Acoustic Echo Cancellation Combined with Deep-Learning-Based Residual Echo Suppression

Eran Shachar

TECHNION
Israel Institute of Technology

M.Sc. Thesis Seminar

The Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering

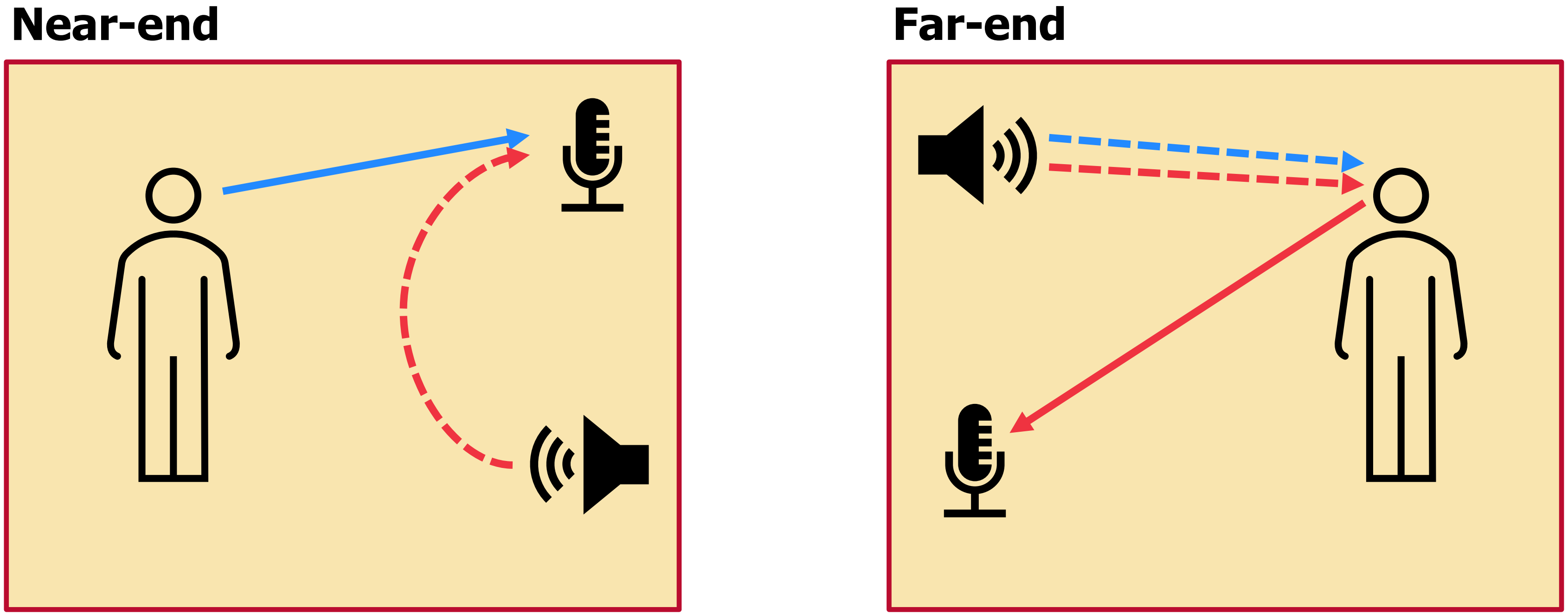
Supervisors: Prof. Israel Cohen, Dr. Baruch Berdugo

TECHNION
Israel Institute of Technology

Background & Motivation

TECHNION
Israel Institute of Technology

Acoustic Echo

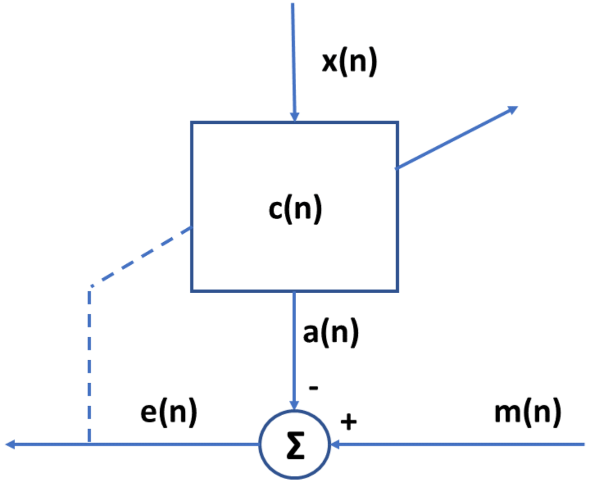


Acoustic Echo

- Degrades conversation quality
- Can appear in any full-duplex telecommunication system
- Common situations:
  - Cell phone conversation when the loudspeaker volume is high
  - Meeting in a conference room with remote participants
- Solution – Acoustic Echo Cancellation

Linear Acoustic Echo Cancellation

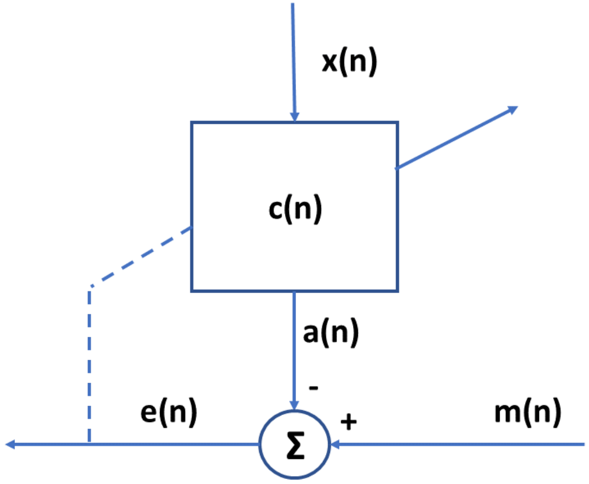
- Traditionally, linear adaptive filters are employed for acoustic echo cancellation



