Deep Learning-based Acoustic-echo Cancellation

Amir Ivry Mark

Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering
Technion – Israel Institute of Technology

Supervised by Prof. Israel Cohen and Dr. Baruch Berdugo
November 1st, 2023

Outline

Acoustic-echo Cancellation: Definition

In a virtual conversation between two ends; a near-end and a far-end, acoustic-echo cancellation (AEC) systems aim to cancel the echo before it returns from the end that received it to the end that created it.

AEC in the Monophonic Case: Schematic View

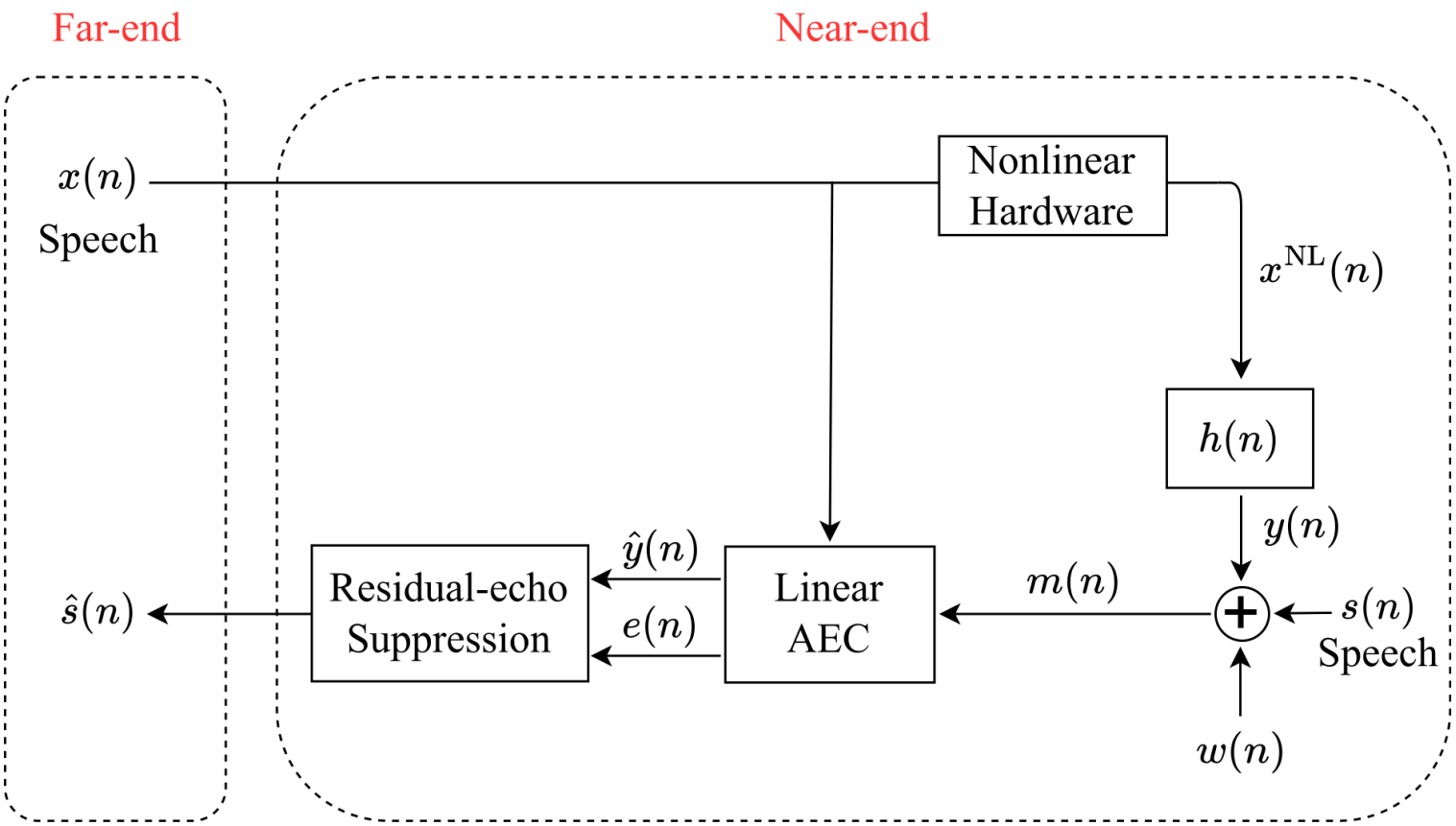

Figure 1: The traditional monophonic acoustic-echo cancellation setup.

Applications of Acoustic-echo Cancellation

Acoustic-echo cancellation is an integral part in many hands-free communication systems:

- Virtual conferencing
- Smartphone communication
- In-car communication
- Smart speakers

The Importance of Acoustic-echo Cancellation

- AEC enhances succeeding systems for speech processing, e.g., speech separation, diarization, and transcription.
- AEC prevents loss of information in full-duplex, which is an inevitable scenario in virtual communication.
- AEC improves conversation intelligibility that contributes to long-term work productivity by preventing fatigue.

The Model of Monophonic AEC: The Near-end Microphone

The near-end microphone signal $m(n)$ is given by:

$$m(n) = s(n) + w(n) + y(n), \tag{1}$$

where:

- $s(n)$ is the near-end speech,
- $w(n)$ represents additive environmental and system noises,
- $y(n)$ is a nonlinear reverberant-echo that is generated from the far-end speech, $x(n)$.

The Model of Monophonic AEC: The Echo

The far-end signal, $x(n)$, is nonlinearly distorted by electrical components that produce $x^{\mathrm{NL}}(n)$.

The microphone captures the nonlinear reverberant-echo $y(n)$ as:

$$y(n) = \left(x^{\mathrm{NL}} * h\right)(n), \tag{2}$$

where $h(n)$ is the linear echo-path from the loudspeaker to the microphone, and $*$ is the convolution operator.

The Model of Monophonic AEC: Estimating the Linear Echo-path

Traditionally, a linear AEC receives $m(n)$ as input and $x(n)$ as reference, and aims to produce the linear-echo estimation $\hat{y}(n)$:

$$\hat{y}(n) = \left(x * \hat{h}\right)(n), \tag{3}$$

where $\hat{h}(n)$ tracks the estimation of the near-end echo path $h(n)$.

The error between $h(n)$ and $\hat{h}(n)$ can be represented with $\tilde{h}(n)$:

$$\tilde{h}(n) = h(n) - \hat{h}(n). \tag{4}$$

The Model of Monophonic AEC: The Adaptation Error

The adaptation error of the linear AEC system is given by $e(n)$:

$$e(n) = m(n) - \hat{y}(n), \tag{5}$$

which can be reformulated using eqs. (1)–(4) as:

$$\begin{aligned} e(n) &= s(n) + w(n) + (y(n) - \hat{y}(n)) && (6)\\ &= s(n) + w(n) + \left(x^{\mathrm{NL}} * h\right)(n) - \left(x * \hat{h}\right)(n) && (7)\\ &= s(n) + w(n) + \left(\left(x^{\mathrm{NL}} - x\right) * h\right)(n) - \left(x * \tilde{h}\right)(n). && (8) \end{aligned}$$

The Goal of the AEC System

To suppress the echo $y(n)$ without distorting the desired speech $s(n)$.

