Robust Acoustic Echo Cancellation in the Short-Time Fourier Transform Domain Using Adaptive Crossband Filters

Jason Wung, Daniele Giacobello, and Joshua Atkins

Beats Electronics, LLC

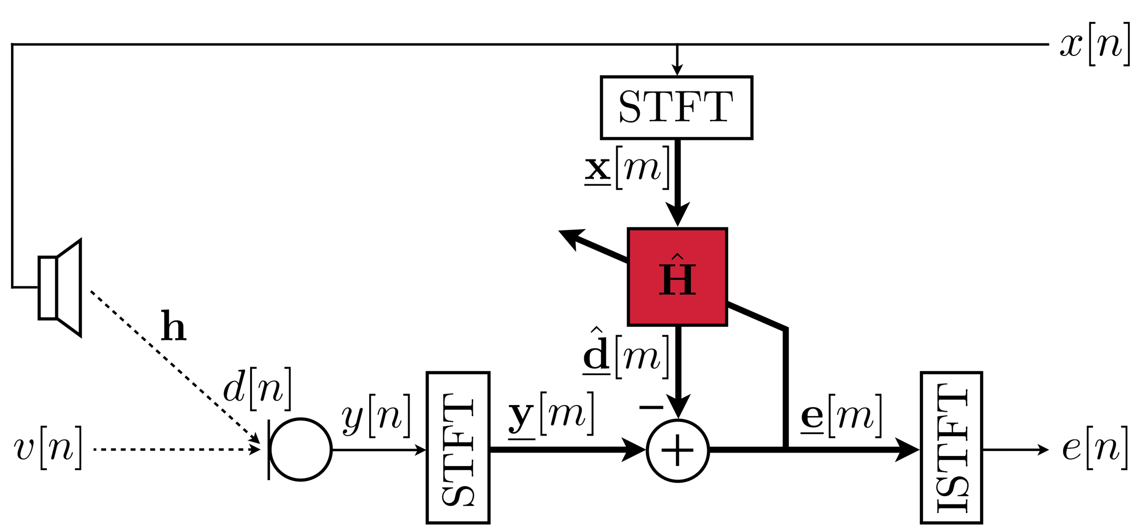
Contact Information:
Beats Electronics, LLC
1601 Cloverfield Blvd.
Suite 5000N
Santa Monica, CA 90404, USA

Email: `jason.wung@beatsbydre.com`

Motivation

- Acoustic echo cancellation (AEC) in the short-time Fourier transform (STFT) domain (1–3) has a simpler system structure than the traditional frequency-domain adaptive filter (FDAF).
  - An FDAF-type algorithm requires several discrete Fourier transforms (DFTs) and inverse DFTs (IDFTs).
  - The STFT-domain processing requires only one DFT and one IDFT for the analysis and the synthesis.
  - AEC in the STFT domain can be easily integrated with a residual echo suppressor (RES).
- The robust AEC (RAEC) provides continuous and stable filter updates during double talk without freezing the adaptive filter but was only used with the FDAF-type algorithms (4).
- In this work, we propose a novel algorithm that combines the simplicity of the STFT-domain AEC with robust adaptive crossband filters.

1 AEC in the STFT Domain



AEC in the STFT domain, where the STFT block represents windowing and transforming to the frequency domain.

- Symbols and definitions:
  - $x[n]$: loudspeaker (far-end) signal
  - $h[n]$: room impulse response
  - $d[n]$: echo signal
  - $v[n]$: near-end speech/noise
  - $y[n]$: near-end microphone signal
  - $e[n]$: error signal
  - $\mathbf{F}$: DFT matrix
  - $\mathbf{w}_A$: analysis window vector
  - $\circ$: Hadamard (element-wise) product
  - $\mathbf{x}[m] = [x[mR], \ldots, x[mR+N-1]]^{\mathrm{T}}$: m^{th} loudspeaker signal vector with frame size N and frame shift R
  - $\underline{\mathbf{x}} = \mathbf{F}(\mathbf{w}_A \circ \mathbf{x}) = [X_0, \ldots, X_{N-1}]^{\mathrm{T}}$: STFT of a signal $\mathbf{x}$
  - $\mathbf{H}$: STFT-domain impulse response matrix
- The STFT-domain echo signal is modeled as (1)

$$\underline{\mathbf{d}}[m] = \sum_{i=0}^{M-1} \mathbf{H}_i[m-1]\underline{\mathbf{x}}[m-i]. \tag{1}$$