TECHNION
Israel Institute of Technology

Linear Acoustic Echo Cancellation

- x(n) – reference signal (far-end speech)
- m(n) – microphone signal
- **c(n)** – filter's coefficients
  vector of length N
- a(n) – filter's output
  (estimated echo signal)
- e(n) – error signal
  (estimated near-end)



Linear Acoustic Echo Cancellation

- The adaptive filter's coefficients are adapted using the error signal according to some (usually iterative) optimization algorithm
- Common algorithm – Least Mean Squares (LMS)

Algorithm 2.1 The LMS algorithm

Parameters: μ - step size, N - number of filter coefficients
for $n = 0, 1, 2, ...$ **do**
$\quad \mathbf{c}(n) = [c_1(n), ..., c_N(n)]^T$
$\quad \mathbf{x}_N(n) = [x(n), x(n-1), ..., x(n-N+1)]^T$
$\quad e(n) = m(n) - a(n) = m(n) - \mathbf{c}^T(n)\mathbf{x}_N(n)$
$\quad \mathbf{c}(n+1) = \mathbf{c}(n) + 2\mu e(n)\mathbf{x}_N(n)$
end for

Linear Acoustic Echo Cancellation

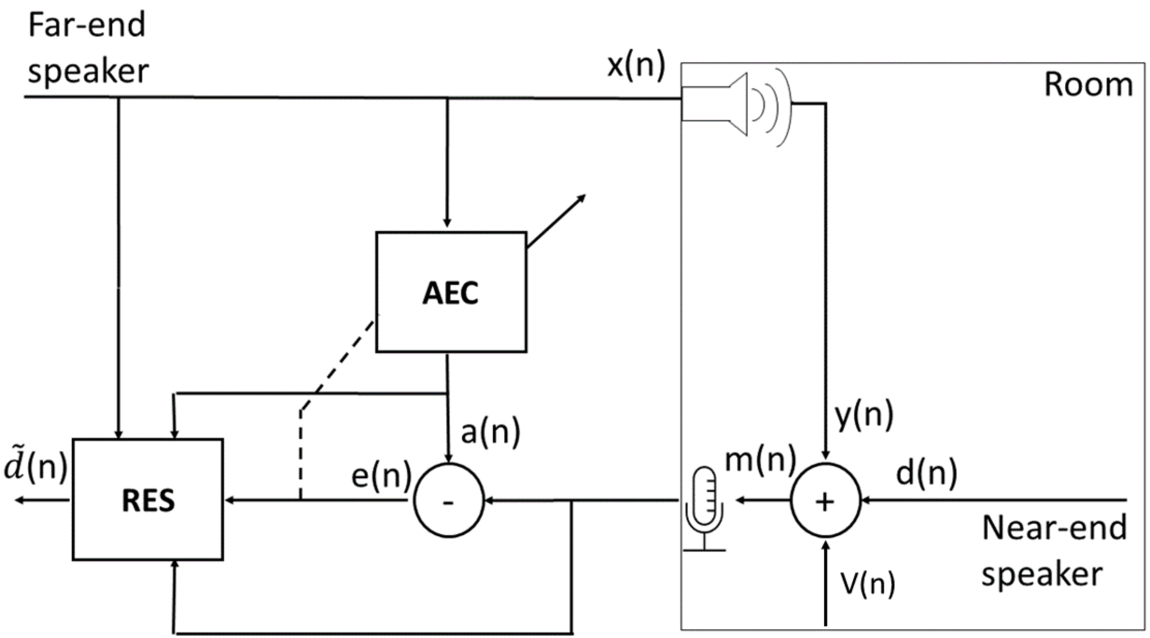
Linear Adaptive Filters

- ✓ Fast
- ✓ Minimum distortion
- ✗ Linear

Residual Echo Suppression

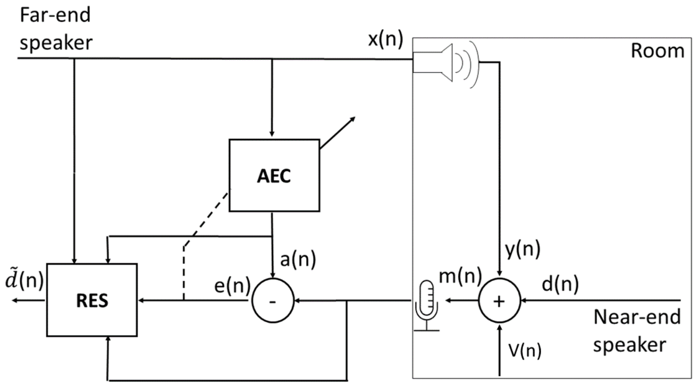
- Residual, nonlinear echo components remain at the output of the linear AEC
- Usually, the residual echo still interferes
- Solution – Residual Echo Suppression (RES)

Residual Echo Suppression



Residual Echo Suppression

- RES operates on the outputs of the linear AEC, and, possibly, also on the reference and microphone signals
- Traditionally, based on nonlinear adaptive filters
- Recently, deep-learning networks



