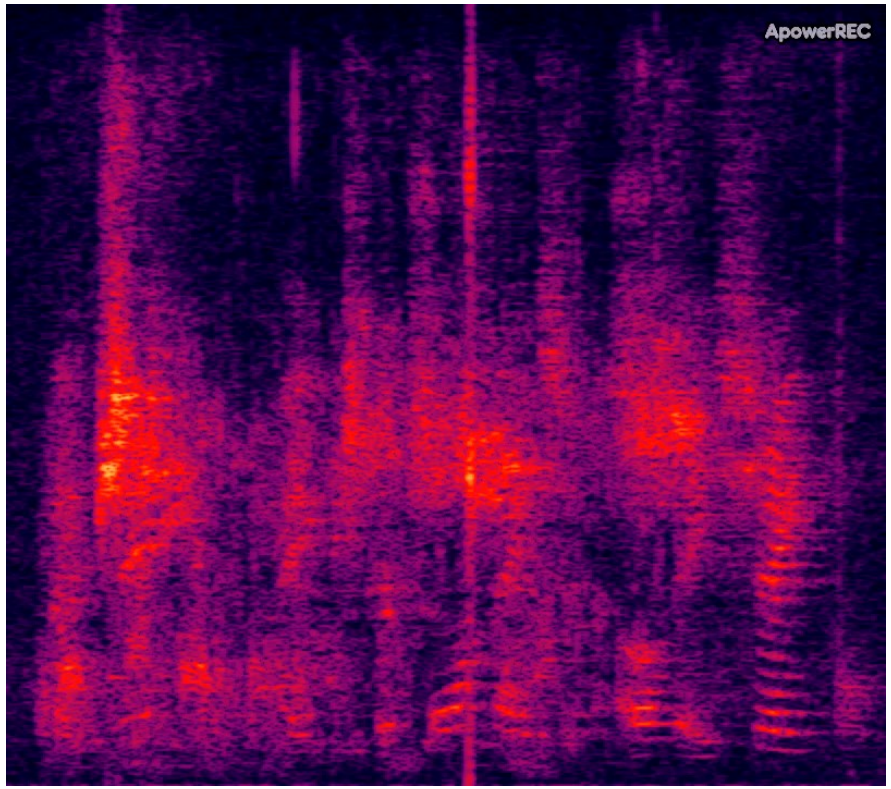
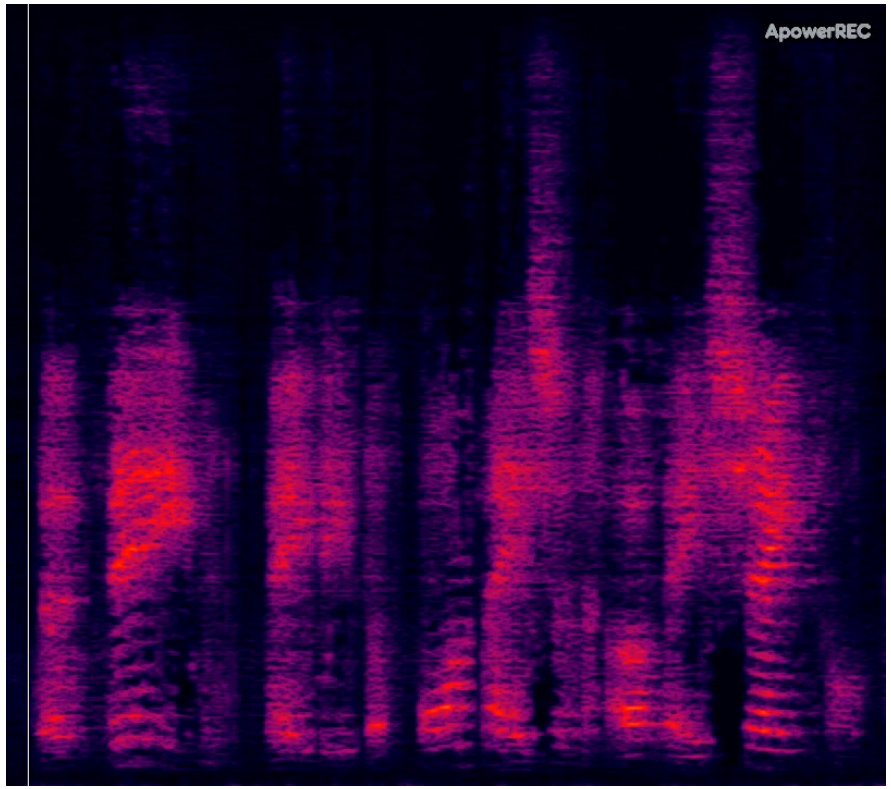
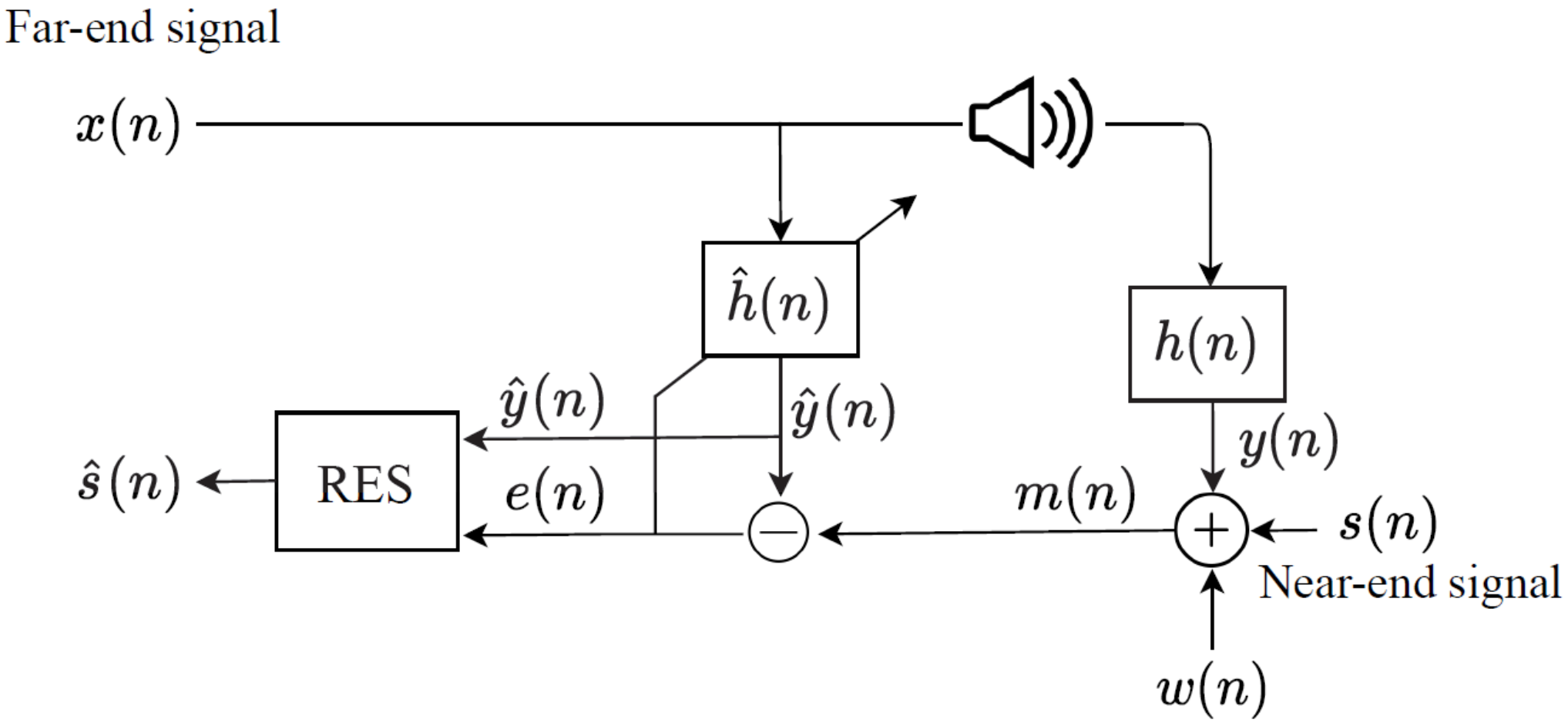
# Deep Residual Echo Suppression with a Tunable Tradeoff Between Signal Distortion and Echo Suppression