The Challenges of AEC

Three main challenges are driven from the adaptation error:

$$e(n) = s(n) + w(n) + \left(\left(x^{\mathrm{NL}} - x \right) * h \right)(n) - \left(x * \tilde{h} \right)(n).$$

- The nonlinearity caused by nonideal hardware in mobile devices imposes $x^{\mathrm{NL}}(n) \neq x(n)$.
- The component $\left(x * \tilde{h} \right)(n)$ represents the mismatch between the real and estimated echo paths.
- The residual-echo components that remain after the nonlinear and linear stages.

Common Nonlinearities in AEC

The ongoing miniaturization of modern hardware induces nonlinearities via nonideal power amplifiers and loudspeakers.

These distortions create nonlinear relations between the microphone and far-end signal, and impede the linear AEC system.

Mismatches in the Linear AEC Stage

The most frequent change of the echo path from the loudspeaker to the microphone is due to natural human movements in the near-end.

When the echo path changes, the linear AEC begins to re-converge and adaptation error increases for this convergence period.

Residual-echo Components after Linear AEC

In double-talk, the linear AEC system often cannot replicate the echo path, and most of the echo remains.

Double-talk scenarios are extremely common in remote talks due to absence of face-to-face queues.

Nonlinear AEC: The Challenge

Electronic components in hands-free devices are constantly going miniaturization.

Miniaturization causes non-negligible nonlinear distortions between the far-end signal and the loudspeaker output.

AEC systems that assume linearity often fail in practice.

A. Ivry, B. Berdugo, and I. Cohen. Nonlinear acoustic echo cancellation with deep learning. In *Proc. Interspeech*, pages 4773—4777, 2021.

Problem Formulation

The signal transmitted to the far-end after the linear AEC is:

$$e\left(n\right)=s\left(n\right)+w\left(n\right)+\left(\left(x^{\mathrm{NL}}-x\right)*h\right)\left(n\right)-\left(x*\tilde{h}\right)\left(n\right). \tag{9}$$

Our goal is to cancel the echo by eliminating the echo components, without distorting the speech signal $s\left(n\right)$.

Existing Solutions

Existing solutions assume a nonlinear relation with fixed order P and fixed number of memory taps:

$$\hat{x}^{\mathrm{NL}}(n) = \sum_{i=0}^{P} c_i \cdot x^i(n - \delta_i), \tag{10}$$

where the coefficients hold $\forall i : c_i \in \mathbb{R}, \delta_i \in \mathbb{Z}_0^+$.

Solutions perform data-driven fit of c_i using neural networks.

Gaps in Existing Solutions

- Their nonlinear modeling does not coincide with the physical behavior of distortions that modern hands-free devices apply.
- Their nonlinear modeling is mostly parametric, i.e., it requires that memory lengths and basis functions are predetermined.
- These limitations cause sub-optimal results in real-life scenarios.

Proposed Solution

We introduced a neural network designed to model the distortions devices induce between receiving and playing the far-end signal.

We construct this network with trainable memory length and nonlinear activation functions that are optimized during training.

The network feeds the linear adaptive filter that tracks the echo path.

Schematic View of the Nonlinear AEC Scenario

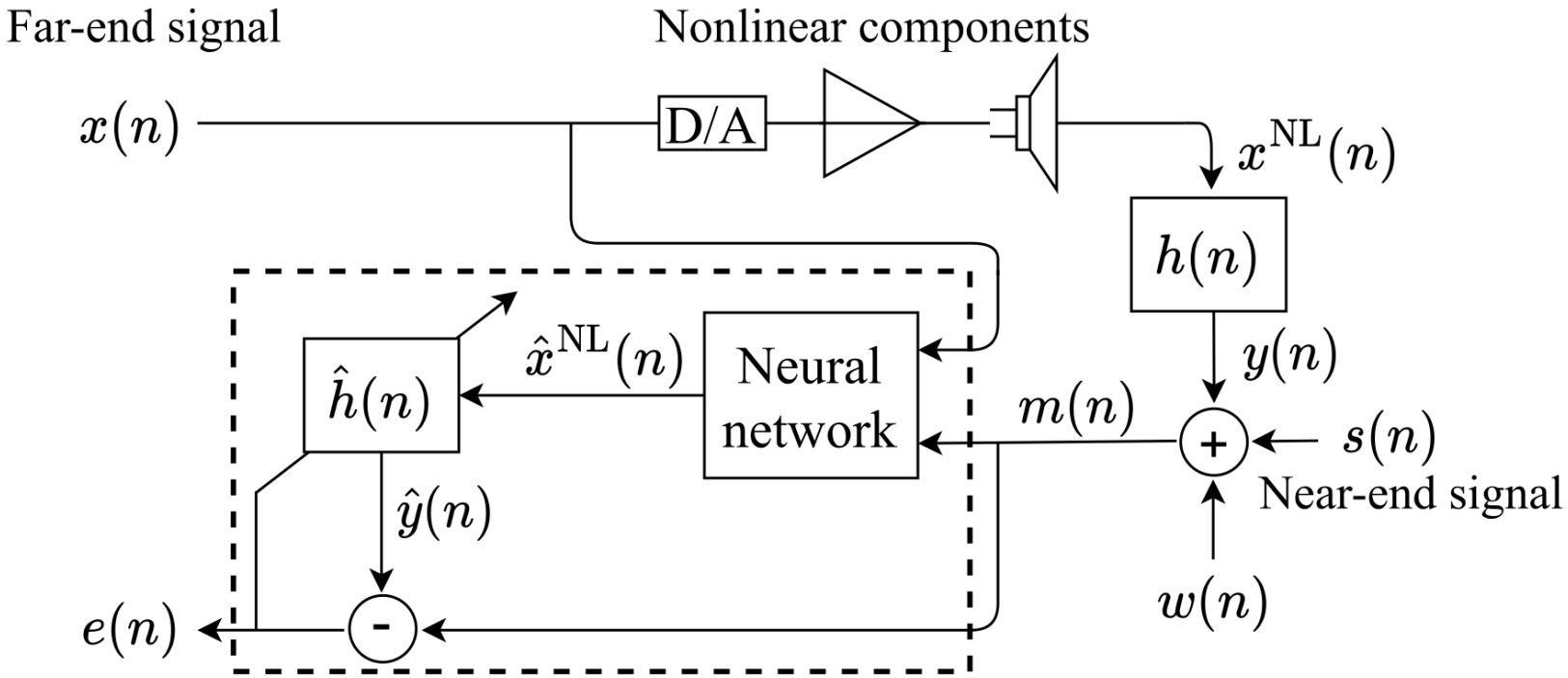

Figure 2: Nonlinear AEC scenario and proposed system (bordered).