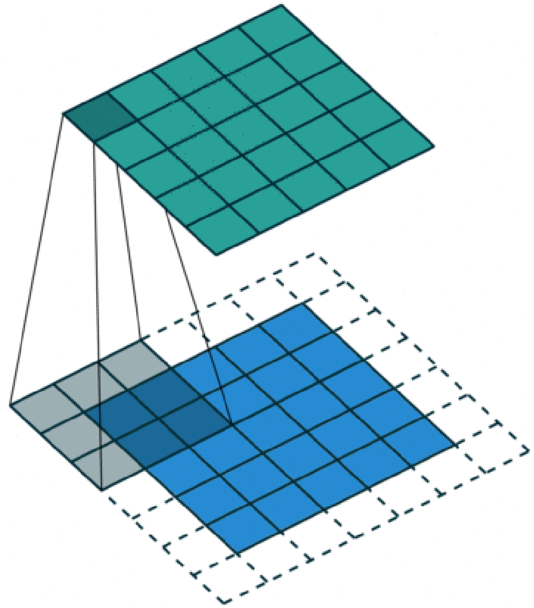
Deep-Learning AEC/RES

- In recent years, DNNs have achieved unprecedented performance in many fields:
  - Computer Vision
  - Natural Language Processing
  - Audio and Speech Processing
  - ...
- Deep-learning based acoustic echo cancellation / residual echo suppression has also seen abundant research

Deep-Learning Components

2D Convolution

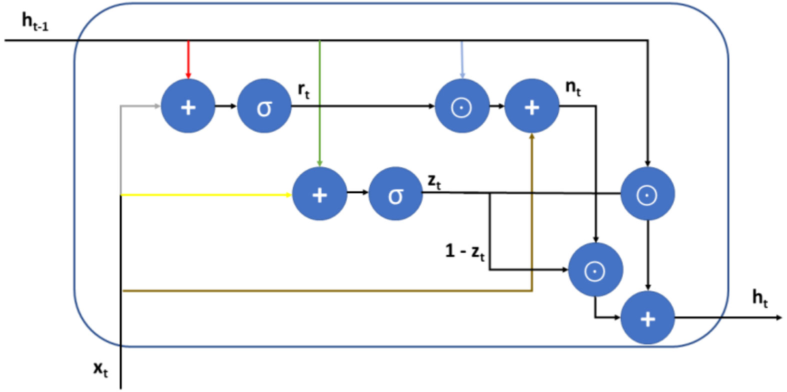
- Basic building block in many DNNs
- Operates on 2D inputs (images, spectrograms, etc.)
- Excel at modeling local spatial relationships
- Unable to memorize previous inputs



Deep-Learning Components

LSTM, GRU

- Recurrent Neural Networks (RNNs) – used to model long sequential data
- Can memorize previous inputs
- Long Short-Term Memory (LSTM) and its lighter version, Gated Recurrent Unit (GRU) are common



TECHNION
Israel Institute of Technology

Deep-Learning Components

Many more…

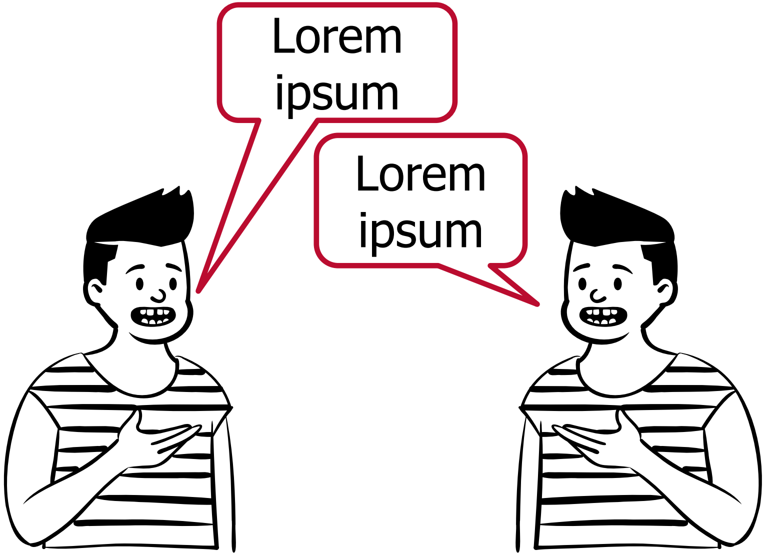
- There are many more important DNN components:
  - Activation functions
  - Normalization layers
  - Optimizers
  - …

Motivation

- Previous studies exhibit excellent performance using sophisticated methods
- Little attention has been paid to the importance of the proper choice of linear AEC in deep-learning-based RES systems
  - Specifically, when employing a pre-trained speech denoiser as an alternative to a RES (more on that later)

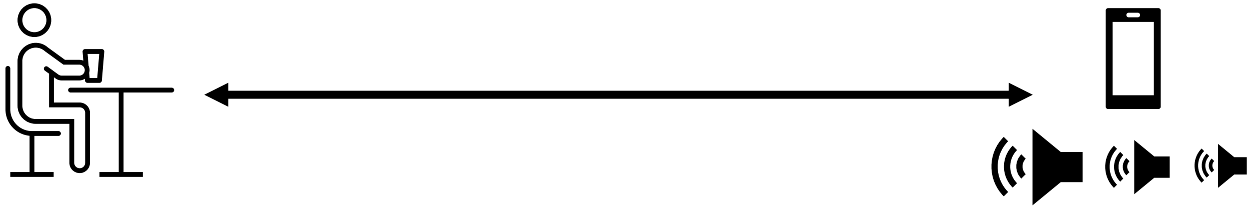
Motivation

- The most challenging situation – double-talk
- Previous studies integrate a double-talk detector (DTD) in RES systems
- None study its efficiency or effect on performance



Motivation

- None of the previous studies focus on the low signal-to-echo ratio (SER) scenario, i.e., when the echo's energy is substantially higher than the near-end speech's energy
- For example – conversation over a cell phone, when the loudspeaker's volume is high, and the near-end speaker stands far away



