INTERSPEECH 2021 Deep Noise Suppression Challenge

Chandan K A Reddy[1], *Harishchandra Dubey*[1], *Kazuhito Koishida*[1], *Arun Nair*[2], *Vishak Gopal*[1],
Ross Cutler[1], *Sebastian Braun*[1], *Hannes Gamper*[1], *Robert Aichner*[1], *Sriram Srinivasan*[1]

[1]Microsoft Corporation, Redmond USA
[2]Johns Hopkins University, USA

`chkarada@microsoft.com, firstname.lastname@microsoft.com`

Abstract

The Deep Noise Suppression (DNS) challenge was designed to unify the research efforts in the area of noise suppression targeted for human perception. We recently organized a DNS challenge special session at INTERSPEECH 2020 and ICASSP 2021. We open-sourced training and test datasets for the wideband scenario along with a subjective evaluation framework based on ITU-T standard P.808, which was used to evaluate participants of the challenge. Many researchers from academia and industry made significant contributions to push the field forward, yet even the best noise suppressor was far from achieving superior speech quality in challenging scenarios. In this version of the challenge organized at INTERSPEECH 2021, we expanded our training and test datasets to accommodate fullband scenarios and challenging test conditions. We used ITU-T P.835 to evaluate the challenge winners as it gives additional information about the quality of processed speech and residual noise. The two tracks in this challenge focused on real-time denoising for (i) wideband, and (ii) fullband scenarios. We also made available a reliable non-intrusive objective speech quality metric for wideband called DNSMOS for the participants to use during their development phase.

Index Terms: DNS Challenge, Deep Noise Suppressor, Speech, Noise, Audio, Speech Quality

1. Introduction

With the explosion in the number of people working remotely due to the pandemic, there has been a surge in the demand for reliable collaboration and real-time communication tools. Excellent speech quality in our audio calls is a need during these times as we try to stay connected and collaborate with people every day. We are easily exposed to a variety of background noises such as a leaf blower, washing machine, dog barking, a baby crying, kitchen noises, 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 and smart devices as well.

Real-time Speech Enhancement (SE) for perceptual quality is a decades-old classical problem and researchers have proposed numerous solutions [1, 2]. In recent years, learning-based approaches have shown promising results [3, 4, 5]. The Deep Noise Suppression (DNS) Challenge organized at INTERSPEECH 2020 [6] and ICASSP 2021 [7] showed great progress, while also indicating that we are still about 1.6 Differential Mean Opinion Score (DMOS) away from the ideal Mean Opinion Score (MOS) of 5 when tested on the challenge test set, which was reasonably representative of realistic scenarios. The DNS Challenge is the first contest that we are aware of using the subjective evaluation to benchmark SE methods using a realistic noisy test set [6].

We open-sourced a large dataset for INTERSPEECH 2020 and ICASSP 2021 DNS challenge[1]. For ease of reference, we will call the INTERSPEECH 2021 challenge DNS Challenge 3, ICASSP 2021 challenge DNS Challenge 2, and the INTERSPEECH 2020 challenge DNS Challenge 1. The DNS Challenge 3 was focused on real-time denoising similar to track 1 of both the DNS challenges 1 and 2. We had 2 tracks in DNS challenge 3 for wideband (sampling rate = 16000 Hz) and fullband (sampling rate = 48000 Hz) scenarios. The datasets include over 760 hours of clean speech including singing voice, emotion data, and non-English languages. Noise data in the training set remains the same as DNS Challenge 2. Both clean speech and noise are made publicly available for both wide and fullband scenarios. We provide over 118,000 room impulse responses (RIR), which includes real and synthetic RIRs from public datasets for wideband. We provide acoustic parameters such as Reverberation Time (RT60) and Clarity (C50) for clean speech and the RIR audio sample. The test set [7] includes a variety of noisy speech utterances in English and non-English in a range of reverberant and noisy scenarios. We also include emotional speech and singing in the presence of background noise.

Unlike ITU-T P.808 that was used for DNS Challenge 1 and 2, we used the implementation of ITU-T P.835[2] [8] for the DNS Challenge 3. In addition to the overall speech quality as in P.808, P.835 provides standalone quality scores of speech and noise. The standalone ratings will help us focus on the areas that require improvement to achieve better overall speech quality. Many noise suppressors are very good in suppressing the background noise but do not improve the quality of speech, which becomes the bottleneck for improving the overall quality. The results of this challenge discussed in the later sections reflect the same. We also provided a non-intrusive objective speech quality metric for wideband scenario called DNSMOS[3] as an Azure service. We showed that DNSMOS is more reliable than other widely used objective metrics such as PESQ, SDR, and POLQA [9]. Also, it does not require reference clean speech and hence can work on real recordings. This paper describes the datasets, challenge results, and the learnings from the challenge in more detail.