The Proposed Deep Architecture

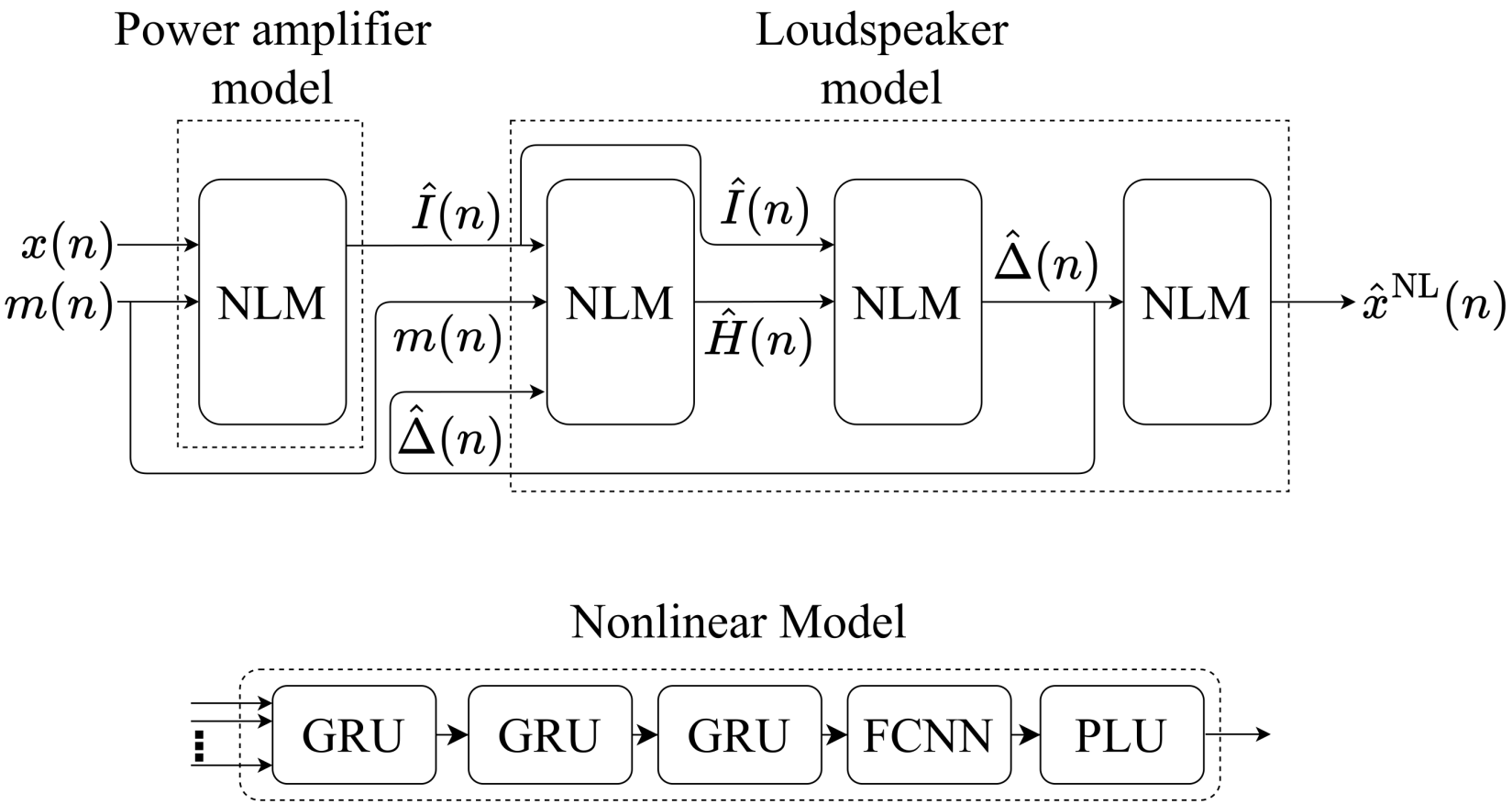

Figure 3: Proposed neural-network architecture and solution.

Experimental Setup

- We utilize 280 hours from the AEC-challenge corpus with over 5,000 different nonlinear devices.

- We consider signal-to-echo-ratios (SERs) in $[-10, 10]$ dB and signal-to-noise-ratios (SNRs) in $[0, 40]$ dB:

$$\mathrm{SER} = 10 \log_{10} \left(\|s(n)\|_2^2 / \|y(n)\|_2^2 \right), \tag{11}$$

$$\mathrm{SNR} = 10 \log_{10} \left(\|s(n)\|_2^2 / \|w(n)\|_2^2 \right). \tag{12}$$

- We measure performance with the perceptual evaluation of speech quality (PESQ) and the signal-to-distortion-ratio (SDR):

$$\mathrm{SDR} = 10 \log_{10} \left. \frac{\|s(n)\|_2^2}{\|\hat{s}(n) - s(n)\|_2^2} \right|_{\text{Double-talk}}. \tag{13}$$

Comparison to Competition and to Linear AEC

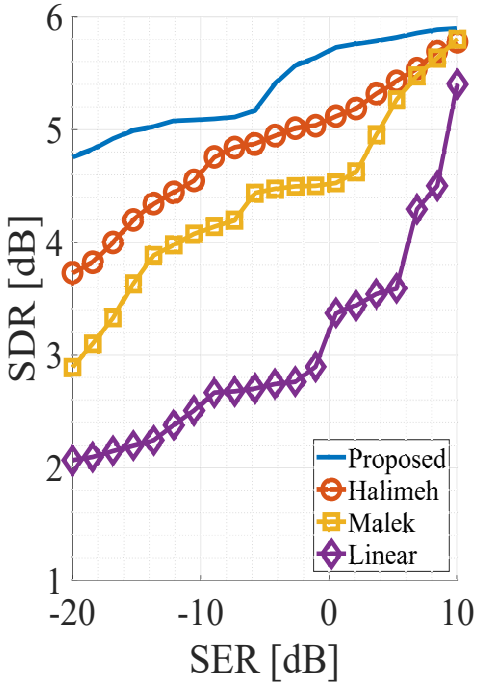

Figure 4: SDR [dB] vs. SER [dB].

