INTERSPEECH 2021 Acoustic Echo Cancellation Challenge

Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Sten Sootla, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sorensen, Robert Aichner, Sriram Srinivasan

Microsoft Corp., USA

firstname.lastname@microsoft.com

Abstract

The INTERSPEECH 2021 Acoustic Echo Cancellation Challenge is intended to stimulate research in the area of acoustic echo cancellation (AEC), which is an important part of speech enhancement and still a top issue in audio communication. Many recent AEC studies report good performance on synthetic datasets where the training and testing data may come from the same underlying distribution. However, AEC performance often degrades significantly on real recordings. Also, most of the conventional objective metrics such as echo return loss enhancement and perceptual evaluation of speech quality do not correlate well with subjective speech quality tests in the presence of background noise and reverberation found in realistic environments. In this challenge, we open source two large datasets to train AEC models under both single talk and double talk scenarios. These datasets consist of recordings from more than 5,000 real audio devices and human speakers in real environments, as well as a synthetic dataset. We also open source an online subjective test framework and provide an online objective metric service for researchers to quickly test their results. The winners of this challenge are selected based on the average Mean Opinion Score achieved across all different single talk and double talk scenarios.

Index Terms: acoustic echo cancellation, deep learning, single talk, double talk, subjective test

1. Introduction

With the growing popularity and need for working remotely, the use of teleconferencing systems such as Microsoft Teams, Skype, WebEx, Zoom, etc., has increased significantly. It is imperative to have good quality calls to make the users' experience pleasant and productive. The degradation of call quality due to acoustic echoes is one of the major sources of poor speech quality ratings in voice and video calls. While digital signal processing (DSP) based AEC models have been used to remove these echoes during calls, their performance can degrade when model assumptions are violated, e.g., fast time-varying acoustic conditions, unknown signal processing blocks or non-linearities in the processing chain, or failure of other models (e.g. background noise estimates). This problem becomes more challenging during full-duplex modes of communication where echoes from double talk scenarios are difficult to suppress without significant distortion or attenuation [1].

With the advent of deep learning techniques, several supervised learning algorithms for AEC have shown better performance compared to their classical counterparts [2, 3, 4]. Some studies have also shown good performance using a combination of classical and deep learning methods such as using adaptive filters and *recurrent neural networks* (RNNs) [4, 5] but only on synthetic datasets. While these approaches provide a good heuristic on the performance of AEC models, there has been no evidence of their performance on real-world datasets with speech recorded in diverse noise and reverberant environments. This makes it difficult for researchers in the industry to choose a good model that can perform well on a representative real-world dataset.

Table 1: *Pearson Correlation Coefficient (PCC) and Spearman's Rank Correlation Coefficient (SRCC) between ERLE, PESQ and P.808 Absolute Category Rating (ACR) results on single talk with delayed echo scenarios (see Section 5).*

	PCC	SRCC
ERLE	0.31	0.23
PESQ	0.67	0.57

Most AEC publications use objective measures such as *echo return loss enhancement* (ERLE) [6, 7] and *perceptual evaluation of speech quality* (PESQ) [8]. ERLE is defined as:

$$ERLE = 10 \log_{10} \frac{\mathbb{E}[y^2(n)]}{\mathbb{E}[\hat{y}^2(n)]} \tag{1}$$

where $y(n)$ is the microphone signal, and $\hat{y}(n)$ is the enhanced speech. ERLE is only appropriate when measured in a quiet room with no background noise and only for single talk scenarios (not double talk). PESQ has also been shown to not have a high correlation to subjective speech quality in the presence of background noise [9]. Using the datasets provided in this challenge we show the ERLE and PESQ have a low correlation to subjective tests (Table 1). In order to use a dataset with recordings in real environments, we can not use ERLE and PESQ. A more reliable and robust evaluation framework is needed that everyone in the research community can use, which we provide as part of the challenge.

