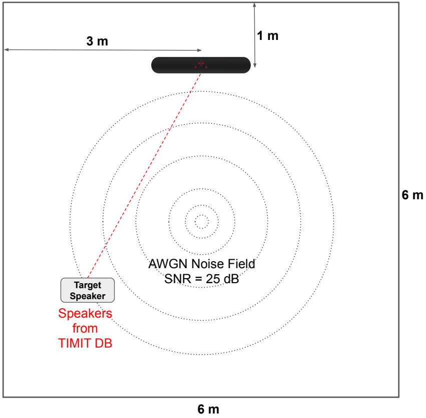
  - $\mu \rightarrow$ adaptation parameter between 0 and 1
  - $P_{\bar{x}_p,m}[\ell] \rightarrow$ PSD of the transformed reference signal
  - $\delta_{p,m,l}[\ell] \rightarrow$ error PSD to reference PSD ratio : $P^2_{e,m}[\ell]/P^2_{\bar{x}_p,l}[\ell]$
  - $\gamma \rightarrow$ tunable regularization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

Simulation Setup



Sampling frequency	16 KHz
Loudspeaker array	Sonos Beam
# of Loudspeakers	5
Frame length	512
Frame overlap (R)	256
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	$\mathcal{N}(67,\ 9)$
SER	$\{-35, -5\}$ dB
Talker distance	$\mathcal{U}(1m,\ 4m)$
Talker azimuth	$\mathcal{U}(0°,\ 180°)$
Talker elevation	$\mathcal{U}(45°,\ 135°)$

Simulation Setup



Sampling frequency	16 KHz
Loudspeaker array	Sonos Beam
# of Loudspeakers	5
Frame length	512
Frame overlap (R)	256
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	$\mathcal{N}(67,\ 9)$
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

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

Simulation Setup

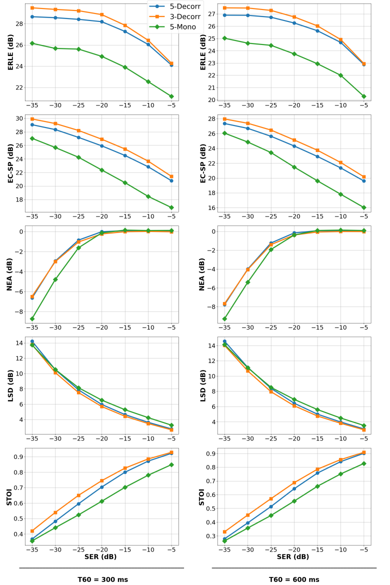
Test Scenarios Configurations:

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:

- Echo return loss enhancement (ERLE): $\frac{\mathbb{E}\{e^2(t)\}}{\mathbb{E}\{y^2(t)\}}$
- Echo cancellation in speech presence (EC-SP): $\frac{\mathbb{E}\{(e(t) - v(t))^2\}}{\mathbb{E}\{(y(t) - v(t))^2\}}$
- Near-end attenuation (NEA): $\frac{\mathbb{E}\{v^2(t)\}}{\mathbb{E}\{e^2(t)\}}$
- Log-spectral distortion (LSD)
- Short-Time Objective Intelligibility (STOI)

Simulation Results



Observations from performance results:

- Improvement in ERLE and EC-SP
- Same NEA and LSD values $\rightarrow$ used them at higher SER values to tune the algorithm
- STOI shows improvement in speech intelligibility when the decorrelation technique is applied
- Lower number of channels $\Rightarrow$ faster convergence and improved robustness and stability during double-talk

Conclusions

A time-domain adaptive decorrelation approach for the reference channels

- Applicable to a varying number of reference channels, and different loudspeaker configurations
- Does not modify the loudspeaker signals $\longrightarrow$ Suitable for Hi-Fi systems
- Very low computational complexity and memory requirements
- Combined this approach with robust AEC methods in the STFT domain
  - Very good ERLE performance
  - Does not significantly distort or attenuate the near-end signal (i.e., the voice command)

Q & A

Thank You For Your Attention!

saeed.sereshki@sonos.com

References I

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. ISSN 1558-7916. doi: 10.1109/TASL.2006.889720.

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

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

J. Benesty, P. Duhamel, and Y. Grenier. A multichannel affine projection algorithm with applications to multichannel acoustic echo cancellation. *IEEE Signal Processing Letters*, 3(2):35–37, 1996b. ISSN 1070-9908. doi: 10.1109/97.484209.

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*, 6(2):156–165, 1998. ISSN 1063-6676. doi: 10.1109/89.661474.

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

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

H. Buchner, J. Benesty, and W. Kellermann. Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication. *Signal Processing*, 85(3):549–570, 2005.

T. Gänsler and J. Benesty. Stereophonic acoustic echo cancellation and two-channel adaptive filtering: an overview. *International Journal of Adaptive Control and Signal Processing*, 14(6):565–586, 2000.

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.

T. Hooley. Single box surround sound. *Acoustical science and technology*, 27(6):354–360, 2006.

References II

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. ISSN 1070-9908. doi: 10.1109/97.404129.

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. ISSN 0096-3518. doi: 10.1109/29.103078.

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.

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

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

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. ISSN 1558-7916. doi: 10.1109/TASL.2011.2159592.

T. S. Wada, J. Wung, and B. H. Juang. Decorrelation by resampling in frequency domain for multi-channel acoustic echo cancellation based on residual echo enhancement. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011.

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

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. doi: 10.1109/ICASSP.2014.6853807.