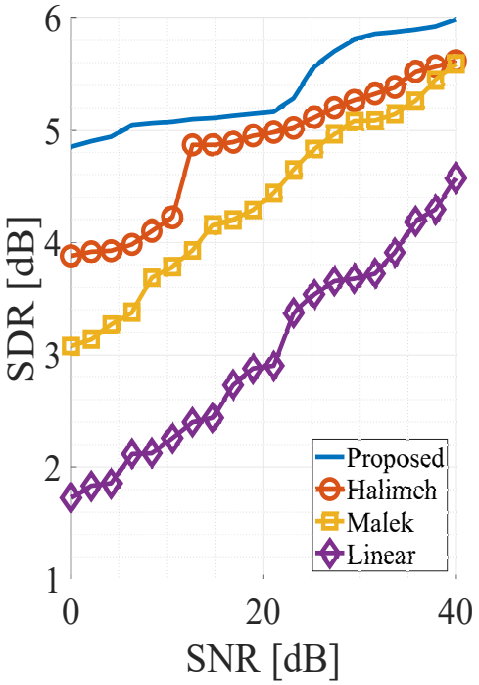

Figure 5: SDR [dB] vs. SNR [dB].

Comparison to Competition and to Linear AEC

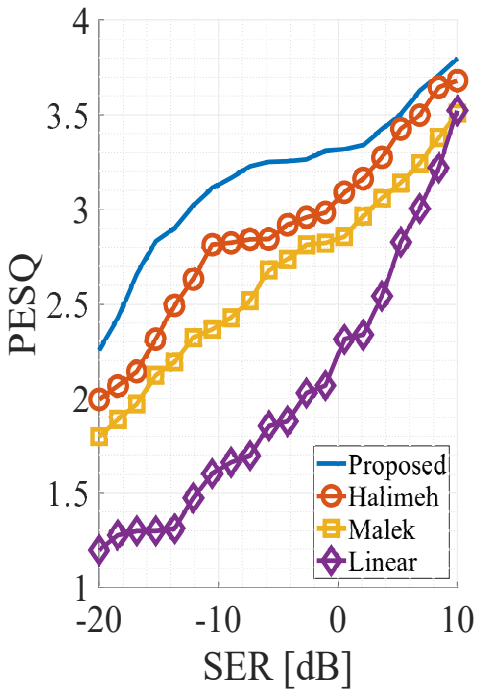

Figure 6: PESQ vs. SER [dB].

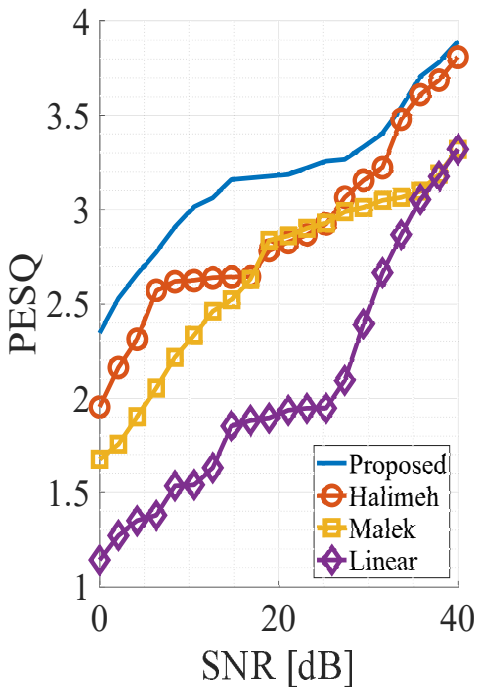

Figure 7: PESQ vs. SNR [dB].

Linear AEC: The Motivation

The popular normalized least mean-square (NLMS) adaptive filter is numerically stable and computationally efficient.

The NLMS integrates the step-size parameter that governs the often conflicting requirements of fast convergence and low misadjustment.

It is highly desirable to control the step-size during adaptation in practical scenarios of time-varying echo paths and double-talk.

A. Ivry, B. Berdugo, and I. Cohen. Deep adaptation control for acoustic echo cancellation. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 741–745, 2022.

Problem Formulation

Let $\mathbf{x}^{\mathrm{NL}}(n)$ denote the L most recent samples of the far-end signal $\mathbf{x}(n)$, after undergoing nonlinear distortions by nonideal components.

Also, let the echo path $\mathbf{h}(n)$ be modeled as a finite impulse response filter with L coefficients:

$$\mathbf{x}^{\mathrm{NL}}(n) = \left[x^{\mathrm{NL}}(n), \ldots, x^{\mathrm{NL}}(n-L+1)\right]^T, \tag{14}$$

$$\mathbf{h}(n) = \left[h_0(n), h_1(n), \ldots, h_{L-1}(n)\right]^T. \tag{15}$$

Problem Formulation

An adaptive filter with L coefficients tracks the echo path estimate $\hat{\mathbf{h}}(n)$ and echo estimate $\hat{y}(n) = \mathbf{x}^T(n)\hat{\mathbf{h}}(n)$:

$$\mathbf{x}(n) = [x(n), x(n-1), \ldots, x(n-L+1)]^T, \tag{16}$$

$$\hat{\mathbf{h}}(n) = \left[\hat{h}_0(n), \hat{h}_1(n), \ldots, \hat{h}_{L-1}(n)\right]^T. \tag{17}$$

Then, an estimate of the near-end speech signal is given by:

$$e(n) = m(n) - \hat{y}(n) = (y(n) - \hat{y}(n)) + s(n) + w(n). \tag{18}$$

Our goal is to estimate $\hat{\mathbf{h}}(n)$ and to cancel the echo by eliminating $y(n) - \hat{y}(n)$, without distorting the speech $s(n)$.

Existing Gaps of Adaptation Control Solutions

Existing approaches make restricting assumptions in practice, e.g., neglecting nonlinearities and assuming a time-invariant echo-path.

Existing methods also often require heuristics that are difficult to control in real-life scenarios.

In reality, these assumptions result in filter misadjustment and slow convergence rates during echo-path changes.

Proposed Solution: The Deep Variable Step-size

Our framework called the deep variable step-size (DVSS) minimizes the misalignment between the actual and estimated echo path.

A neural network learns the relation between the far-end, microphone, and a priori adaptation error, and the optimal step-size.

This data-driven method avoids acoustic assumptions and heuristics.

General NLMS Filter Model in Double-talk

The a priori and a posteriori error signals of the NLMS adaptation process are, respectively, given by:

$$\epsilon\left(n\right) = \mathbf{x}^{\mathrm{NL}^T}\left(n\right)\mathbf{h}\left(n\right) - \mathbf{x}^T\left(n\right)\hat{\mathbf{h}}\left(n-1\right) + s\left(n\right) + w\left(n\right), \tag{19}$$

$$e\left(n\right) = \mathbf{x}^{\mathrm{NL}^T}\left(n\right)\mathbf{h}\left(n\right) - \mathbf{x}^T\left(n\right)\hat{\mathbf{h}}\left(n\right) + s\left(n\right) + w\left(n\right). \tag{20}$$

General NLMS Filter Model in Double-talk

NLMS-type adaptive filters follow the update rule:

$$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + \mu(n)\,\mathbf{x}(n)\,\epsilon(n), \quad \hat{\mathbf{h}}(0) = \mathbf{0}^T, \tag{21}$$

where $\mu(n)$ is the step-size and $\forall n : \mu(n) \geq 0, \mu(n) \in \mathbb{R}$.

