THE SECOND DEEP NOISE SUPPRESSION CHALLENGE: IEEE ICASSP 2021

Chandan K A Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, Sriram Srinivasan

Microsoft Corporation, Redmond, USA

ABSTRACT

The Deep Noise Suppression (DNS) challenge is designed to foster innovation in the area of noise suppression to achieve superior perceptual speech quality. We recently organized a DNS challenge special session at INTERSPEECH 2020. We open sourced training and test datasets for researchers to train their noise suppression models. We also open sourced a subjective evaluation framework and used the tool to evaluate and pick the final winners. Many researchers from academia and industry made significant contributions to push the field forward. We also learned that as a research community, we still have a long way to go in achieving superior speech quality in challenging noisy real-time conditions. In this challenge, we are expanding both our training and test datasets. We will have two tracks with one track focusing on real-time denoising and the other track focusing on real-time personalised deep noise suppression. We also open source a non-intrusive objective speech quality metric called DNSMOS for participants to use during their development stages. The final evaluation will be based on subjective tests.

***Index Terms*—** Speech, Perceptual Speech Quality, Objective Metric, Deep Noise Suppressor, Metric.

1. INTRODUCTION

In the recent times, remote work has become the "new normal" as the number of people working remotely has exponentially increased due to pandemic. There has been a surge in the demand for reliable collaboration and real-time communication tools. Audio calls with superior speech quality is a need during these times as we try to stay connected and collaborate with people over stretch everyday. We are easily exposed to a variety of background noises such as dog barking, baby crying, air conditioner, traffic, kitchen noise, etc. Background noise significantly degrades the quality and intelligibility of the perceived speech leading to fatigue. Background noise poses a challenge in other applications such as hearing aids, smart devices etc.

Real-Time Speech Enhancement (SE) for perceptual quality is decades old classical problem for which researchers have proposed numerous solutions [1, 2, 3]. In the recent years, learning based approaches have shown promising results[4, 5, 6]. The Deep Noise Suppression (DNS) Challenge organized at Interspeech 2020 showed promising results, while also indicating that we are still about 1.4 Differential Mean Opinion Score (DMOS) behind the ideal Mean Opinion Score (MOS) of 5 when tested on the DNS Challenge test set [7, 8]. The DNS Challenge is the first contest that we are aware of that used subjective evaluation to benchmark SE methods using a realistic noisy test set [9]. We open sourced clean speech and noise corpus with configurable scripts to generate noisy-clean speech pairs suitable to train a supervised noise suppression model. There were two tracks, real-time and non-real-time based on the computational complexity of the inference. We received an overwhelming response to the challenge with participation from a diverse group of researchers, developers, students, and hobbyists from both academia and industry. We also received positive response from the participants as many found the open sourced data sets quite useful.

The Deep Noise Suppression (DNS) Challenge at ICASSP 2021 is intended to stimulate research in the area of real-time noise suppression. For ease of reference, we will call the ICASSP 2021 challenge as DNS Challenge 2 and the Interspeech 2020 challenge as DNS Challenge 1. The DNS challenge 2 will have a real-time denoising track similar to the one in DNS Challenge 1. In addition, we will have a personalized DNS track focused on using speaker information to achieve better perceptual quality. In addition to the data sets we open sourced for DNS Challenge 1, we will be adding over 20 hrs of clean speech with singing and provide more information about the characteristics of the noise. We will also provide about 100000 room impulse responses (RIR) curated from other data sets. For DNS Challenge 1, we open sourced a subjective evaluation framework based on ITU-T P.808. However, not many participants used the tool during their development phase as the subjective evaluation takes time and costs money. Most of the commonly used objective metrics like PESQ, POLQA and SDR require a reference clean speech to compute the scores. Hence, we cannot use them on real recordings. Also, these metrics are shown to correlate poorly with subjective scores. We are providing a non-intrusive objective metric called Deep Noise Suppression MOS (DNSMOS) predictor as an Azure service. DNSMOS is shown to correlate very well with subjective ratings and helps to stack rank the developed noise suppressors. The final evaluation of the participating models will be done based on subjective evaluation using P.808 subjective testing framework.