  - M is the filter length in the STFT domain.
  - If $\mathbf{H}$ is diagonal, (1) reduces to the multiplicative transfer function approximation (2) but is not accurate due to the finite analysis window length.
  - The modeling accuracy can be improved by adding $2K$ cross-terms without significantly increasing the computational complexity (1).
- The adaptive filter matrix can be updated using

$$\hat{\mathbf{H}}_i[m] = \hat{\mathbf{H}}_i[m-1] + \mathbf{G} \circ \Delta\hat{\mathbf{H}}_i[m], \quad i = 0, \ldots, M-1.$$

  - $\mathbf{G} = \sum_{k=-K}^{K} \mathbf{P}^k$ selects $2K+1$ diagonal bands.
  - $\mathbf{P} \equiv \begin{bmatrix} \mathbf{0}_{1\times N-1} & 1 \\ \mathbf{I}_{N-1\times N-1} & \mathbf{0}_{N-1\times 1} \end{bmatrix}$ is a permutation matrix.
  - $\mathbf{G}$ limits the number of crossband filters that are useful for the STFT-domain AEC (1,3).
- The least mean square (LMS) update matrix is (3)

$$\Delta\hat{\mathbf{H}}_i^{\text{LMS}}[m] = \mu\underline{\mathbf{e}}[m]\underline{\mathbf{x}}^{\mathrm{H}}[m-i]. \tag{2}$$

  - $\underline{\mathbf{e}}[m] = \underline{\mathbf{y}}[m] - \hat{\underline{\mathbf{d}}}[m]$ is the STFT-domain error signal.
  - $\mu > 0$ is a step-size.
  - Eq. (2) takes into account the cross-frequency components of $\underline{\mathbf{x}}$ without relying on the DFT and the IDFT for the gradient constraint in the FDAF.

2 Robust Acoustic Echo Cancellation

- The RAEC uses error recovery nonlinearity (ERN), noise-robust step-size, and iterative adaptation (4).
- The ERN is given by (5)

$$\phi(E_k[m]) = \begin{cases} \frac{T_k[m]}{|E_k[m]|}E_k[m], & |E_k[m]| \geq T_k[m], \\ E_k[m], & \text{otherwise.} \end{cases}$$

  - The ERN limits the error signal when its magnitude is above a certain threshold $T_k[m]$.
  - The threshold is given by $T_k[m] = \sqrt{S_{ee,k}[m]}$, where

$$S_{ee,k}[m] = \beta S_{ee,k}[m-1] + (1-\beta)|E_k[m]|^2.$$

- The noise-robust step-size is given by (6)

$$\mu_k[m] = \mu\frac{S_{xx,k}[m]}{S_{xx,k}^2[m] + \gamma S_{ee,k}^2[m]} = \mu\frac{1}{S_{xx,k}[m] + \delta_k[m]}. \tag{3}$$

  - $\delta_k[m] = \gamma S_{ee,k}^2[m]/S_{xx,k}[m]$ is an adaptive regularization term, where γ is a tuning parameter.
  - The frequency-dependent regularization term scales down the step-size automagically when the near-end interference $v[n]$ is large.

3 Proposed Algorithm

- The normalized LMS (NLMS) update matrix is

$$(\Delta\hat{\mathbf{H}}_i^{\text{NLMS}}[m])_{k+1,l+1} = \mu\frac{E_k[m]X_l^*[m-i]}{S_{xx,l}[m] + \delta}. \tag{4}$$

- Given (3) and (4), the robust step-size extends to a *cross-frequency dependent* regularization term $\delta_{k,l}[m] = \gamma S_{ee,k}^2[m]/S_{xx,l}[m]$ in the STFT domain.
- The proposed update matrix is given by

$$(\Delta\hat{\mathbf{H}}_i[m])_{k+1,l+1} = \mu\frac{\phi(E_k[m])X_l^*[m-i]}{S_{xx,l}[m] + \delta_{k,l}[m]}. \tag{5}$$

Proposed RAEC algorithm in the STFT domain.