From (19)–(21), we have:

$$e(n) = \epsilon(n)\left(1 - \mu(n)\,\mathbf{x}^T(n)\,\mathbf{x}(n)\right). \tag{22}$$

General NLMS Filter Model in Double-talk

We impose echo cancellation to the a posteriori error:

$$e(n) = s(n) + w(n). \tag{23}$$

Assuming uncorrelated $s(n)$ and $w(n)$, and substituting (23) into (22):

$$s(n) + w(n) = \epsilon(n)\left(1 - \mu(n)\,\mathbf{x}^T(n)\,\mathbf{x}(n)\right). \tag{24}$$

Data-driven Generation of the Optimal Step-Size

The normalized misalignment $\mathcal{D}(n)$ is, in dB:

$$\begin{aligned}\mathcal{D}(n) &= 20\log_{10}\left(\frac{\|\mathbf{h}(n)-\hat{\mathbf{h}}(n)\|_2}{\|\mathbf{h}(n)\|_2}\right) \\ &= 20\log_{10}\left(\frac{\|\mathbf{h}(n)-\hat{\mathbf{h}}(n-1)-\mu(n)\,\mathbf{x}(n)\,\epsilon(n)\|_2}{\|\mathbf{h}(n)\|_2}\right).\end{aligned} \tag{25}$$

The optimal step-size $\mu^*(n)$ is the solution of the problem:

$$\mu^*(n) = \underset{0<\mu(n)<1}{\operatorname{argmin}}\ \mathcal{D}(n), \tag{26}$$

where $0 < \mu(n) < 1$ is a stability condition of NLMS-type algorithms.

Schematic View of the Linear AEC Scenario

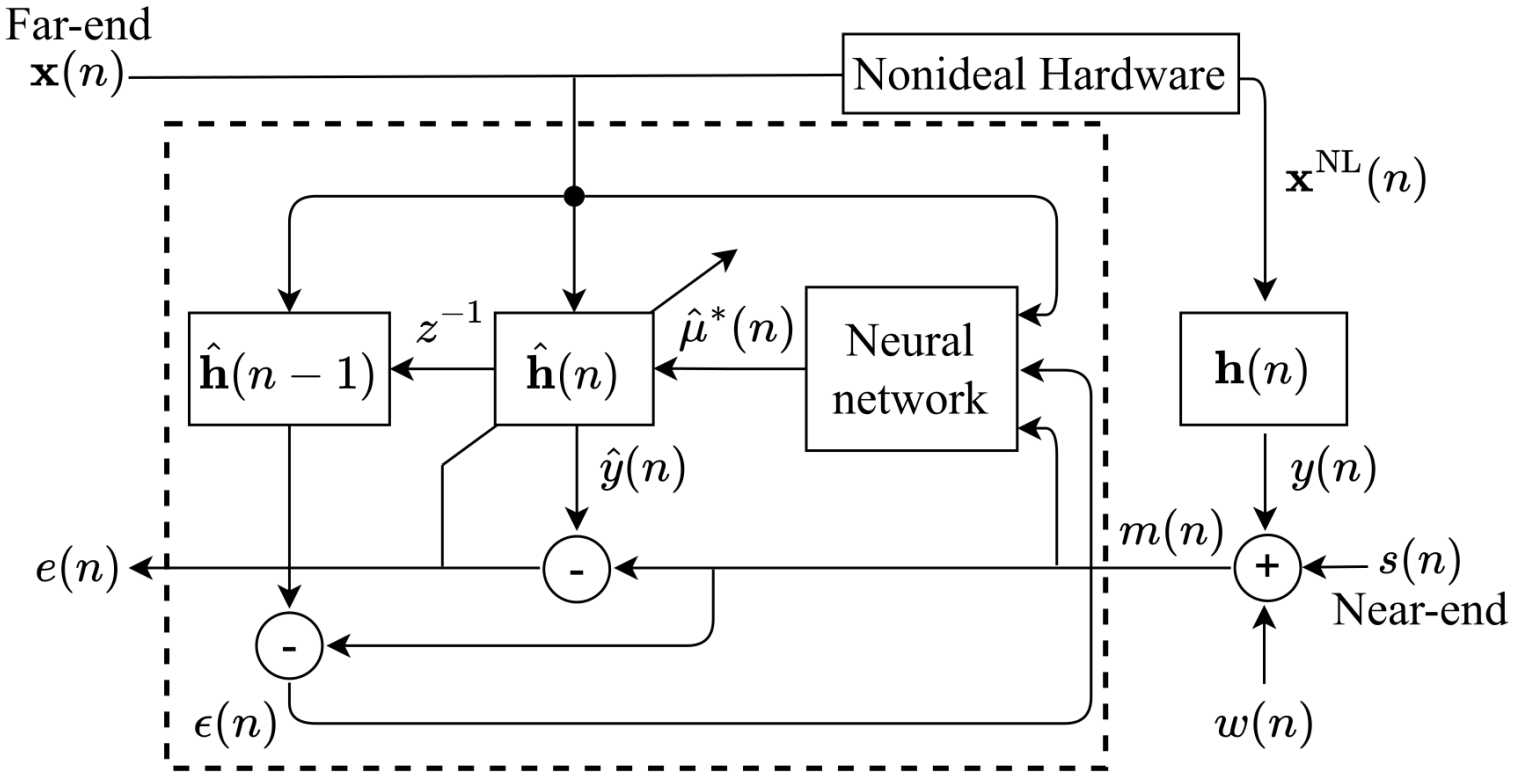

Figure 8: Linear AEC scenario and proposed system (bordered).

Experimental Settings

- We used 100 hours from the AEC-Challenge database with double-talk and frequent echo-path changes.
- Echo-to-speech-ratios (ESRs) in $[-10, 10]$ dB and echo-to-noise-ratios (ENRs) in $[0, 40]$ dB are considered:

$$\mathrm{ESR} = 10 \log_{10} \left(\|y(n)\|_2^2 / \|s(n)\|_2^2 \right), \tag{27}$$

$$\mathrm{ENR} = 10 \log_{10} \left(\|y(n)\|_2^2 / \|w(n)\|_2^2 \right). \tag{28}$$