Amir Ivry, Prof. Israel Cohen, and Dr. Baruch Berdugo | IEEE ICASSP ‘21

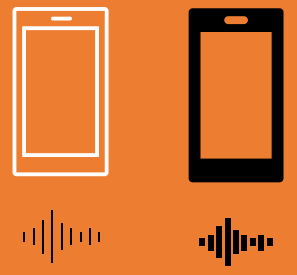


# Traditional System Setup



# The Challenge



Double-talk scenarios (full-duplex)

Traditional AECs reduce linear echo components with limited resources

!

RES should jointly achieve low signal distortion and high echo suppression

# Proposed Solution

UNet for error-to-target regression

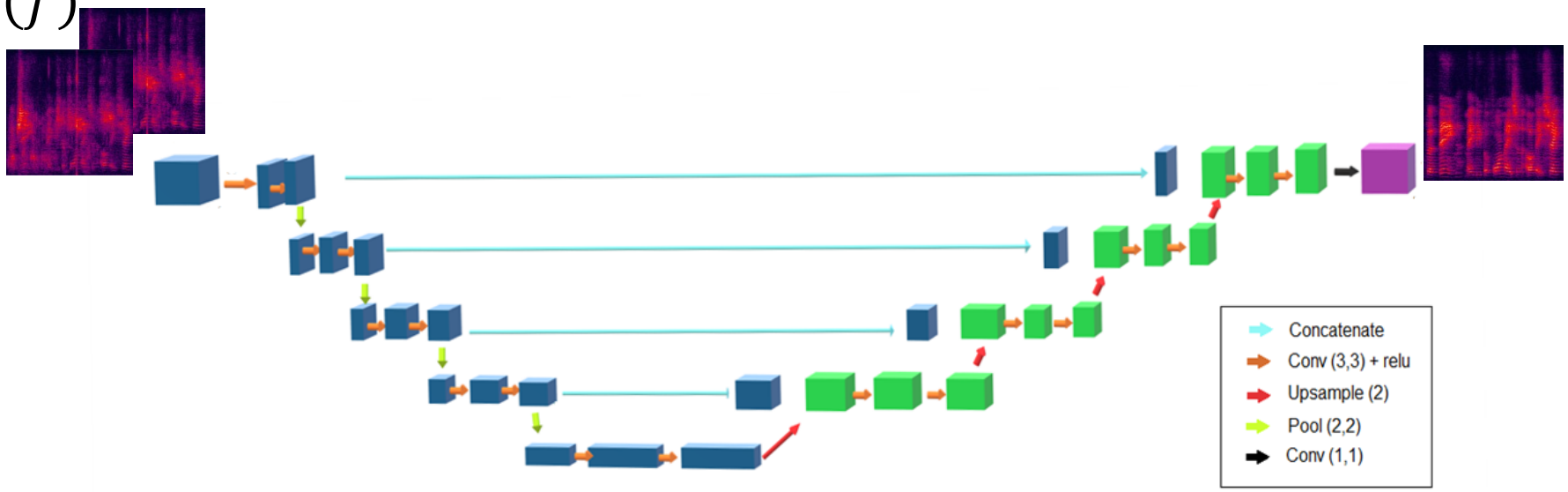
Over 160 hours of real and simulated data

Tunable design parameter for distortion-echo tradeoff

# RES System Training

(Adaptive) $A(f)$

(Error) $E(f)$




$S(f)$ (Target)

Objective to minimize: $\left\|S(f) - \hat{S}\,(f)\right\|^2 + \boldsymbol{\alpha} \cdot \left\|\hat{S}\,(f)\right\|^2 + 0.1 \cdot \sigma^2_{\hat{s}\,(f)}$

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

# AEC Challenge Database

10,000

synthetic scenarios

2,500+

real-life environments

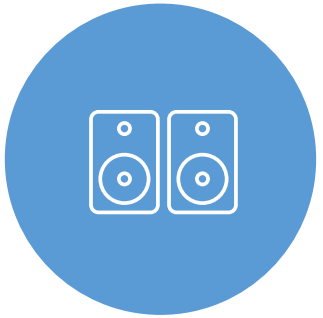
[-10, 10] dB SER

[0, 40] dB SNR

# Independent Recordings

Mouth simulator

(Bruel & Kjaer)



External speaker

(Logitech)

Speaker-phone

(Phoenix Audio)