This AEC challenge is designed to stimulate research in the AEC domain by open sourcing a large training dataset, test set, and subjective evaluation framework. We provide two new open source datasets for training AEC models. The first is a real dataset captured using a large-scale crowdsourcing effort. This dataset consists of real recordings that have been collected from over 5,000 diverse audio devices and environments. The second dataset is synthesized from speech recordings, room impulse responses, and background noise derived from [10]. An initial test set will be released for the researchers to use during development and a blind test set near the end, which will be used to decide the final competition winners. We believe these datasets are large enough to facilitate deep learning and representative enough for practical usage in shipping telecommunication products.

This is the second AEC challenge we have conducted. The first challenge was held at ICASSP 2021 [11] and included 17 participants with entries ranging from pure deep models, hybrid linear AEC + deep echo suppression, and DSP methods. The results show that the deep and hybrid models far outperformed DSP methods, with the winner being a hybrid model. However, there is still much room for improvement. To improve the challenge and further stimulate research in this area we have made the following changes:

- The dataset has increased from 2,500 devices and environments to 5,000 to provide additional training data.
- The test set has been significantly improved to include more real-world issues that challenge echo cancellers, such as clock drift, gain variations on the near end, more severe echo path changes, glitches in the mic/speaker signal, and more devices with poor onboard AEC's. This test set should be more challenging than the first challenge.
- The test framework has been improved to increase the accuracy of echo impairment ratings in the presence of background noise.
- The challenge includes a real-time and non-real-time track.
- Additional time is given to complete the challenge.
- A new Azure Service based objective metric is provided that has a high correlation to human ratings (see Table 2).

The training dataset is described in Section 2, and the test set in Section 3. We describe a DNN-based AEC method in Section 4. The online subjective evaluation framework is discussed in Section 5, and the objective service in Section 6. The results are given in Section 7. The challenge rules are described in https://aka.ms/aec-challenge.

2. Training datasets

The challenge will include two new open source datasets, one real and one synthetic. The datasets are available at https://github.com/microsoft/AEC-Challenge.

2.1. Real dataset

The first dataset was captured using a large-scale crowdsourcing effort. This dataset consists of more than 30,000 recordings from 5,000 different real environments, audio devices, and human speakers in the following scenarios:

1. Far end single talk, no echo path change
2. Far end single talk, echo path change
3. Near end single talk, no echo path change
4. Double talk, no echo path change
5. Double talk, echo path change
6. Sweep signal for RT60 estimation

For the far end single talk case, there is only the loudspeaker signal (far end) played back to the users and users remain silent (no near end signal). For the near end single talk case, there is no far end signal and users are prompted to speak, capturing the near end signal. For double talk, both the far end and near end signals are active, where a loudspeaker signal is played and users talk at the same time. Echo path change was incorporated by instructing the users to move their device around or bring themselves to move around the device. The near end single talk speech quality is given in Figure 1. The RT60 distribution for 2678 environments in the real dataset for which impulse response measurements were available is estimated using a method by Karjalainen et al. [12] and shown in Figure 2. The RT60 estimates can be used to sample the dataset for training.

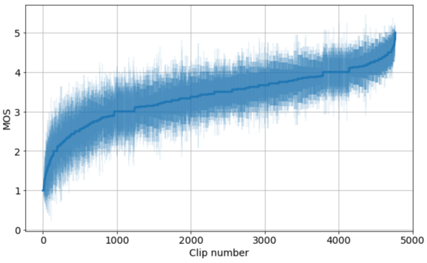


Figure 1: *Sorted near end single talk clip quality (P.808) with 95% confidence intervals.*

We use *Amazon Mechanical Turk* as the crowdsourcing platform and wrote a custom HIT application that includes a custom tool that raters download and execute to record the six scenarios described above. The dataset includes only Microsoft Windows devices. Each scenario includes the microphone and loopback signal (see Figure 3). Even though our application uses the WASAPI raw audio mode to bypass built-in audio effects, the PC can still include Audio DSP on the receive signal (e.g., equalization and Dynamic Range Compression (DRC)); it can also include Audio DSP on the send signal, such as AEC and noise suppression.