TECHNION
Israel Institute of Technology

Acoustic Echo Cancellation with the Normalized Sign-Error Least Mean Squares Algorithm and Deep Residual Echo Suppression

TECHNION
Israel Institute of Technology

Overview

- Normalized Sign-error Least Mean Squares (NSLMS) vs. Normalized Least Mean Squares (NLMS)
- Deep Complex Convolution Recurrent Network (DCCRN) as RES
- Pre-trained speech denoiser as RES

NLMS

$$\mathbf{c}(n+1) = \mathbf{c}(n) + \frac{\alpha(n)e(n)\mathbf{x}_N(n)}{||\mathbf{x}_N(n)||^2}$$

TECHNION
Israel Institute of Technology

NSLMS

$$\mathbf{c}(n+1) = \mathbf{c}(n) + \frac{\alpha(n)\,\mathrm{sgn}(e(n))\,\mathbf{x}_N(n)}{\|\mathbf{x}_N(n)\|^2}$$

NLMS vs. NSLMS

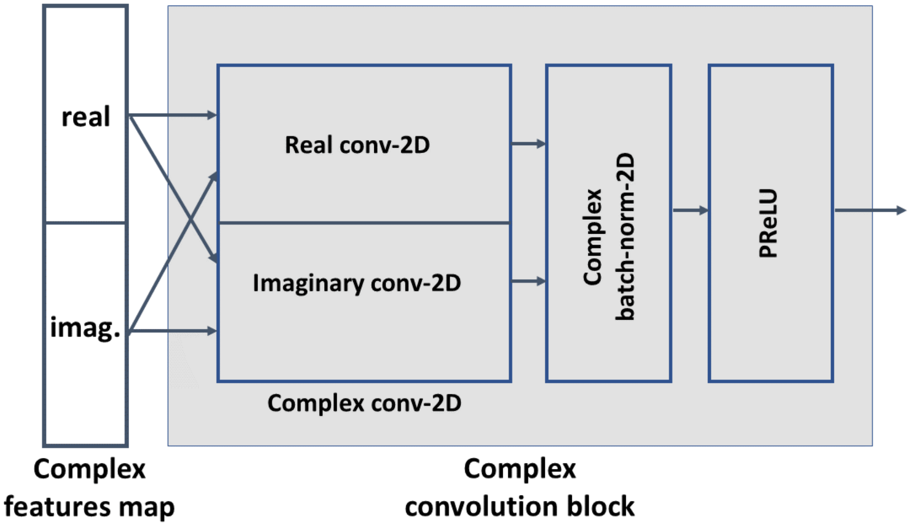
Freire and Douglas "Adaptive Cancellation of Geomagnetic Background Noise Using a Sign-error Normalized LMS Algorithm"	**Pathak et al.** "Real Time Speech Enhancement for the Noisy MRI Environment"
• Cancellation of geomagnetic background noise in magnetic anomaly detection systems • Demonstrated the superiority of NSLMS over NLMS	• Speech enhancement in noisy MRI environments • Demonstrated the superiority of NSLMS over NLMS • Residual noise produced by NSLMS has characteristics of white noise, NLMS output is more structured

DCCRN

Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement,” preprint arXiv, 2020.

- Proposed for speech enhancement
- Operates in the STFT domain
- Complex, convolutional, encoder-decoder structure, and a complex LSTM
- Estimates a complex ratio mask (CRM)

Complex 2D Convolution Block

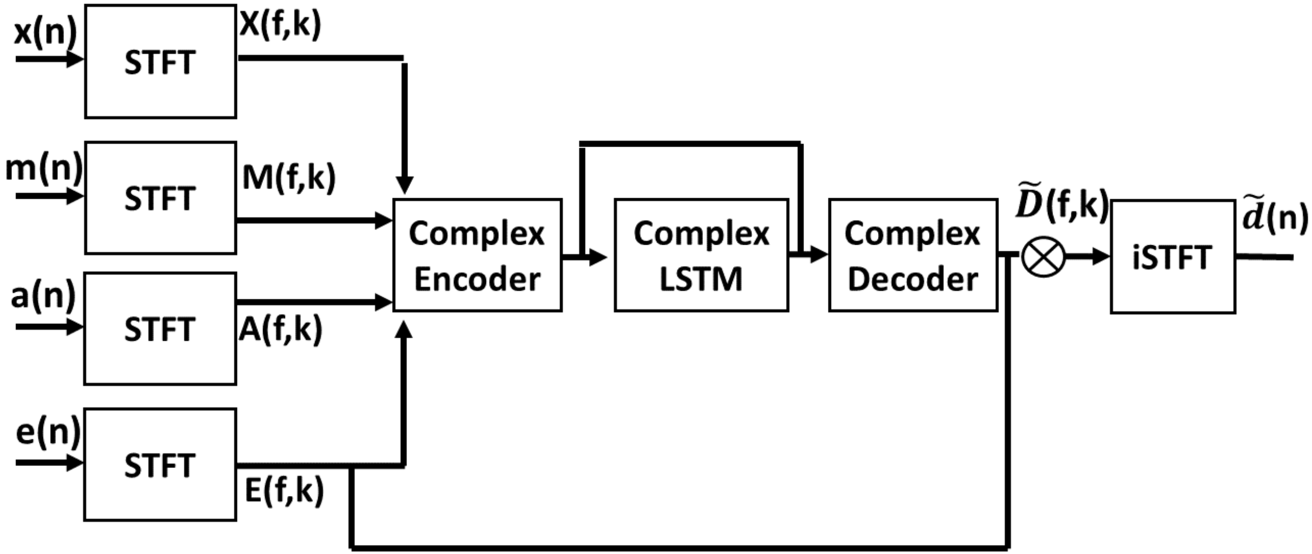


$$O_c = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r)$$

Complex LSTM

$$F_c = (\mathrm{LSTM}_r(X_r) - \mathrm{LSTM}_i(X_i)) + j(\mathrm{LSTM}_i(X_r) + \mathrm{LSTM}_r(X_i)) \,.$$

