AECMOS: A Perceptual Objective Speech Quality Metric for Echo Impairment

Marju Purin **Sten Sootla** **Mateja Sponza** **Ando Saabas** **Ross Cutler**

Microsoft Corporation

April 13, 2022

Acoustic Echo

- Acoustic echo arises when a near end microphone picks up the near end loudspeaker signal and a far end user hears their own voice.
- The presence of acoustic echo is a top complaint in user ratings of audio call quality.

Goal: Replace humans as call quality raters

- Human rating process is time-consuming, costly, and not scalable
- Capture human subjective opinion with a neural network
- Rate call degradation in two categories: echo and other (noise, reverberation, artifacts)
- Monitor real calls

Model https://github.com/microsoft/AEC-Challenge

Available as an Azure service and a .onnx model:

- Input: 3 processed signals (STFT window 512, hop=0.5) and optional scenario marker
- Output: echo DMOS and other degradation DMOS
- Model handles variable length inputs natively

Architecture

Table: AECMOS architecture

Layer	Output Dimensions
Input: $3 \times 541 \times 257$	
Conv: 32, (3×3), LeakyReLU MaxPool: (2 x 2), Dropout(0.4)	$(32, 270, 128)$
Conv: 64, (3×3), LeakyReLU MaxPool: (2 x 2), Dropout(0.4)	$(64, 135, 64)$
Conv: 64, (3×3), LeakyReLU MaxPool: (2 x 2), Dropout(0.4)	$(64, 67, 32)$
Conv: 128, (3×3), LeakyReLU MaxPool: (2 x 2), Dropout(0.4)	$(128, 33, 16)$
Global MaxPool	$(1, 128)$
Bidirectional GRU: 128, NumLayers 2 HiddenUnits 64, Drouput(0.2)	$(1, 128)$
Dense: 64, LeakyReLU Dropout(0.4)	$(1, 64)$
Dense: 64, LeakyReLU Dropout(0.4)	$(1, 64)$
Dense: 2, $1 + 4 \cdot$sigmoid	$(1, 2)$

Test Set

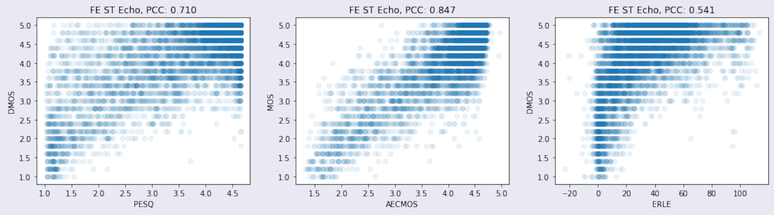
Get 17,600 enhanced signals

- 300 double talk, 300 far-end single talk, and 200 near-end single talk examples
- 14 Interspeech 2021 contest models and 8 in-house models

Per Clip (test set)

Table: Per Clip PCC for AECMOS, and other commonly used metrics: DNSMOS, ERLE, PESQ, EQUEST.

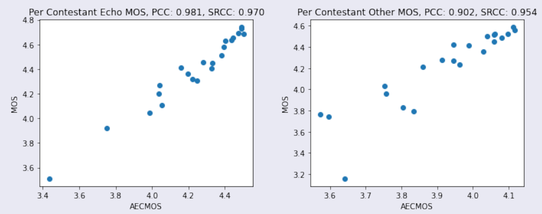
	AEC MOS	DNS MOS	ERLE	PESQ	EQUEST
ST far end DMOS	0.847		0.541	0.710	0.686
ST near-end MOS	0.611	0.640			
DT Echo DMOS	0.582				
DT Other DMOS	0.751				



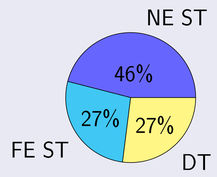
Per contestant (test set)

Table: AECMOS Per Contestant PCC and SRCC: All Scenarios refers to far end single talk and double talk for echo; and near end single talk MOS and double talk Other MOS.

	PCC	SRCC
All Scenarios Echo DMOS	0.981	0.970
All Scenarios (Other) MOS	0.902	0.954
ST far end Echo DMOS	0.996	0.969
ST near end MOS	0.923	0.831
DT Echo DMOS	0.898	0.863
DT Other DMOS	0.927	0.955



Training Data



- 64K samples
- A sample consists of three signals: microphone, loopback, and enhanced signal
- Ground truth labels:1-5, where 1=Very annoying ... 5=Imperceptible

Ablation Study

Table: Per Clip Pearson Correlation Coefficients: Baseline Convolutional Model; add scenario markers to model input; remove a convolution layer and add a GRU layer to obtain AECMOS.

	Baseline	+ scenario	+ GRU
All Scenarios Echo DMOS	0.732	0.746	0.797
All Scenarios (Other) MOS	0.735	0.775	0.802
ST far end Echo DMOS	0.780	0.825	0.847
ST near end MOS	0.434	0.534	0.611
DT Echo DMOS	0.458	0.422	0.582
DT Other DMOS	0.577	0.657	0.751

Ablation Study: Mel features

Table: Per Clip Pearson Correlation Coefficients: AECMOS; AECMOS trained with Mel spectrogram features.

	AECMOS	AECMOS Mel
All Scenarios Echo DMOS	0.797	0.742
All Scenarios (Other) MOS	0.802	0.819
ST far end Echo DMOS	0.847	0.739
ST near end MOS	0.611	0.604
DT Echo DMOS	0.582	0.553
DT Other DMOS	0.751	0.772