For clean speech far end signals, we use the speech segments from the Edinburgh dataset [13]. This corpus consists of short single speaker speech segments (1 to 3 seconds). We used a *long short term memory* (LSTM) based gender detector to select an equal number of male and female speaker segments. Further, we combined 3 to 5 of these short segments to create clips of length between 9 and 15 seconds in duration. Each clip consists of a single gender speaker. We create a gender-balanced far end signal source comprising of 500 male and 500 female clips. Recordings are saved at the maximum sampling rate supported by the device and in 32-bit floating point format; in the released dataset we down-sample to 16kHz and 16-bit using automatic gain control to minimize clipping.

For noisy speech far end signals we use 2000 clips from the near end single talk scenario that were rated between MOS 3 and 4 using ITU-T P.808 subjective testing framework. Clips are gender balanced to include an equal number of male and female voices.

For near end speech, the users were prompted to read sentences from TIMIT [14] sentence list. Approximately 10 seconds of audio is recorded while the users are reading.

2.2. Synthetic dataset

The second dataset provides 10,000 synthetic scenarios, each including single talk, double talk, near end noise, far end noise, and various nonlinear distortion scenarios. Each scenario includes a far end speech, echo signal, near end speech, and near end microphone signal clip. We use 12,000 cases (100 hours of audio) from both the clean and noisy speech datasets derived in [10] from the LibriVox project[1] as source clips to sample

[1] https://librivox.org

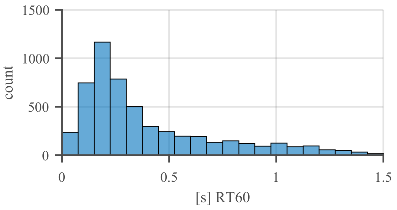


Figure 2: *Distribution of reverberation time (RT60).*

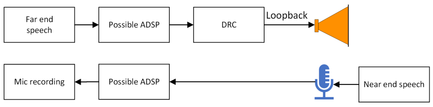


Figure 3: *The custom recording application recorded the loopback and microphone signals.*

far end and near end signals. The LibriVox project is a collection of public domain audiobooks read by volunteers. [10] used the online subjective test framework ITU-T P.808 to select audio recordings of good quality ($4.3 \leq$ MOS ≤ 5) from the LibriVox project. The noisy speech dataset was created by mixing clean speech with noise clips sampled from Audioset [15], Freesound[2] and DEMAND [16] databases at signal to noise ratios sampled uniformly from [0, 40] dB.

To simulate a far end signal, we pick a random speaker from a pool of 1,627 speakers, randomly choose one of the clips from the speaker, and sample 10 seconds of audio from the clip. For the near end signal, we randomly choose another speaker and take 3-7 seconds of audio which is then zero-padded to 10 seconds. Of the selected far end and near end speakers, 71% are female and 67% are male. To generate an echo, we convolve a randomly chosen room impulse response from a large internal database with the far end signal. The room impulse responses are generated by using Project Acoustics technology[3] and the RT60 ranges from 200 ms to 1200 ms. In 80% of the cases, the far end signal is processed by a nonlinear function to mimic loudspeaker distortion. For example, the transformation can be clipping the maximum amplitude, using a sigmoidal function as in [17], or applying learned distortion functions, the details of which we will describe in a future paper. This signal gets mixed with the near end signal at a signal to echo ratio uniformly sampled from -10 dB to 10 dB. The signal to echo ratio is calculated based on the clean speech signal (*i.e.* a signal without near end noise). The far end and near end signals are taken from the noisy dataset in 50% of the cases. The first 500 clips can be used for validation as these have a separate list of speakers and room impulse responses. Detailed metadata information can be found in the repository.