DCCRN RES



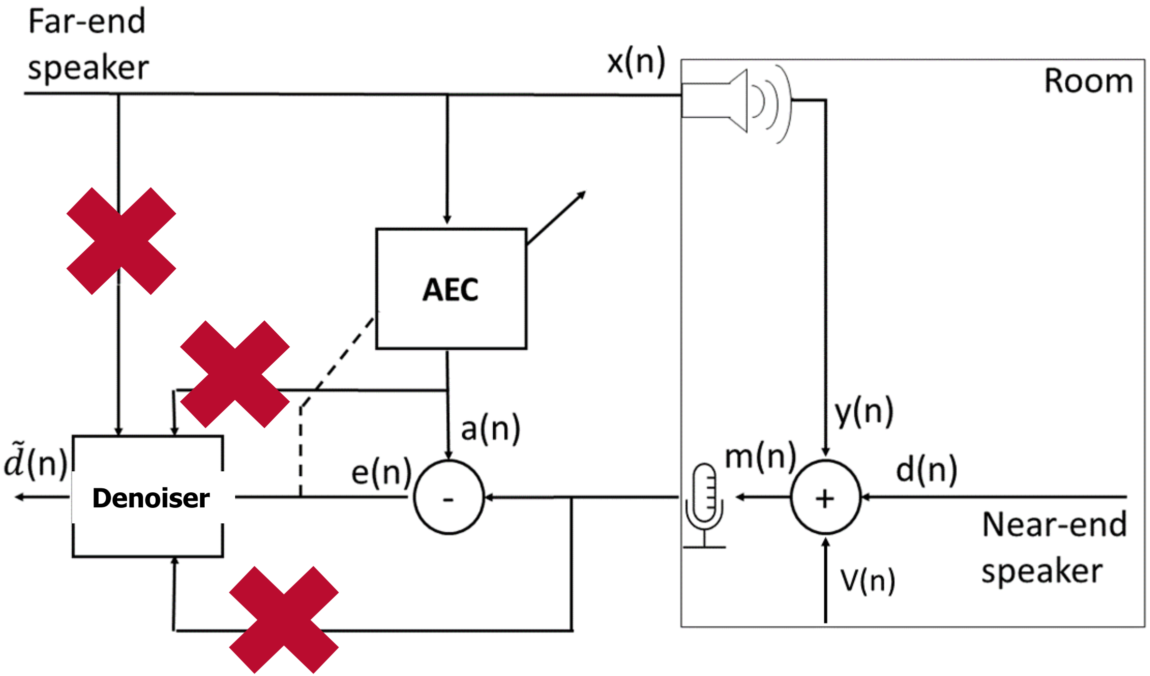
x(n)
STFT
X(f,k)
m(n)
STFT
M(f,k)
a(n)
STFT
A(f,k)
e(n)
STFT
E(f,k)
Complex Encoder
Complex LSTM
Complex Decoder
$\widetilde{D}$(f,k)
iSTFT
$\widetilde{d}$(n)

Speech Denoiser

A. Defossez, G. Synnaeve, and Y. Adi, “Real Time Speech Enhancement in the Waveform Domain,” preprint arXiv, 2020

- Proposed for speech enhancement
- Operates in the time domain
- Real, convolutional, encoder-decoder structure with an LSTM
- Pre-trained on a large and diverse corpus with many types of noise and diverse conditions
- Fine-tuned on the RES dataset

Speech Denoiser RES



Performance Measures

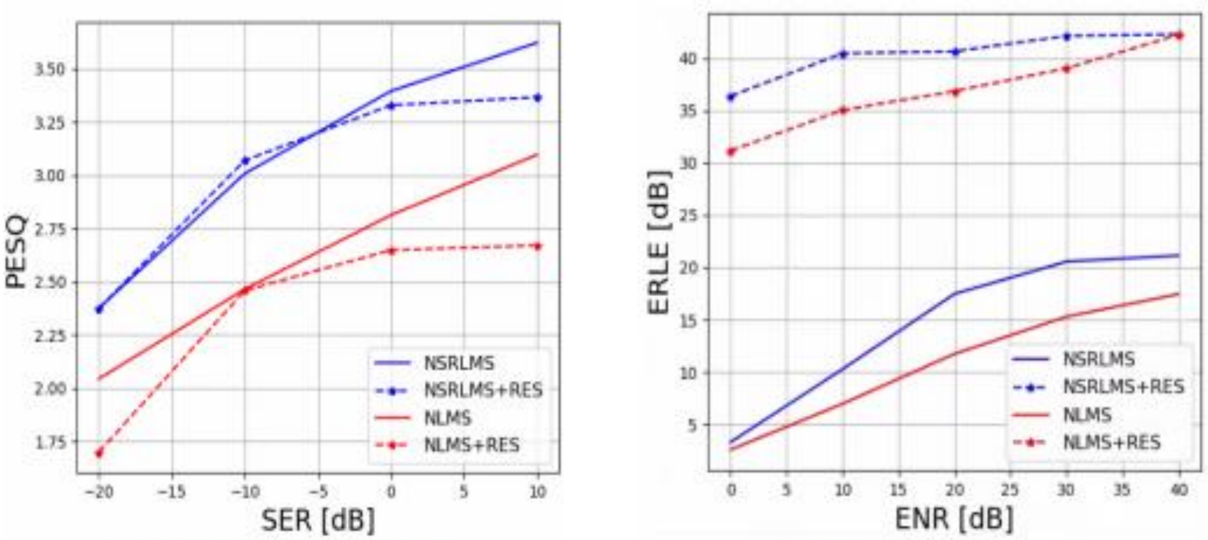
Far-end only	Near-end only / Double-talk	
ERLE	DNSMOS	PESQ
• Echo Return Loss Enhancement • Measured in dB • Measures echo reduction between the microphone and enhanced signals $\mathrm{ERLE} = 10\log_{10}\frac{\|\|m(n)\|\|^2}{\|\|\hat{d}(n)\|\|^2}$	• Deep Noise Suppression Mean Opinion Score • Developed for noise suppressors • DNN trained to predict subjective human ratings • Non-intrusive • Range: [1, 5]	• Perceptual Evaluation of Speech Quality • Based on an algorithm designed to approximate a subjective evaluation of a degraded audio sample • Intrusive • Range: [-0.5, 4.5]