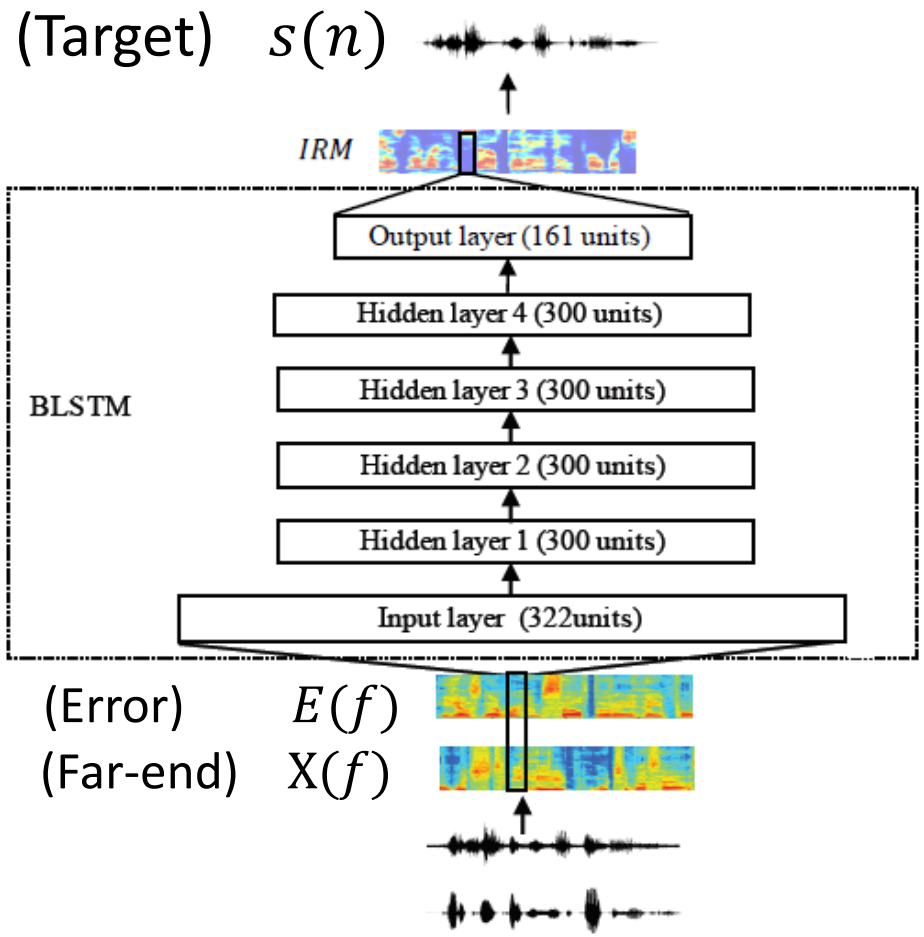
As low as -20 dB SER

# Performance Measures

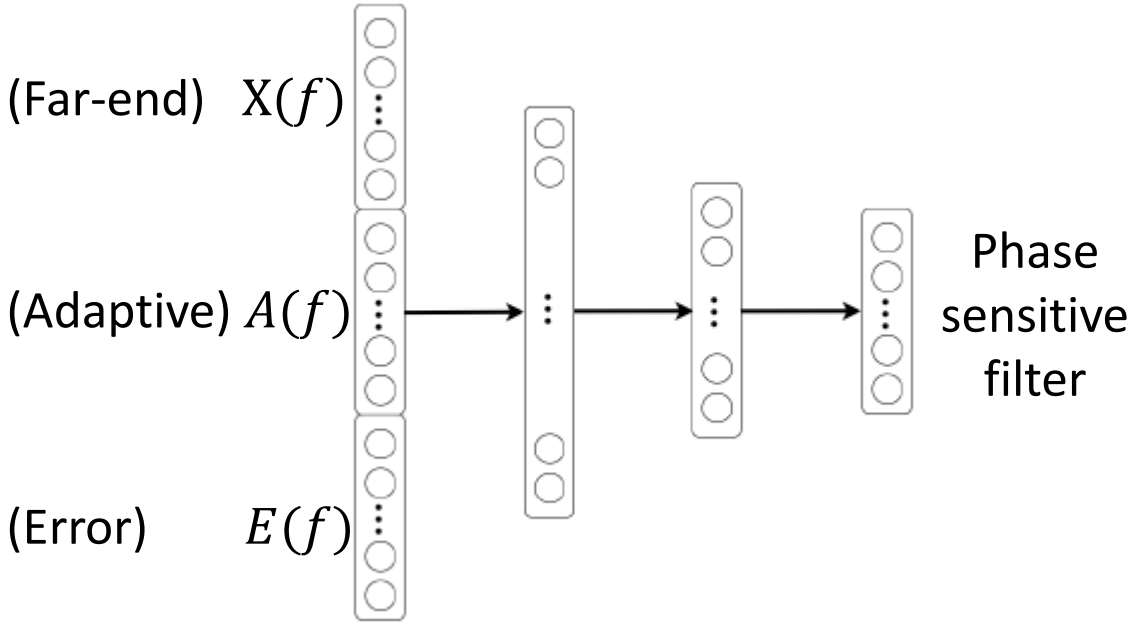
| Measure name | Abbr. | Definition | Scenario |
|---|---|---|---|
| Echo return loss enhancement | **ERLE** [dB] | $10\log_{10}\frac{\|\|error\|\|^2}{\|\|prediciton\|\|^2}$ | Single-talk<br>Far-end only |
| Signal-to-artifacts-ratio | **SAR** [dB] | $10\log_{10}\frac{\|\|error\|\|^2}{\|\|target-prediciton\|\|^2}$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