3. Test set

Two test sets are included, one at the beginning of the challenge and a blind test set near the end. Both consist of 800 real world recordings, between 30-45 seconds in duration. The datasets include the following scenarios that make echo cancellation more challenging:

- Long- or varying delays, i.e., files where the delay between loopback and mic-in is atypically long or varies during the recording.
- Strong speaker and/or mic distortions.
- Stationary near-end noise.
- Non-stationary near-end noise.
- Recordings with audio DSP processing from the device, such as AEC.
- Glitches, i.e., files with ”choppy” audio, for example, due to very high CPU usage.
- Gain variations, i.e., recordings where far-end level changes during the recording (2.1), sampled randomly.

4. Baseline AEC Method

We adapt a noise suppression model developed in [18] to the task of echo cancellation. Specifically, a recurrent neural network with gated recurrent units takes concatenated log power spectral features of the microphone signal and far end signal as input, and outputs a spectral suppression mask. The short-time Fourier transform is computed based on 20 ms frames with a hop size of 10 ms, and a 320-point discrete Fourier transform. We use a stack of two gated recurrent unit layers, each of size 322 nodes, followed by a fully-connected layer with a sigmoid activation function. The model has 1.3 million parameters. The estimated mask is point-wise multiplied with the magnitude spectrogram of the microphone signal to suppress the far end signal. Finally, to resynthesize the enhanced signal, an inverse short-time Fourier transform is used on the phase of the microphone signal and the estimated magnitude spectrogram. We use a mean squared error loss between the clean and enhanced magnitude spectrograms. The Adam optimizer with a learning rate of 0.0003 is used to train the model. The model and the inference code is available in the challenge repository. [4]

5. Online subjective evaluation framework

We have extended the open source P.808 Toolkit [19] with methods for evaluating the echo impairments in subjective tests. We followed the *Third-party Listening Test B* from ITU-T Rec. P.831 [20] and ITU-T Rec. P.832 [21] and adapted them to our use case as well as for the crowdsourcing approach based on the ITU-T Rec. P.808 [22] guidance.

A third-party listening test differs from the typical listening-only tests (according to the ITU-T Rec. P.800) in the way that listeners hear the recordings from the *center* of the connection rather in the former one in which the listener is positioned at one end of the connection [20]. Thus, the speech material should be recorded by having this concept in mind. During the test session, we use different combinations of single- and multi-scale ACR ratings depending on the speech sample under evaluation. We distinguish between single talk and double talk scenarios. For the near end single talk, we ask for the overall quality. For the far end single talk and double talk scenario, we ask for an echo annoyance and for impairments of other degradations in two separate questions[5]. Both impairments are rated on the degradation category scale (from 1:*Very annoying*, to 5: *Imperceptible*). The impairments scales leads to a Degradation Mean Opinion Scores (DMOS). Note that we do not use the

[2] https://freesound.org

[3] https://www.aka.ms/acoustics

[4] https://github.com/microsoft/AEC-Challenge/tree/main/baseline/interspeech2021

[5] Question 1: How would you judge the degradation from the echo? Question 2: How would you judge other degradations (noise, missing audio, distortions, cut-outs)?

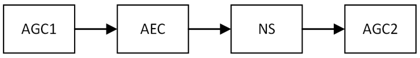


Figure 4: *The audio processing pipeline used in the challenge.*

Table 2: *AECMOS Pearson rank correlation coefficient (PCC) for our current model and the old model.*

Scenario	PCC	Old PCC
Far end single talk echo DMOS	0.97	0.99
Near end single talk other DMOS	0.92	0.80
Double talk echo DMOS	0.93	0.94
Double talk other DMOS	0.91	0.52

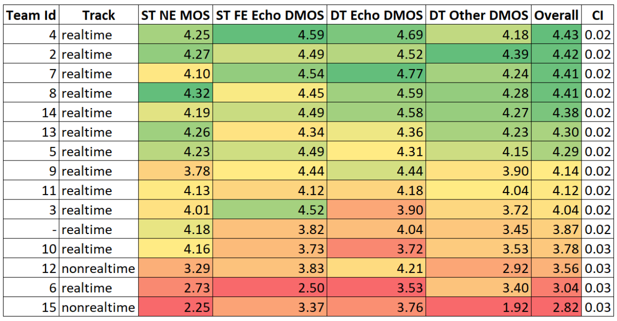
Other degradation category for far end single talk for evaluating echo cancellation performance, since this metric mostly reflects the quality of the original far end signal. However, we have found that having this component in the questionnaire helps increase the accuracy of echo degradation ratings (when measured against expert raters). Without the Other category, raters can sometimes assign degradations due to noise to the Echo category.