Results

Table 3.1: Performance comparison of the different systems. FE stands for far-end only, NE stands for near-end only, and DT stands for double-talk.

	ERLE	DNSMOS		PESQ	
	FE	DT+NE	DT	DT+NE	DT
NLMS	16.60	2.81	2.62	3.33	2.42
NSLMS	**21.17**	**2.86**	**2.71**	**3.66**	**2.98**
NLMS+ Denoiser	32.63	2.72	2.44	3.23	2.32
NSLMS+ Denoiser	**39.44**	**2.84**	**2.65**	**3.63**	**3.13**
NLMS+ RES	38.55	2.76	2.46	3.34	2.53
NSLMS+ RES	**40.34**	**2.84**	**2.64**	**3.70**	**3.11**

Results



(a) PESQ in double-talk only scenario

(b) ERLE in far-end only scenario

Figure 3.3: Comparison of the linear AECs with or without RES.

Summary

- An echo suppression system based on the NSLMS-AEC and the DCCRN speech enhancement model
- NSLMS outperforms the common NLMS
- NSLMS produces residual echo that is more akin to noise than speech
- DCCRN RES outperforms the larger, pre-trained speech denoiser
- NSLMS brings bigger performance improvement over NLMS for the speech denoiser (more akin to noise…)

Double-Talk Detection-Aided Residual Echo Suppression via Spectrogram Masking and Refinement

Eran Shachar, Israel Cohen, and Baruch Berdugo. Double-talk detection-aided residual echo suppression via spectrogram masking and refinement. *Acoustics,* 4(3):637-655, 2022

TECHNION
Israel Institute of Technology

Overview

- Two-stage residual echo suppression system focused on the low SER scenario
- 1st stage – spectrogram masking and double-talk detection
- Study proper integration of DTD with the masking mode
- 2nd stage – spectrogram refinement

Masking and Inpainting for Speech Enhancement

"Masking and Inpainting: A Two-Stage Speech Enhancement Approach for Low SNR and Non Stationary Noise," in IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 6959–6963, 2020.

- A two-stage approach to low signal-to-noise ratio (SNR) speech enhancement
- 1st stage – speech spectrogram masking, removes most of the noise, and some speech
- 2nd stage – spectrogram inpainting, reconstruct speech that was lost in the masking stage

1st Stage: Spectrogram Masking, Double-talk Detection

- Employs the U-Net architecture – lighter, faster, provides the same performance
- Integrates a DTD
- DTD operates prior to masking. A feature representation is learned from the DTD predictions, and is used by the masking network

1^{st} Stage: Spectrogram Masking, Double-talk Detection

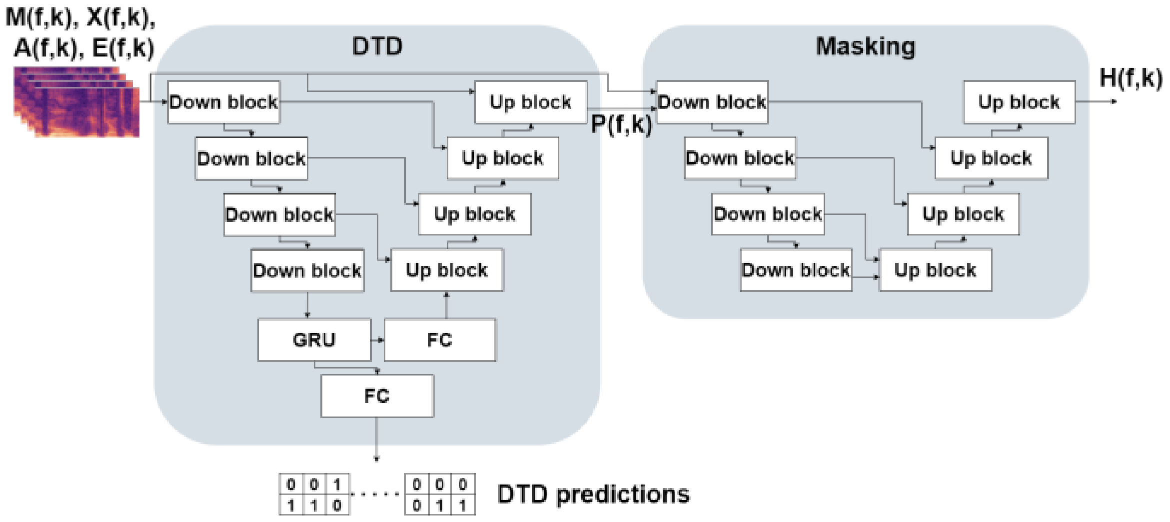


Figure 4.1: Structure of the double-talk detector (DTD) and masking model architecture. FC stands for fully connected.