DVSS vs. Competition and Traditional NLMS

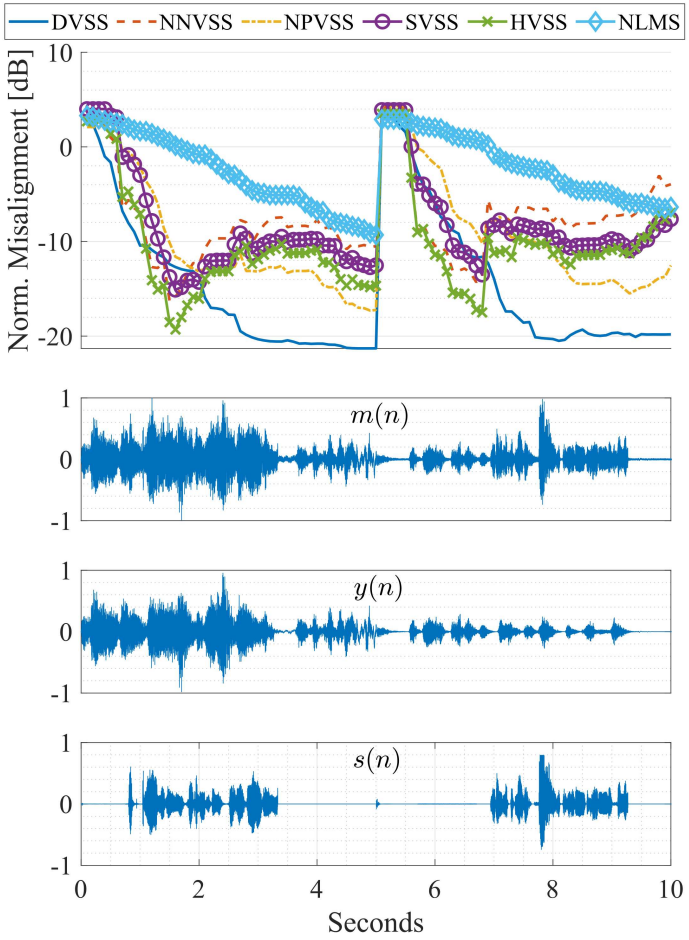

Figure 9: Convergence comparison to abrupt echo-path change at 5 s.

Residual-echo Suppression: Motivation

Residual-echo suppression (RES) is essential because the linear AEC struggles in double-talk and re-convergence, mainly due to:

- long reverberation times
- high echo-levels
- imperfect solutions

A. Ivry, B. Berdugo, and I. Cohen. Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 126–130, 2021.

Ivry, B. Berdugo, and I. Cohen. Off-the-Shelf deep integration for residual-Echo suppression. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 746–750, 2022.

Residual-echo Suppression: an Overview

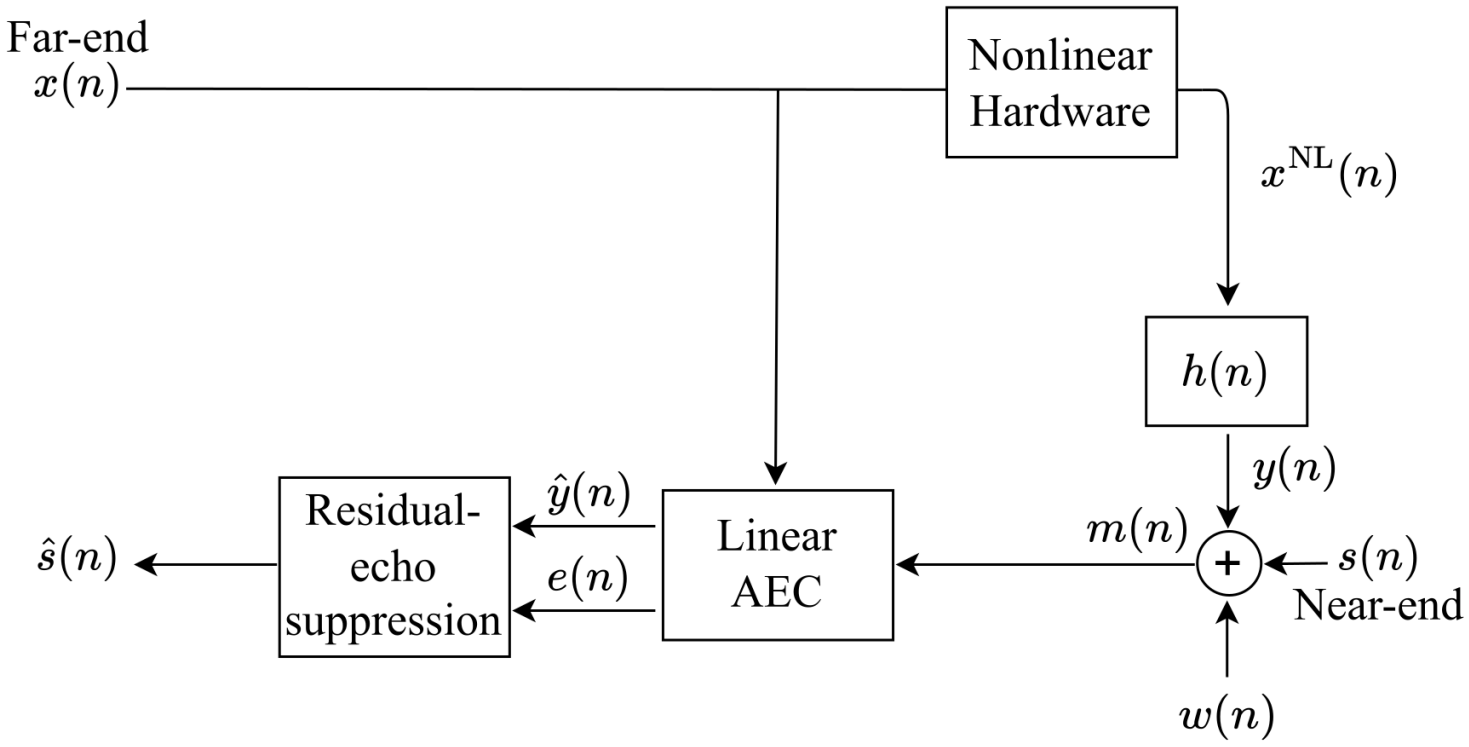

Figure 10: Residual-echo suppression scenario.

Existing Studies

Deep learning has been naturally dominating RES systems due to the non-linear nature of the problem.

Current RES systems do not address user needs, i.e., to balance speech distortion and echo suppression levels.

Our Proposed Solution

An objective function with a tunable design parameter α:

$$J(\alpha) = \left\| \widehat{S}(f) - S(f) \right\|_2^2 + \alpha \cdot \left\| \widehat{S}(f) \right\|_2^2 + \sigma^2_{\widehat{S}(f)} \cdot \mathbb{I}_{\alpha>0} \,, \tag{29}$$

where:

- $\alpha \geq 0, \alpha \in \mathbb{R}$,
- $\hat{S}(f)$ and $S(f)$ are the predicted and desired speech spectra amplitudes,
- σ^2 is the variance operator,
- $\mathbb{I}$ is the indicator function,

Objective Quality Assessment: Motivation

Human perception of speech quality is optimally evaluated using human subjective evaluation.

RES systems during double-talk are traditionally evaluated using the objective signal-to-distortion-ratio (SDR).

The SDR is affected by both desired-speech distortion and residual-echo presence, which renders it ambiguous.

A. Ivry, B. Berdugo, and I. Cohen. A user-centric approach for deep residual-echo suppression in double-talk. submitted to *IEEE Transactions of Acoustic, Speech, and Language Processing*, 2023.