2. CHALLENGE TRACKS

The challenge will have the following two tracks:

1. Track 1: Real-Time Denoising track
   - The noise suppressor must take less than the stride time T_s (in ms) to process a frame of size T (in ms) on an Intel Core i5 quad-core machine clocked at 2.4 GHz or equivalent processors. For example, $T_s = T/2$ for 50% overlap between frames. The total algorithmic latency allowed including the frame size T, stride time T_s and any look ahead must be less than or equal to 40ms. For example, if you use a frame length of 20ms with a stride of 10ms resulting in an algorithmic delay of 30ms, then you satisfy the latency requirements. If you use a frame size of 32ms with a stride of 16ms resulting in an algorithmic delay of 48ms, then your method does not satisfy the latency requirements as the total algorithmic latency exceeds 40ms. If your frame size plus stride $T_1 = T + T_s$ is less than 40ms, then you can use up to $(40 - T_1)$ms future information.
2. Track 2: Personalized Deep Noise Suppression (pDNS) track
   - Satisfy Track 1 requirements.
   - You will have access to 2 minutes speech of a particular speaker to extract speaker related information that might be useful to improve the quality of the noise suppressor. The enhancement must be done on the noisy speech test segment of the same speaker.
   - The enhanced speech using speaker information must be of better quality than enhanced speech without using the speaker information.

3. TRAINING DATASETS

The goal of releasing the clean speech and noise datasets is to provide researchers with an extensive and representative dataset to train their SE models. We initially released MSSNSD with a focus on extensibility, but the dataset lacked the diversity in speakers and noise types. We published a significantly larger and more diverse data set with configurable scripts for DNS Challenge 1 [9]. Many researchers found this dataset useful to train their noise suppression models and achieved good results. However, the training and the test datasets did not include clean speech with emotions and singing. Also, the dataset is biased towards English language. For DNS Challenge 2, we are adding speech clips with other emotions and including about 10 non-English languages. The details about the clean and noisy dataset are described in the following sections.

3.1. Clean Speech

Clean speech consists of three subsets: (i) Neutral clean speech; (ii) Singing clean speech; (iii) Emotional clean speech; and (iv) Non-english clean speech. First subset is derived from the public audiobooks dataset called Librivox[1]. It is available under the permissive creative commons 4.0 license [10]. It has recordings of volunteers reading over 10,000 public domain audiobooks in various languages, majority of which are in English. In total, there are 11,350 speakers. Many of these recordings are of excellent speech quality, meaning that the speech was recorded using good quality microphones in a silent and less reverberant environments. But there are many recordings that are of poor speech quality as well with speech distortion, background noise, and reverberation. Hence, it is important to clean the data set based on speech quality.

We used the online subjective test framework ITU-T P.808 [11] to sort the book chapters by subjective quality. The audio chapters in Librivox are of variable length ranging from few seconds to several minutes. We randomly sampled 10 audio segments from each book chapter, each of 10 seconds in duration. For each clip, we had 2 ratings, and the MOS across all clips was used as the book chapter MOS. Figure 1 shows the results, which show the quality spanned from very poor to excellent quality.

Second subset consists of high-quality audio recordings of singing voice recorded in noise-free conditions by professional singers. This subset is derived from *VocalSet* corpus [12] with Creative Commons Attribution 4.0 International License (CC BY 4.0). license. It has 10.1 hours of clean singing voice recorded by 20 professional singers: 9 Male, and 11 female. This data was recorded on a range of vowels, diverse set of voices on several standard and extended vocal techniques, and sung in contexts of scales, arpeggios, long tones, and excerpts. We downsampled the mono .WAV files from 44.1kHz to 16kHz and added it to clean speech used by training data synthesizer.

Third subset consists of emotion speech recorded in noise-free conditions. This is derived from Crowd-sourced Emotional Mutimodal Actors Dataset (CREMA-D) [13] made available under the Open Database License. It consists of 7,442 audio clips from 91 actors: 48 male, and 43 female. Age of actors were in range 20 to 74 years with diverse

[1] https://librivox.org/

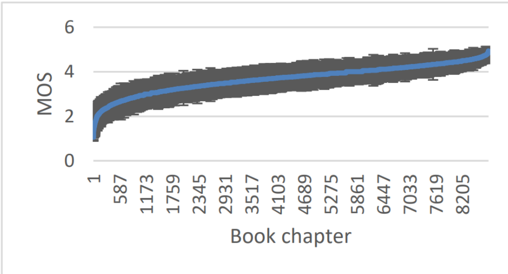

Fig. 1: Sorted near end single-talk clip quality (P.808) with 95% confidence intervals.

ethnic backgrounds including African America, Asian, Caucasian, Hispanic, and Unspecified. Actors read from a pool of 12 sentences for generating this emotional speech dataset. It accounts for six emotions: Anger, Disgust, Fear, Happy, Neutral, and Sad at four intensity levels: Low, Medium, High, Unspecified. The recorded audio clips were annotated by multiple human raters in three modalities: audio, visual, and audio-visual. Categorical emotion labels and real-value emotion level values of perceived emotion were collected using crowd-sourcing from 2,443 raters. This data was provided as 16 kHz .WAV files so we added it to our clean speech as it is.