2^{nd} Stage: Spectrogram Refinement

- Masking alone is not sufficient to both suppress the echo and preserve near-end speech quality
- Contrary to speech enhancement, here we want to separate speech from speech and not noise from speech
- Renders the inpainting operation much more challenging. Instead, we perform spectrogram refinement

2^{nd} Stage: Spectrogram Refinement

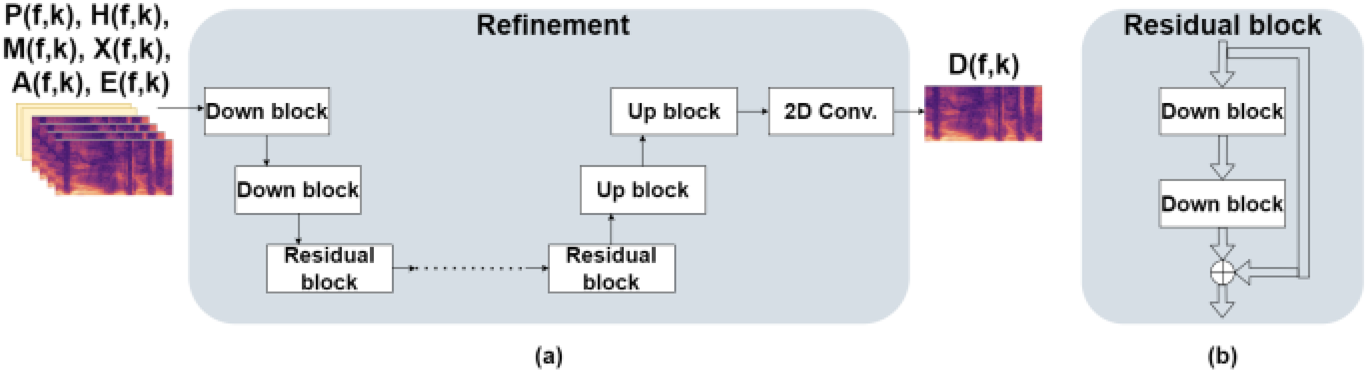


Figure 4.2: Structure of refinement model architecture and residual blocks. (**a**) Refinement model architecture. (**b**) Structure of the residual blocks.

2^{nd} Stage: Spectrogram Refinement

- This stage is focused on improving speech quality
- Loss function – PMSQE

  J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A Deep Learning Loss Function Based on the Perceptual Evaluation of the Speech Quality," IEEE Signal Process. Lett., vol. 25, no. 11, pp. 1680–1684, 2018.
- Approximates PESQ

$$l_{\mathrm{MSE}} = \frac{1}{n}\sum_{f}\sum_{k}(\log_{10}(\tilde{D}(f,k)+\epsilon) - \log_{10}(D(f,k)+\epsilon))^2$$

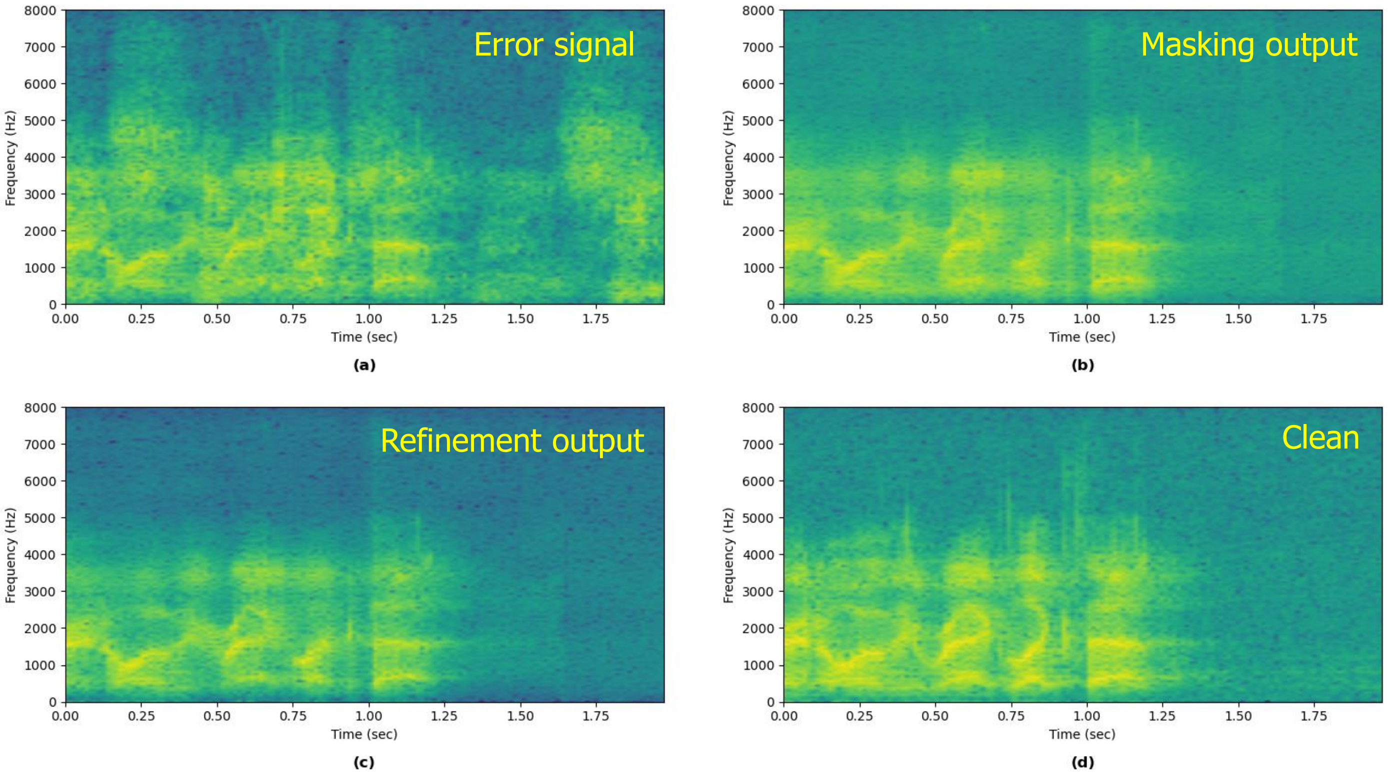
$$l = l_{\mathrm{PESQ}} + \lambda_{\mathrm{MSE}} l_{\mathrm{MSE}}$$