In the current challenge, for the far end single talk scenario, we evaluate the second half of each clip, to avoid initial degradations from initialization, convergence periods, and initial delay estimation. For the double talk scenario, we evaluate roughly the final third of the audio clip.

The audio pipeline used in the challenge is shown in Figure 4. In the first stage (AGC1) a traditional automatic gain control is used to target a speech level of -24 dBFS. The output of AGC1 is saved in the test set. The next stage is an AEC, which participants will process and upload to the challenge submission site. The next stage is a traditional noise suppressor (DMOS $<$ 0.1 improvement) to reduce stationary noise. Finally, a second AGC is run to ensure the speech level is still -24 dBFS. The subjective test framework with an AEC extension is available at https://github.com/microsoft/P.808. A more detailed description of the test framework and its validation is given in [23].

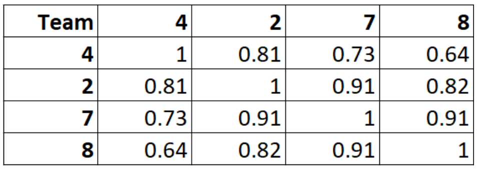
6. Azure service objective metric

We have developed an objective perceptual speech quality metric called AECMOS. It can be used to stack rank different AEC methods based on Mean Opinion Score (MOS) estimates with high accuracy. It is a neural network-based model that is trained using the ground truth human ratings obtained using our online subjective evaluation framework. The audio data used to train the AECMOS model is gathered from the numerous subjective tests that we conducted in the process of improving the quality of our AECs as well as the first AEC challenge results. Our model has improved greatly since the start of the contest. The performance of AECMOS on stack ranking models is given in Table 2 compared with subjective human ratings on the 14 submitted models from the Second AEC Challenge. We are still working on making the model generalize better on the new challenge test set using methods described in [24]. Sample code and details of the evaluation API can be found on https://aka.ms/aec-challenge.



Team Id	Track	ST NE MOS	ST FE Echo DMOS	DT Echo DMOS	DT Other DMOS	Overall	CI
4	realtime	4.25	4.59	4.69	4.18	4.43	0.02
2	realtime	4.27	4.49	4.52	4.39	4.42	0.02
7	realtime	4.10	4.54	4.77	4.24	4.41	0.02
8	realtime	4.32	4.45	4.59	4.28	4.41	0.02
14	realtime	4.19	4.49	4.58	4.27	4.38	0.02
13	realtime	4.26	4.34	4.36	4.23	4.30	0.02
5	realtime	4.23	4.49	4.31	4.15	4.29	0.02
9	realtime	3.78	4.44	4.44	3.90	4.14	0.02
11	realtime	4.13	4.12	4.18	4.04	4.12	0.02
3	realtime	4.01	4.52	3.90	3.72	4.04	0.02
-	realtime	4.18	3.82	4.04	3.45	3.87	0.02
10	realtime	4.16	3.73	3.72	3.53	3.78	0.03
12	nonrealtime	3.29	3.83	4.21	2.92	3.56	0.03
6	realtime	2.73	2.50	3.53	3.40	3.04	0.03
15	nonrealtime	2.25	3.37	3.76	1.92	2.82	0.03

Figure 5: *AEC challenge results*



Team	4	2	7	8
4	1	0.81	0.73	0.64
2	0.81	1	0.91	0.82
7	0.73	0.91	1	0.91
8	0.64	0.82	0.91	1

Figure 6: *ANOVA results for the top 4 participants*

7. Results

We received 14 submissions for the challenge. Each team submitted processed files from the blind test set (see Section 3). We batched all submissions into three sets:

- Near end single talk files for MOS test (NE ST MOS).
- Far end single talk files for Echo and Other degradation DMOS test (FE ST Echo/Other DMOS).
- Double talk files for Echo and Other degradation DMOS test (DT Echo/Other DMOS).