Fourth subset has clean speech from non-English languages. It consists of both tonal and non-tonal languages including Chinese (Mandarin), German and Spanish. Mandarin data consists of OpenSLR18 [2] THCHS-30 [14] and OpenSLR33 [3] AISHELL [15] datasets, both with Apache 2.0 license. THCHS30 was published by Center for Speech and Language Technology (CSLT) at Tsinghua University for speech recognition. It consists of 30+ hours of clean speech recorded at 16-bit 16 kHz in noise-free conditions. Native speakers of standard Mandarin read text prompts chosen from a list of 1000 sentences. We added entire THCHS-30 data in our clean speech for training set. It consisted of 40 speakers: 9 male, 31 female in age range 19-55 years. It has total 13,389 clean speech audio files [14]. AISHELL dataset was created by Beijing Shell Shell Technology Co. Ltd. It has clean speech recorded by 400 native speakers (47% male and 53% female) of Mandarin with different accents. Audio was recorded in noise-free conditions using high fidelity microphones. It is provided as 16-bit 16kHz .wav files. It is one of the largest open-source Mandarin speech dataset. We added the entiore AISHELL corpus with 141,600 utterances spanning 170+ hours of clean Mandarin speech to our training set.

3.2. Noise

The noise clips were selected from Audioset [4] [16] and Freesound [5]. Audioset is a collection of about 2 million human labeled 10s sound clips drawn from YouTube videos and belong to about 600 audio events. Like the Librivox data, certain audio event classes are over-represented. For example, there are over a million clips with audio classes music and speech and less than 200 clips for classes such as toothbrush, creak, etc. Approximately 42% of the clips have a single class, but the rest may have 2 to 15 labels. Hence, we developed a sampling approach to balance the dataset in such a way that each class has at least 500 clips. We also used a speech activity detector to remove the clips with any kind of speech activity. The reason is to avoid suppression of speech by the noise suppression model trained to suppress speech-like noise. The resulting dataset has about 150 audio classes and 60,000 clips. We also augmented an additional 10,000 noise clips downloaded from Freesound and DEMAND databases [17]. The chosen noise types are more relevant to VOIP applications.

3.3. Room Impulse Responses

We provide 3076 real and approximately 115,000 synthetic rooms impulse responses (RIRs) where we can choose either one or both types of RIRs for convolving with clean speech. Noise is then added to reverbed clean speech while DNS models are expected to take noisy reverbed speech and produce clean reverbed speech. Challenge participants can do both de-reverb and denosing with their models if they prefer. These RIRs are chosen from openSLR26 [18] [6] and openSLR28 [18] [7] datasets, both released with Apache 2.0 License.

3.4. Acoustic Parameters

We provide two acoustic parameters: (i) Reveberation time, T60 and (ii) Clarity, C50 for all audio clips in clean speech of training set. We provide T60, C50 and isReal Boolean flag for all RIRs where isReal is 1 for real RIRs and 0 for synthetic ones. The two parameters are correlated. A RIR with low C50 can be described as highly reverberant and vice versa [19, 20]. These parameters are supposed to provide flexibility to researchers for choose a sub-set of provided data for controlled studies.

4. TEST SET

In DNS Challenge 1, the test set consisted of 300 real recordings and 300 synthesized noisy speech clips. The real clips

[2] http://www.openslr.org/18/
[3] http://www.openslr.org/33/
[4] https://research.google.com/audioset/
[5] https://freesound.org/
[6] http://www.openslr.org/26/
[7] http://www.openslr.org/28/

were recorded internally at Microsoft and also using crowd sourcing tools. Some of the clips were taken from Audioset. The synthetic clips were divided into reverberant and less reverberant clips. This utterances were predominantly in English. All the clips are sampled at 16 kHz with average clip length of 12 secs. The development phase test set is in the **"ICASSP_dev_test_set"** directory in the DNS Challenge repository.

For this challenge, the primary focus is to make the test set as realistic and diverse as possible.