# Competing RES Methods

## Zhang, Interspeech, ‘18




## Carbajal, ICASSP, ‘18




$$\text{Phase sensitive filter} \triangleq \frac{S(f)}{E(f)} \cdot \cos(\theta_S - \theta_E)$$

# Performance
# Comparison to Competing Methods

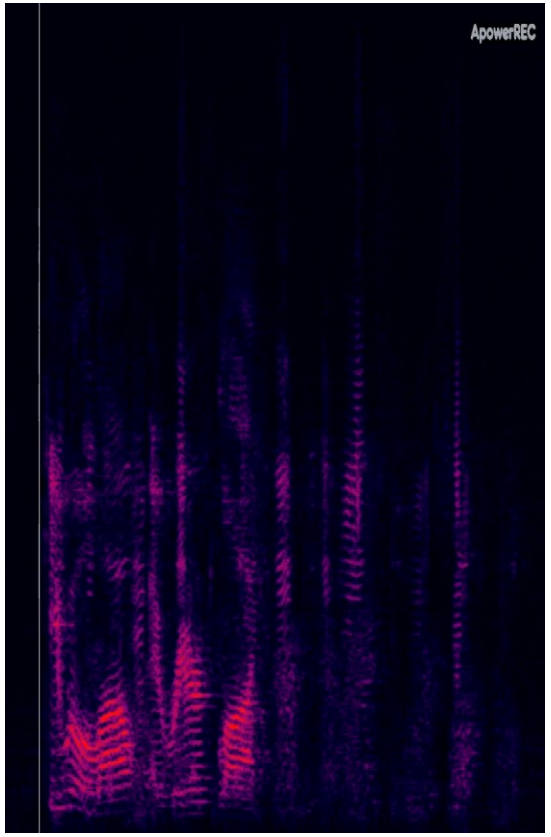
## No echo path change

| | 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 |

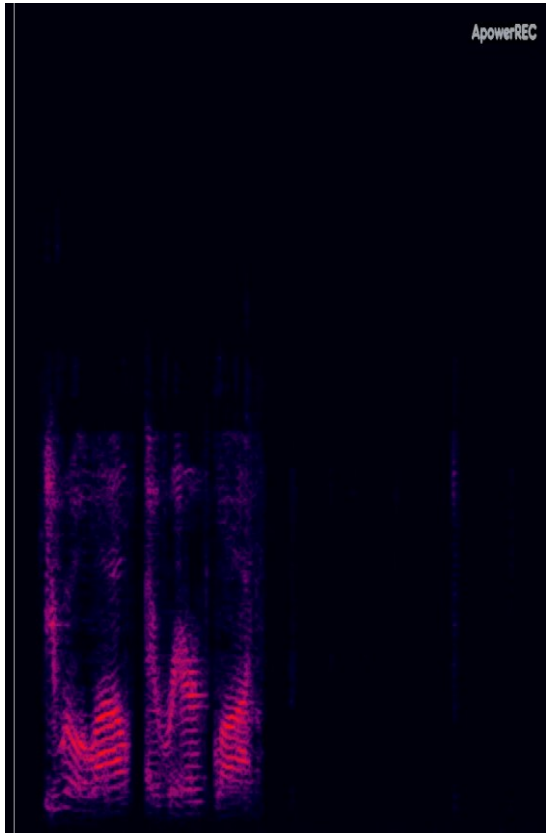
## Echo path change

| | 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 |

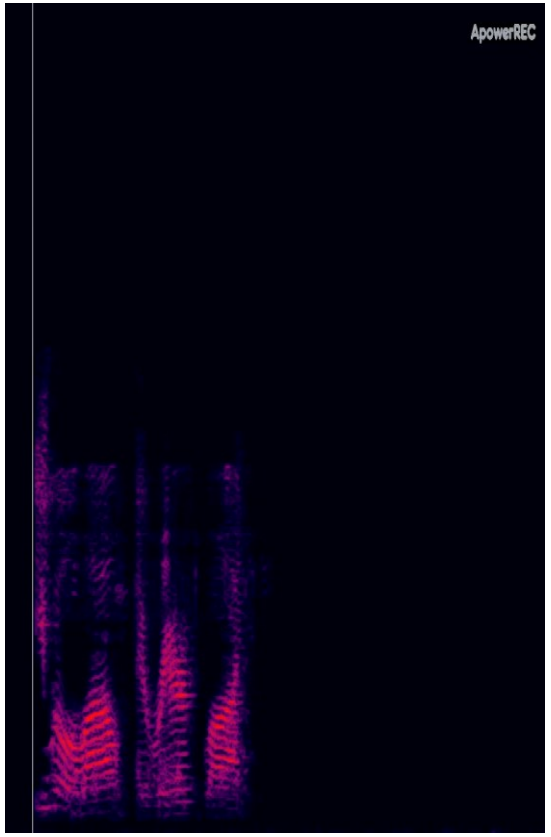
$E(f)$ – Error



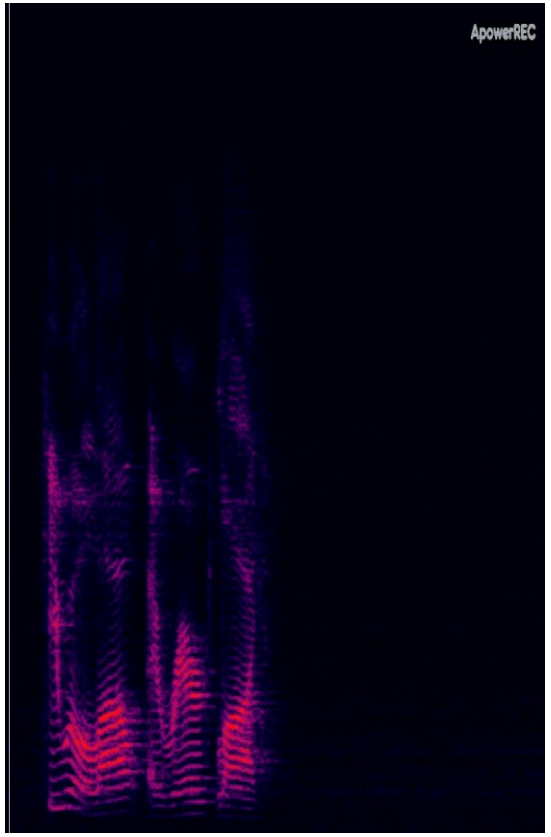

Zhang, Interspeech, 18’




Carbajal, ICASSP, 18’




Proposed, ICASSP, 21’




# Performance Comparison to Competing Methods

Audio Examples

# Design Parameter

| | $\alpha = 0$ | | $\alpha = 0.5$ | | $\alpha = 1$ | |
|---|---|---|---|---|---|---|
| | 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 |

# Resources Analysis

| Resource | Value | Notes |
|---|---|---|
| Parameters | **130 K** | Applicable for integration on-device |
| Memory | **4.16 MB** | |
| Floating point operations per second (flops) | **1.6 G-flops** | |
| System latency | **30 ms** | Complies with real-time communication standards |
| Network inference time | **8 ms** | |


Thank You