Deep Residual Echo Suppression with a Tunable Tradeoff Between Signal Distortion and Echo Suppression

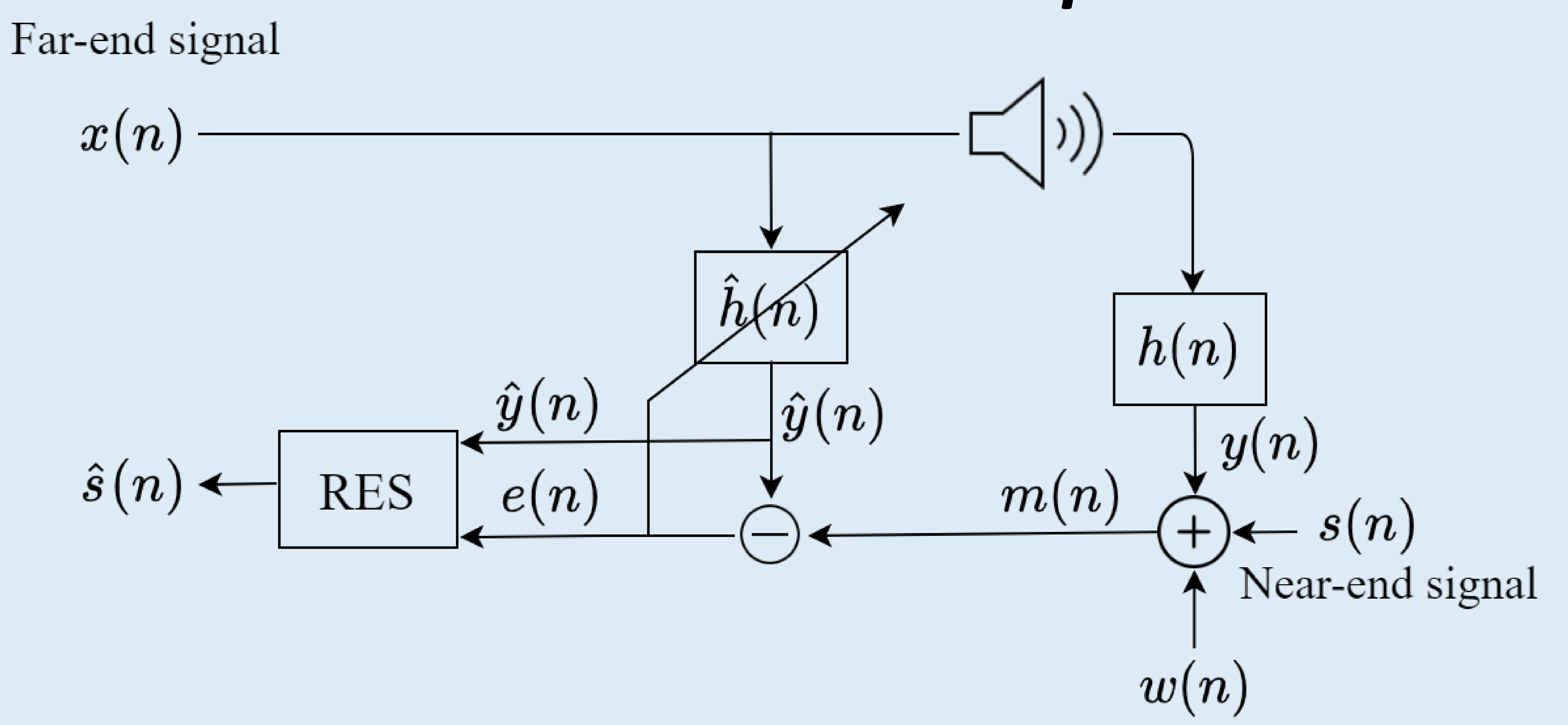
Amir Ivry, Israel Cohen, and Baruch Berdugo | IEEE ICASSP '21

We propose a residual echo suppression method using a UNet neural network that directly maps the outputs of a linear acoustic echo canceler to the desired signal in the spectral domain. This system embeds a design parameter that allows a tunable tradeoff between the desired-signal distortion and residual echo suppression in double-talk scenarios. The system employs 136 thousand parameters and requires 1.6 Giga floating-point operations per second and 10 Mega-bytes of memory. This implementation satisfies both the timing requirements and the computational and memory limitations of on-device applications.