A. Ivry, B. Berdugo, and I. Cohen. Objective metrics to evaluate residual-echo suppression during double-talk. In *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 101–105, 2021.

The Ambiguity of the SDR Metric

Let us consider the deep RES system as a time-varying gain $g(n)$:

$$g(n) = \left.\frac{\widehat{s}(n)}{e(n)}\right|_{\text{Double-talk}}. \tag{30}$$

The SDR is defined as:

$$\begin{aligned} \text{SDR} &= 10\log_{10}\left.\frac{\|s(n)\|_2^2}{\|s(n)-\widehat{s}(n)\|_2^2}\right|_{\text{Double-talk}} \\ &= 10\log_{10}\left.\frac{\|s(n)\|_2^2}{\|s(n)-g(n)\,e(n)\|_2^2}\right|_{\text{Double-talk}}. \end{aligned} \tag{31}$$

The SDR cannot distinct when $g(n)\,e(n)$ comprises distortion-free speech and echo, or distorted speech without echo.

Our Objective Metrics: The DSML

The desired-speech maintained level (DSML) is calculated by:

$$\mathrm{DSML} = 10 \log_{10} \left. \frac{\left\| \tilde{s}\left(n\right) \right\|_2^2}{\left\| \tilde{s}\left(n\right) - g\left(n\right) s\left(n\right) \right\|_2^2} \right|_{\text{Double-talk}}, \tag{32}$$

where $\tilde{s}\left(n\right) = \widehat{g}\left(n\right) s\left(n\right)$ compensates for biased attenuation introduced by the neural network:

$$\widehat{g}\left(n\right) = \frac{\left\langle g\left(n\right) s\left(n\right), s\left(n\right) \right\rangle}{\left\| s\left(n\right) \right\|_2^2}. \tag{33}$$

Our Objective Metrics: The RESL

The residual-echo suppression level (RESL) is derived by estimating the noisy residual-echo as $r(n) = e(n) - s(n)$, and calculating:

$$\mathrm{RESL} = 10 \log_{10} \left. \frac{\|r(n)\|_2^2}{\|g(n)\, r(n)\|_2^2} \right|_{\text{Double-talk}} . \tag{34}$$

The Parameter α and the DSML and RESL Metrics

Recalling the objective function we use for our RES system:

$$J(\alpha) = \left\| \widehat{S}(f) - S(f) \right\|_2^2 + \alpha \cdot \left\| \widehat{S}(f) \right\|_2^2 + \sigma^2_{\widehat{S}(f)} \cdot \mathbb{I}_{\alpha>0} \,, \tag{35}$$

When $\alpha = 0$, the error between the desired-speech prediction and ground truth is minimized, which reduces the speech distortion.

As α increases, smaller prediction values are generated, which reduces the level of residual echo.

The DNSMOS as an Objective Gold-standard

The objective deep noise-suppression mean-opinion score (DNSMOS) metric estimates human ratings with great accuracy.

The DNSMOS is a deep system that has learned the relation between acoustic scenarios and their subjective human evaluations.

DNSNOS Correlations vs. α Values

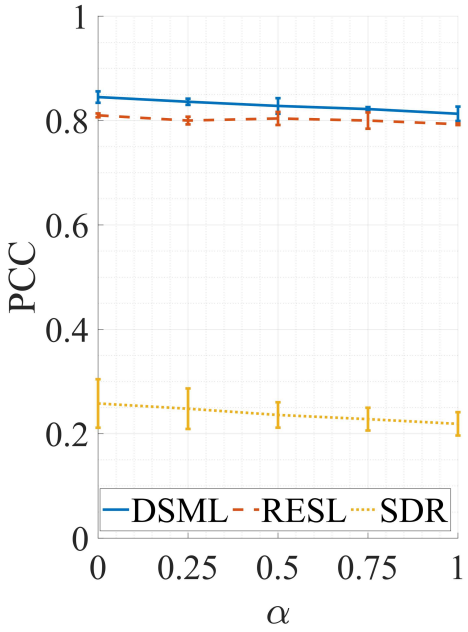

Figure 11: Pearson coeff.

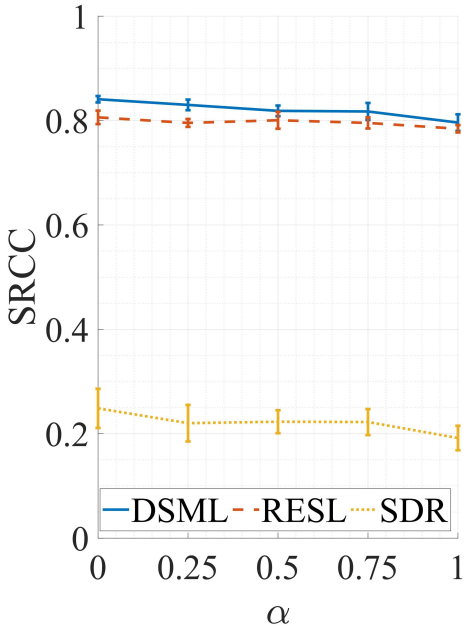

Figure 12: Spearman's rank coeff.

The DSML-RESL Trade-off vs. α Values

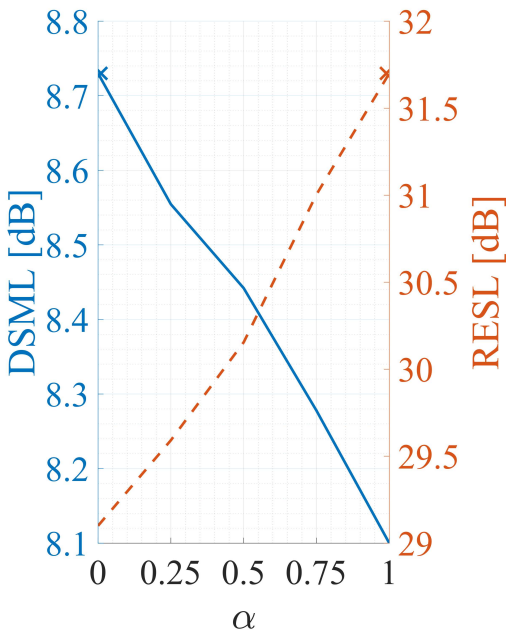

Figure 13: No echo-path change.