Definitions

$$(\mathbf{F})_{k+1,n+1} \equiv e^{-j\frac{2\pi}{N}kn}, \quad k,n = 0, \ldots, N-1$$
$$\phi(\underline{\mathbf{e}}[m]) \equiv [\phi(E_0[m]), \ldots, \phi(E_{N-1}[m])]^{\mathrm{T}}$$

Echo cancellation

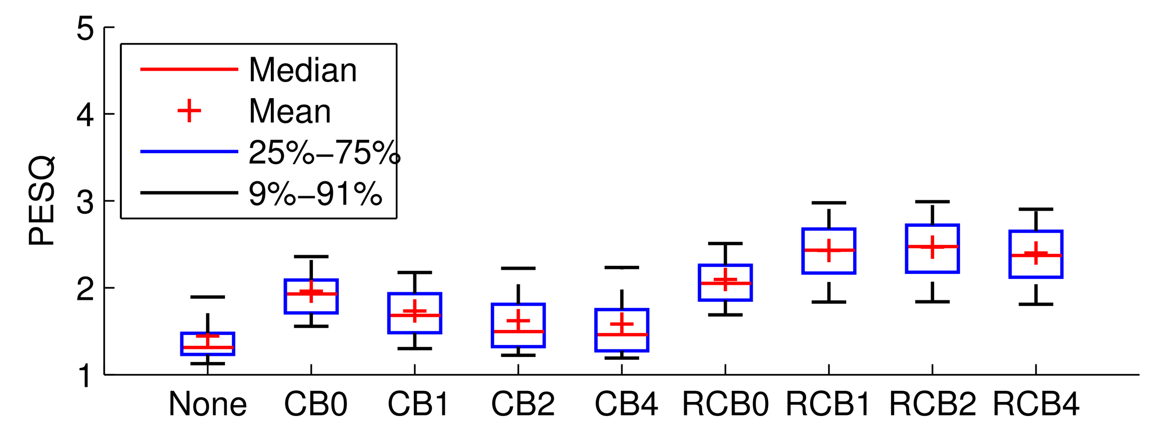
$$\underline{\mathbf{x}}[m] = \mathbf{F}(\mathbf{w}_A \circ [x[mR], \ldots, x[mR+N-1]]^{\mathrm{T}})$$
$$\underline{\mathbf{y}}[m] = \mathbf{F}(\mathbf{w}_A \circ [y[mR], \ldots, y[mR+N-1]]^{\mathrm{T}})$$
$$\underline{\mathbf{e}}[m] = \underline{\mathbf{y}}[m] - \sum_{i=0}^{M-1} \hat{\mathbf{H}}_i[m-1]\underline{\mathbf{x}}[m-i]$$

Filter adaptation

$$\underline{\mathbf{s}}_{xx}[m] = \beta\underline{\mathbf{s}}_{xx}[m-1] + (1-\beta)(\underline{\mathbf{x}}[m] \circ \underline{\mathbf{x}}^*[m])$$
$$\underline{\mathbf{s}}_{ee}[m] = \beta\underline{\mathbf{s}}_{ee}[m-1] + (1-\beta)(\underline{\mathbf{e}}[m] \circ \underline{\mathbf{e}}^*[m])$$
$$(\mathbf{M}[m])_{k+1,l+1} = \frac{S_{xx,l}[m]}{S_{xx,l}^2[m] + \gamma S_{ee,k}^2[m]}, \quad k,l = 0, \ldots, N-1$$
$$\Delta\hat{\mathbf{H}}_i[m] = \mu\mathbf{M}[m] \circ \{\phi(\underline{\mathbf{e}}[m])\underline{\mathbf{x}}^{\mathrm{H}}[m-i]\}, \quad i = 0, \ldots, M-1$$
$$\hat{\mathbf{H}}_i[m] = \hat{\mathbf{H}}_i[m-1] + \mathbf{G} \circ \Delta\hat{\mathbf{H}}_i[m], \quad i = 0, \ldots, M-1$$