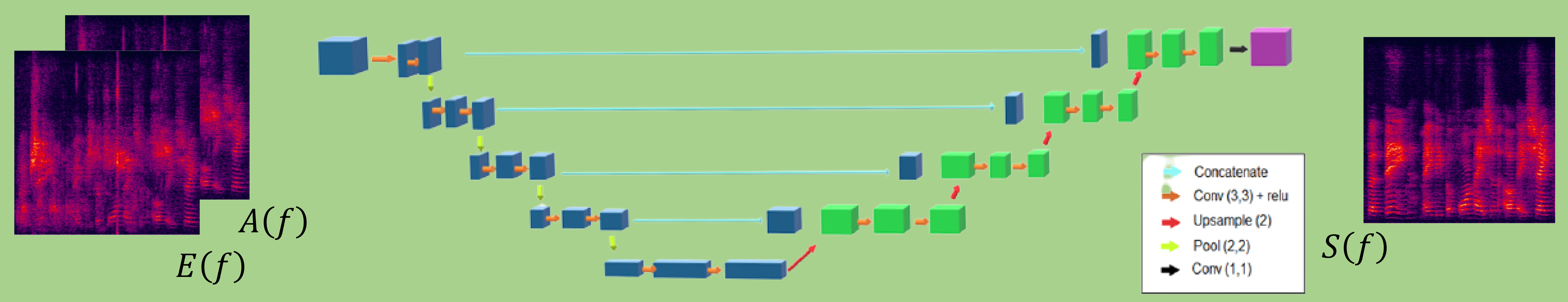
1 *Motivation*

- The presence of acoustic echo can lead to degradation in intelligibility and quality of conversation, since the far-end speaker can hear their own voice while speaking, and near-end speech can be screened.
- Conventional acoustic echo cancelers (AECs) do not model non-linearities in the echo path, and generally introduce a mismatch between true and estimated echo paths during convergence. Thus, the residual echo must be suppressed by a dedicated system.
- We introduce a residual echo suppression (RES) method with a dual-channel input and single-channel output UNet neural network that directly maps the outputs of a linear AEC to the desired near-end signal in the short-time Fourier transform (STFT) domain.
- A design parameter that allows balance between echo reduction and signal distortion is embedded in the UNet objective function, minimized during the training process.
- We conduct experiments with over 160 h of data that was acquired from the AEC challenge database and from independent recordings in real conditions.

2 *Acoustic Echo Cancellation Setup*



3 *Proposed System* in the STFT Domain



- Objective function to minimize is given by:

$$\left\|S(f)-\hat{S}(f)\right\|^2+\boldsymbol{\alpha}\cdot\left\|\hat{S}(f)\right\|^2+0.1\cdot\sigma^2_{\hat{S}(f)}$$

- $\boldsymbol{\alpha}\geq 0$ is the tunable design parameter

4 *Performance Measures*

Measure name	Abbr.	Definition	Scenario
Echo return loss enhancement	**ERLE** [dB]	$10\log_{10}\frac{\|e\|^2}{\|\hat{s}\|^2}$	Single-talk Far-end only
Signal-to-artifacts-ratio	**SAR** [dB]	$10\log_{10}\frac{\|e\|^2}{\|s-\hat{s}\|^2}$	Single-talk Near-end only
Signal-to-distortion-ratio	**SDR** [dB]	$10\log_{10}\frac{\|error\|^2}{\|s-\hat{s}\|^2}$	Double-talk

5 *Database*

- AEC challenge database:
  - 10k synthetic scenarios
  - Real recordings with over 2,500 devices
  - SER between [-10, 10] dB
- Real recordings in lab with SER between [-20,-10] dB
- Overall – over 160 hours of real and simulated data

6 *Results*

No echo path change	UNet		Zhang		Carbajal	
	mean	std	mean	std	mean	std
PESQ	**3.61**	**0.24**	2.51	0.41	2.47	0.55
SDR	**7.1**	**0.8**	4.3	1.4	4.1	1.6
ERLE	**40.1**	**2.1**	35.7	3.3	21.5	3.6
SAR	**8.8**	**0.8**	4.8	1.1	4.5	1.1

Echo path change	UNet		Zhang		Carbajal	
	mean	std	mean	std	mean	std
PESQ	**3.3**	**0.25**	2.35	0.45	2.05	0.7
SDR	**7**	**0.8**	2.71	1.9	2.8	1.65
ERLE	**38.5**	**2.45**	28.3	3.9	18	4
SAR	**8.8**	**0.95**	4.3	1.35	4.4	1.3

Comparison of α values	$\alpha=0$		$\alpha=0.5$		$\alpha=1$	
	mean	std	mean	std	mean	std
PESQ	3.61	0.24	3.54	0.29	3.45	0.35
SDR	7.1	0.8	6.9	0.95	6.8	1.1
ERLE	40.1	2.1	41.9	2.2	43.5	2.2
SAR	8.8	0.8	8.4	0.8	8.2	0.9

Before convergence	UNet		Zhang		Carbajal	
	mean	std	mean	std	mean	std
PESQ	**2.88**	**0.5**	2.02	0.8	1.91	0.95
SDR	**4.9**	**1.4**	2.6	2.1	1.1	1.7
ERLE	**31.8**	**2.9**	23.3	4.1	15.2	4.9
SAR	**8.5**	**1**	3.7	1.45	3.7	2.7

7 *Conclusion*

- We introduced an RES method based on a UNet neural network that receives the outputs of a linear AEC in the STFT domain.
- Consists of 136k parameters that require 1.6 Gflops and 10 MB of memory. Satisfies hands-free communication timing constraints on neural processor.
- In addition, we integrate into the system a tunable tradeoff between echo suppression and signal distortion using a built-in design parameter.