4.1. Track 1

Similar to DNS Challenge 1, the test set for DNS Challenge 2 is divided into real recordings and synthetic categories. However, the synthetic clips are mainly composed of the scenarios that we were not able to collect in realistic conditions. The track 1 test clips can be found in the **track_1** sub-directory of **ICASSP_dev_test_set**.

4.1.1. Real Recordings

The real recordings consists of non-English and English segments. The English segment will have 300 clips that are from the blind test set from DNS Challenge 1. These clips were collected using the crowd sourcing platform and internally at Microsoft using a variety of devices, acoustic and noisy conditions. The English segment of the test clips is accompanied by annotations to compute "Word Error Rate (WER)". We would encourage all the participants to report WER in their paper. The non-English segment comprises of 100 clips in the following languages: Portuguese, Russian, Spanish, Mandarin, Cantonese, Punjabi and Vietnamese. The last 4 are tonal languages. In total, there are 400 real test clips.

4.1.2. Synthetic test clips

It consists of 200 noisy clips obtained by mixing clean speech (non-English, emotional speech and singing) with noise. The testset noises is taken from Audioset and Freesound [9]. The 100 non-English test clips include German, French and Italian languages from librivox audio books. Emotion clean speech consists of laughter, yelling and crying chosen from Freesound and mixed with testset noise to generate 50 noisy clips. Similarly, clean singing voice from Freesound were used to generate 50 noisy clips for singing testset.

4.2. Track 2

For pDNS track, we provide 2 minutes of clean adaptation data for each primary speaker with goal to suppress neighboring speakers and background noise. pDNS models are expected to leverage speaker-aware training and speaker-adapted inference. There are two motivations to provide clean speech for primary speaker: (1) speaker models are sensitive to false-alarms in speech activity detection (SAD) [21]. Clean speech can be used for obtaining accurate SAD labels; (2) speaker adaptation is expected to work well using multi-conditioned data. Clean speech can be used for generating reverberant and noisy data for speaker adaptation.

4.2.1. Real Recordings

Development testset contains 100 real recordings from 20 primary speakers collected using crowd sourcing. Each primary speaker has noisy testclips for three scenarios: (i) primary speaker in presence of neighboring speaker; (ii) primary speaker in presence of background noise; and (iii) primary speaker in presence of both background noise and neighboring speaker.

4.2.2. Synthetic test clips

The synthetic clips include 500 noisy clips from 100 primary speakers. Each primary speaker has 2 minutes of clean adaptation data. All clips has varying levels of neighboring speaker and noise. TestSet noise from Track 1 was mixed with primary speech extracted from VCTK corpus [22]. We used VoxCeleb2 [23] corpus for neighboring speakers.

5. DNSMOS: A NO-REFERENCE OBJECTIVE SPEECH QUALITY METRIC

We developed a robust objective perceptual speech quality metric called DNSMOS. It can be used to stack rank different DNS methods based on Mean Opinion Score (MOS) estimates with great accuracy and hence the name DNSMOS. It is a neural network based model that is trained using the ground truth human ratings obtained using ITU-T P.808 [11, 9]. The audio data used to train the DNSMOS model is gathered from the numerous P.808 experiments that we conducted in the process of improving the quality of noise suppressor. DNSMOS is trained on the largest and the most extensive data set that we are aware of. We were able to achieve a Spearman rank correlation coefficient of 0.96 with subjective human ratings.

Sample code and details of the evaluation API can be found on `https://dns-challenge.azurewebsites.net/dnsmos/ICASSP2021`

6. EVALUATION METHODOLOGY

Most DNS evaluations use objective measures such as and PESQ [24], SDR and POLQA [25]. However, these metrics are shown to not have a high correlation to subjective speech quality in the presence of background noise [26]. DNSMOS is a useful metric to stack rank noise suppressors with high accuracy. However, subjective evaluation is the gold standard. Hence, the final evaluation will be done on the blind test set that is similar to development stage test set using the crowd-sourced subjective speech quality metric based on ITU P.808 [11] to determine the DNS quality. The DNS submission with the highest average P.808 Mean Opinion Score (MOS) across all scenarios and satisfies the computational complexity requirements will be the winner. To help provide a reference for evaluation, we will provide a baseline DNS model based on [27].