4 Experimental Evaluation

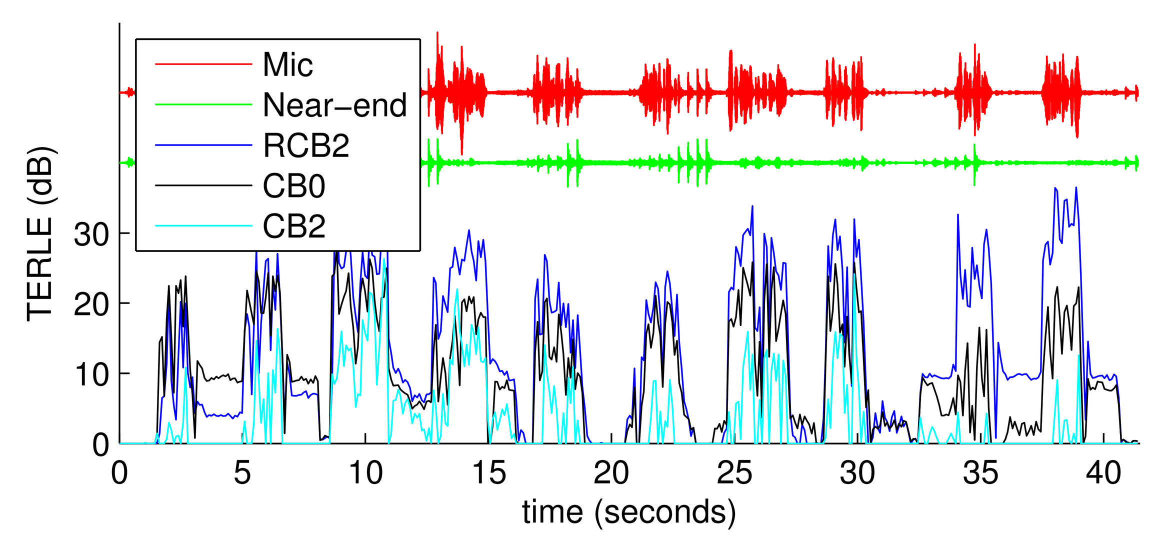
- Impulse response measurement:
  - Two *Beats by Dr. Dre Pill* speakers spaced 1 meter apart were used for the measurement.
  - The sound pressure level (SPL) was calibrated to 85 dB_C at 1 meter away with a -20 dBFS narrow-band (500 Hz to 2 kHz) pink noise.
  - The microphone was placed closely to one of the speakers to measure the room impulse response $\mathbf{h}$ and the impulse response from the other speaker to the microphone.
  - The SPL of the echo signal was about 20 dB stronger than the near-end signal.
- Speech files and noise files from the ITU-T P.501 test signals were randomly selected.
  - Noise was added to speech with a segmental signal-to-noise ratio (SSNR) of -5, 0, 5, and 10 dB.
  - The near-end speech plus noise and the far-end speech were constantly overlapped.
  - 100 utterances were generated with an averaged length of 40 seconds for each utterance.
- Perceptual Evaluation of Speech Quality (PESQ):



Mean opinion score using PESQ.

  - Clean near-end speech as reference for PESQ
  - CB: crossband filters with the NLMS update (4)
  - RCB: robust adaptive crossband filters (5)
  - The number after CB represents K.
- True echo return loss enhancement (TERLE):

$$\text{TERLE (dB)} \equiv 10\log_{10}\left(\frac{\sum_n |y[n]-v[n]|^2}{\sum_n |e[n]-v[n]|^2}\right).$$



TERLE plot comparing the STFT-domain AEC with and without the robustness constraint.

References

(1) Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

(2) ——, "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.

(3) ——, "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.

(4) T. S. Wada and B.-H. Juang, "Enhancement of residual echo for robust acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 20, no. 1, pp. 175–189, Jan. 2012.

(5) J. Wung, T. S. Wada, B.-H. Juang, B. Lee, M. Trott, and R. W. Schafer, "A system approach to acoustic echo cancellation in robust hands-free teleconferencing," in *Proc. IEEE WASPAA*, pp. 101–104, Oct. 2011.

(6) T. S. Wada and B.-H. Juang, "Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement," in *Proc. IEEE WASPAA*, pp. 205–208, Oct. 2009.