To obtain the final overall rating, we averaged the results of NE ST MOS, FE ST Echo DMOS, and DT Echo/Other DMOS, weighting them equally. The final standings are shown in Figure 5. The resulting scores show a wide variety in model performance. The score differences in near end, echo, and double talk scenarios for individual models highlight the importance of evaluating all scenarios, since in many cases, performance in one scenario comes at a cost in another scenario.

For the top four teams, we ran an ANOVA test to determine statistical significance (Figure 6). The differences between the teams were not statistically significant, and per the challenge rules, the winners will be picked based on the computational complexity of the models.

8. Conclusions

The results of this challenge show considerable improvement over the previous challenge [11], even though this test set was significantly more challenging. Nearly all participants exceeded the baseline model, and this year's top performer exceeded the baseline by 0.13 DMOS more than the previous challenge.

9. References

[1] "IEEE 1329-2010 Standard method for measuring transmission performance of handsfree telephone sets," 2010.

[2] A. Fazel, M. El-Khamy, and J. Lee, "CAD-AEC: Context-aware deep acoustic echo cancellation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6919–6923.

[3] M. M. Halimeh and W. Kellermann, "Efficient multichannel nonlinear acoustic echo cancellation based on a cooperative strategy," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 461–465.

[4] Lu Ma, Hua Huang, Pei Zhao, and Tengrong Su, "Acoustic echo cancellation by combining adaptive digital filter and recurrent neural network," *arXiv preprint arXiv:2005.09237*, 2020.

[5] Hao Zhang, Ke Tan, and DeLiang Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions.," in *INTERSPEECH*, 2019, pp. 4255–4259.

[6] "ITU-T recommendation G.168: Digital network echo cancellers," Feb 2012.

[7] Gerald Enzner, Herbert Buchner, Alexis Favrot, and Fabian Kuech, "Acoustic echo control," in *Academic press library in signal processing*, vol. 4, pp. 807–877. Elsevier, 2014.

[8] "ITU-T recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb 2001.

[9] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 631–635.

[10] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[11] Kusha Sridhar, Ross Cutler, Ando Saabas, Tanel Parnamaa, Hannes Gamper, Sebastian Braun, Robert Aichner, and Sriram Srinivasan, "ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP*, 2021.

[12] Matti Karjalainen, Poju Antsalo, Aki Mäkivirta, Timo Peltonen, and Vesa Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc*, vol. 50, no. 11, pp. 867, 2002.

[13] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks.," in *Interspeech*, 2016, pp. 352–356.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[15] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[16] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[17] Chul Min Lee, Jong Won Shin, and Nam Soo Kim, "DNN-based residual echo suppression," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] Yangyang Xia, Sebastian Braun, Chandan KA Reddy, Harishchandra Dubey, Ross Cutler, and Ivan Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 871–875.

[19] Babak Naderi and Ross Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," in *INTERSPEECH*, 2020.

[20] "ITU-T P.831 Subjective performance evaluation of network echo cancellers ITU-T P-series recommendations," 1998.

[21] ITU-T Recommendation P.832, *Subjective performance evaluation of hands-free terminals*, International Telecommunication Union, Geneva, 2000.

[22] "ITU-T P.808 supplement 23 ITU-T coded-speech database supplement 23 to ITU-T P-series recommendations (previously ccitt recommendations)," 1998.

[23] Ross Cutler, Babak Nadari, Markus Loide, Sten Sootla, and Ando Saabas, "Crowdsourcing approach for subjective evaluation of echo impairment," in *ICASSP*, 2021.

[24] Chandan K A Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2021.