7. CHALLENGE RULES

- The participants must adhere to the requirements specified in section 2 for each track.
- Participants can use any data of their choice to train their models.
- Participants can use build a signal processing based or a learning based deep model. No restrictions on the algorithm.
- Submission must follow instruction on `https://dns-challenge.azurewebsites.net/`. Use Shift+F5 (Windows) or Cmd+R(Mac) to get the latest updates on that site.
- Winners will be picked based on the subjective evaluation using ITU-T P.808.
- Blind test for both the tracks will be made available on October 2nd 2020.
- For track 1, the participants must send the results (audio clips) achieved by their developed models to the organizers.
- For track 2, the participants must send the audio clips enhanced with and without using speaker information and must show that quality is better with using speaker information.
- We will use the submitted clips with no alteration to conduct ITU-T P.808 subjective evaluation and pick the winners based on the results. Participants are forbidden from using the blind test set to retrain or tweak their models. They should not submit results using other DNS methods that they are not submitting to ICASSP 2021. Failing to adhere to these rules will lead to disqualification from the challenge.
- Participants should report the computational complexity of their model in terms of the number of parameters and the time it takes to infer a frame on a particular CPU (preferably Intel Core i5 quad core machine clocked at 2.4 GHz). Among the submitted proposals differing by less than 0.1 MOS, the lower complexity model will be given a higher ranking.
- Each participating team must submit an ICASSP paper that summarizes the research efforts and provide all the details to ensure reproducibility. Authors may choose to report additional objective/subjective metrics in their paper.
- Submitted papers will undergo standard peer-review process of ICASSP 2021. The paper needs to be accepted to the conference for the participants to be eligible for the challenge.

8. TIMELINE

- September 8, 2020: Release of the datasets and scripts for training and testing.
- October 2, 2020: Blind test set released to participants.
- October 9, 2020: Deadline for participants to submit their results for objective and P.808 subjective evaluation on the blind test set.
- October 16, 2020: Organizers will notify the participants about the results.
- October 19, 2020: Regular paper submission deadline for ICASSP 2021.
- January 22, 2021: Paper acceptance/rejection notification
- January 25, 2021: Notification of the winners with winner instructions, including a prize claim deadline.

9. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] C. Karadagur Ananda Reddy, N. Shankar, G. Shreedhar Bhat, R. Charan, and I. Panahi, "An individualized super-gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1601–1605, 2017.

[4] S. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information*

Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 006–012.

[5] Hyeong-Seok Choi, Hoon Heo, Jie Hwan Lee, and Kyogu Lee, "Phase-aware single-stage speech denoising and dereverberation with u-net," *arXiv preprint arXiv:2006.00687*, 2020.

[6] Yuichiro Koyama, Tyler Vuong, Stefan Uhlich, and Bhiksha Raj, "Exploring the best loss function for dnn-based low-latency speech enhancement with temporal convolutional networks," *arXiv preprint arXiv:2005.11611*, 2020.

[7] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," *arXiv preprint arXiv:2008.04259*, 2020.

[8] Umut Isik, Ritwik Giri, Neerad Phansalkar, Jean-Marc Valin, Karim Helwani, and Arvindh Krishnaswamy, "Poconet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," *arXiv preprint arXiv:2008.04470*, 2020.

[9] 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," in *ISCA INTERSPEECH*, 2020.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE ICASSP*, 2015, pp. 5206–5210.

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

[12] Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo, "Vocalset: A singing voice dataset.," in *ISMIR*, 2018, pp. 468–474.

[13] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[14] Zhiyong Zhang Dong Wang, Xuewei Zhang, "Thchs-30 : A free chinese speech corpus," 2015.

[15] Xingyu Na Bengu Wu Hao Zheng Hui Bu, Jiayu Du, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA 2017*, 2017, p. Submitted.

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP*, 2017, pp. 776–780.

[17] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, pp. 3591, 05 2013.

[18] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE ICASSP*, 2017, pp. 5220–5224.

[19] Poju Antsalo, Aki Makivirta, Vesa Valimaki, Timo Peltonen, and Matti Karjalainen, "Estimation of modal decay parameters from noisy response measurements," in *Audio Engineering Society Convention 110*. Audio Engineering Society, 2001.

[20] Hannes Gamper, "Blind c50 estimation from single-channel speech using a convolutional neural network"," in *IEEE MMSP*, 2020, pp. 136–140.

[21] John HL Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[22] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[23] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *ISCA INTERSPEECH*, 2018.

[24] "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.

[25] John Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part II-perceptual model," *AES: Journal of the Audio Engineering Society*, vol. 61, pp. 385–402, 06 2013.

[26] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *IEEE ICASSP*, 2019, pp. 631–635.

[27] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *IEEE ICASSP*, 2020, pp. 871–875.