Results – Ablation Study

Table 4.3: Ablation study results. M stands for masking, D for DTD, and R for refinement.

	Far-end only		Double-talk	
	ERLE	AECMOS	PESQ	AECMOS
AEC	18.80	4.67	2.25	4.15
AEC+M	40.39	4.67	2.74	4.66
AEC+M+D	42.28	4.67	2.84	4.69
AEC+R	40.69	4.66	2.75	4.57
AEC+M+D+R	**44.32**	**4.68**	**2.94**	**4.71**

Results - Ablation Study



Results – DTD Comparisons

Table 4.4: Study of different configurations of the masking model with a DTD. Conf. stands for configuration.

	Far-end only		Double-talk	
	ERLE	**AECMOS**	**PESQ**	**AECMOS**
No DTD	40.39	**4.67**	2.74	4.66
Conf. 1	41.07	4.61	2.69	4.56
Conf. 2	39.88	4.66	2.75	4.60
Conf. 3	41.17	4.66	2.72	4.65
Proposed	**42.28**	**4.67**	**2.84**	**4.69**

Results – Comparative

Table 4.6: Comparison of the proposed, the Residual-U-Net (U-Net), and the Complex-Masking (Masking) systems. Param. stands for parameters and Mem. for memory.

	Far-end only		Double-talk		# Param.	Mem. (Bytes)	RTF
	ERLE	**AECMOS**	**PESQ**	**AECMOS**			
U-Net	39.39	4.62	2.56	4.04	0.14 M	0.5 M	0.03
Masking	**44.54**	4.67	2.73	4.55	1.86 M	7.0 M	0.32
Proposed	44.32	**4.68**	**2.94**	**4.71**	5.1 M	21.3 M	0.04

Results – Different SERs

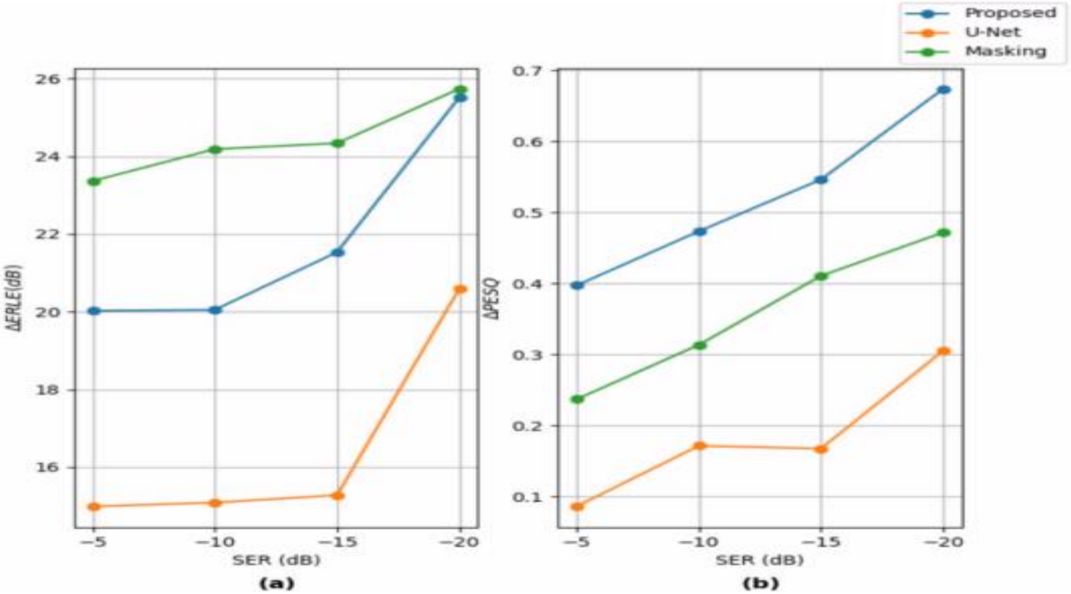


Figure 4.5: Systems' performance in different SERs. **(a)** Echo return loss enhancement (ERLE) difference between the systems' outputs and the error signal. **(b)** Perceptual evaluation of speech quality (PESQ) difference between the systems' outputs and the error signal.

TECHNION
Israel Institute of Technology

Results – Some Examples

Microphone	Error	Clean	Estimated

TECHNION
Israel Institute of Technology

Summary

- A two-stage deep-learning RES and DTD system focused on the low SER scenario
- Proposed DTD configuration outperforms competition from previous studies
- Novel spectrogram refinement stage
- Proposed systems outperforms competing systems
- Most efficient in the lower SERs