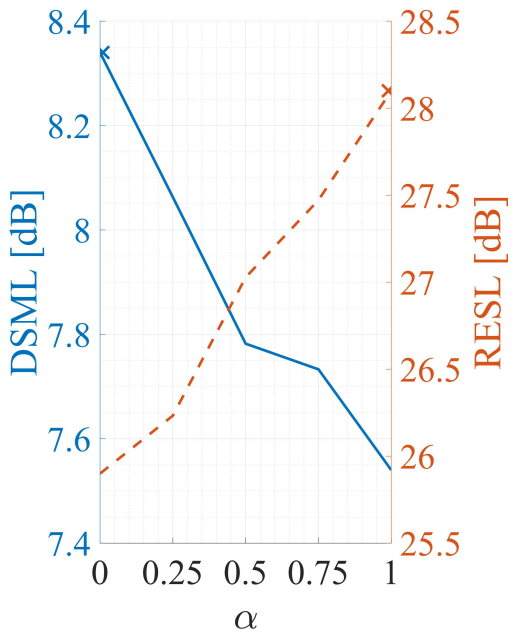

Figure 14: With echo-path change.

The DSML and RESL in Various SER Values

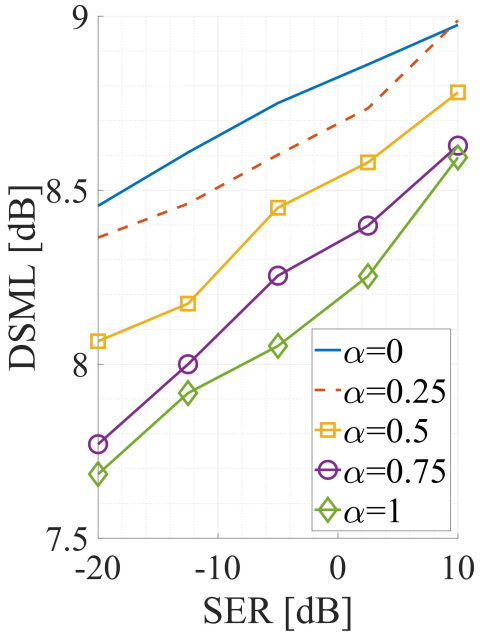

Figure 15: DSML [dB] vs. SER [dB].

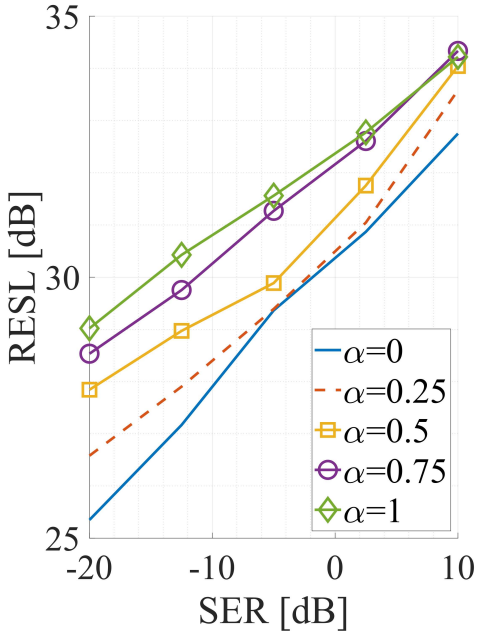

Figure 16: RESL [dB] vs. SER [dB].

The DSML and RESL in Various SNR Values

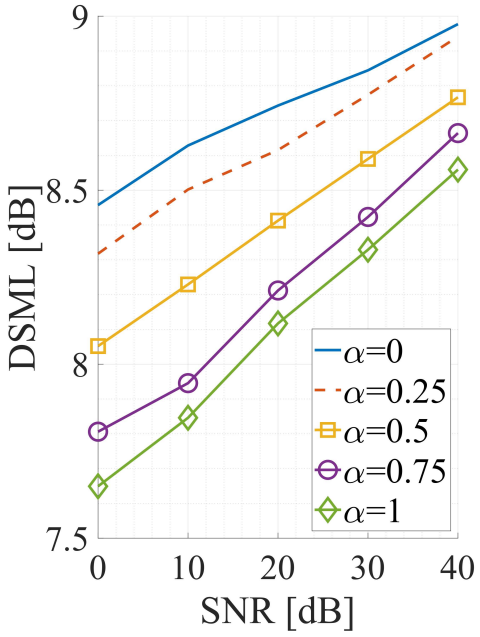

Figure 17: DSML [dB] vs. SNR [dB].

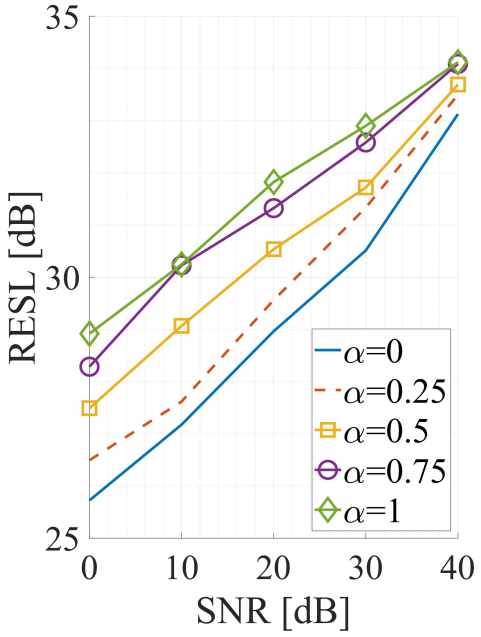

Figure 18: RESL [dB] vs. SNR [dB].

Stereophonic AEC

We introduced a DVSS framework that derives the optimal step-size in the complex domain using the widely-linear model.

We extended the RESL and DSML metrics and showed they outperform the SDR using the acoustic-echo cancellation MOS.

A. Ivry, B. Berdugo, and I. Cohen. Objective metrics to evaluate residual-Echo suppression during double-talk in the Stereophonic Case. In *Proc. Interspeech*, pages 5348–5352, 2022.

A. Ivry, B. Berdugo, and I. Cohen. Deep adaptation control for stereophonic acoustic echo cancellation. accepted to *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

Voice-activity Detection

We extended our work for Voice-Activity Detection (VAD) by introducing a VAD system that separates speech from silence using nonlinear dimensionality reduction and geometric patterns.

Our VAD has shown promising results in real acoustic environments of reverberation, noises, and transients.

A. Ivry, I. Cohen, and B. Berdugo. Voice activity detection for transient noisy environment based on diffusion nets. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):254–264, 2019.

A. Ivry, B. Berdugo, and I. Cohen. Evaluation of deep-learning-based voice activity detectors and room impulse response models in reverberant environments. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 406–410, 2020.

Examine Real-valued Speech Signal Representations

Decomposing the speech waveform signal into its frequency sub-bands using a real-valued transform can be efficient to:

- Enable a utilization of waveform-based deep learning models.
- Preserve phase information.
- Associate every sub-band with a lower sample frequency than the original signal, which may reduce computational cost.

Develop a framework for Real-time Waveform-based Processing

Equipped with a sub-band decomposition of the signal, one can:

- Decompose existing speech-based systems into smaller and more efficient sub-systems.
- Process each sub-band separately and independently by a smaller waveform-based architecture, and merge their outcomes.
- Utilize a smaller computational load for each sub-band adequate for embedding on real-time mobile platforms.

Thank you!
Any Questions?