2. Challenge Tracks

The following were the algorithmic and computational requirements that each participant had to satisfy to be eligible for the challenge.

1. Track 1: Real-Time Denoising track for wideband

---

[1]`https://github.com/microsoft/DNS-Challenge`
[2]`https://github.com/microsoft/P.808`
[3]`https://github.com/microsoft/DNS-Challenge/tree/master/DNSMOS`

- 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 processor. 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 $\leq$ 40ms. For example, for a real-time system that receives 20ms audio chunks, if you use a frame length of 20ms with a stride of 10ms resulting in an algorithmic latency of 30ms, then you satisfy the latency requirements. If you use a frame of size 32ms with a stride of 16ms resulting in an algorithmic latency 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: Real-Time Denoising track for fullband

- Satisfy Track 1 requirements.

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 [10] with a focus on extensibility, but the dataset lacked the diversity in speakers, emotions, languages, and noise types. We published a significantly larger and more diverse data set with configurable scripts for DNS Challenge 1 and 2 [6]. Many researchers found this dataset useful to train their noise suppression models and achieved good results. However, the training and the test datasets again lacked more clips with emotions such as crying, yelling, laughter or singing. Also, the dataset only included the clips in the English language. For DNS Challenge 3, we added speech clips with other emotions and included about 10 non-English languages. The clean speech in the training set resulted in a total of 760.53 hours: read speech (562.72 hours), singing voice (8.80 hours), emotion data (3.6hours), Chinese mandarin data (185.41 hours). We have grown clean speech to 760.53 hours as compared to 562.72 hours in DNS Challenge 1. 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) Read speech recorded in clean conditions; (ii) Singing clean speech; (iii) Emotional clean speech; and (iv) Non-English clean speech. The first subset is derived from the public audiobooks dataset called Librivox[4]. It is available under the permissive Creative Commons 4.0 license [11]. It has recordings of volunteers reading over 10,000 public domain audiobooks in various languages, the 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 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 [12] 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. We only chose the top 25% of the clips, which totalled to 562 hours of clean speech. All the files are resampled to 16 kHz.

The second subset consists of high-quality audio recordings of singing voices recorded in noise-free conditions by professional singers. This subset is derived from *VocalSet* corpus [13] 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 males, and 11 females. This data was recorded on a range of vowels, a diverse set of voices on several standards and extended vocal techniques, and sung in contexts of scales, arpeggios, long tones, and excerpts. For wideband, we downsampled the audio files from 44.1 kHz to 16 kHz and added them to the clean speech corpus used by the training data synthesizer.

The third subset consists of emotional speech recorded in noise-free conditions. This is derived from Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [14] made available under the Open Database License. It consists of 7,442 audio clips from 91 actors: 48 male, and 43 female accounting for a total of 3.5 hours of audio. The age of the actors was in the range of 20 to 74 years with diverse ethnic backgrounds including African American, 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 sampled at 16 kHz.

The fourth subset has a 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 [5] THCHS-30 [15] and OpenSLR33 [6] AISHELL [16] datasets, both with the Apache 2.0 license. THCHS30 was published by the 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 the entire THCHS-30 data in our clean speech for the training set. It consisted of 40 speakers: 9 male, 31 female in the age range of 19-55 years. It has total 13,389 clean speech audio files [15]. The 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. The audio was recorded in noise-free conditions using high-fidelity microphones. It is provided as 16-bit 16kHz files. It is one of the largest open-source Mandarin speech datasets. We added the entire AISHELL corpus with 141,600 utterances spanning 170+ hours of clean Mandarin speech to our training set. Spanish data is 46 hours of clean

---

[4]https://librivox.org/

[5]http://www.openslr.org/18/
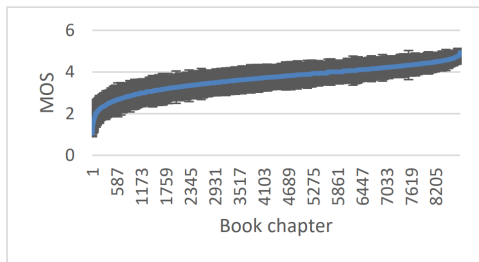[6]http://www.openslr.org/33/

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

speech derived from OpenSLR39, OpenSLR61, OpenSLR71, OpenSLR73, OpenSLR74 and OpenSLR75 where re-sampled all files from 48 kHz to 16 kHz to use them as wideband signals. German data is derived from four corpora namely (i) The Spoken Wikipedia Corpora [17], (ii) Telecooperation German Corpus for Kinect [18], (iii) M-AILABS data [19], (iv) zamia-speech forschergeist corpora. Complete German data constitute 636 hours. Italian (128 hours), French (190 hours), Russian (47 hours) are taken from M-AILABS data [19]. M-AILABS Speech Dataset is a publicly available multi-lingual corpora for training speech recognition and speech synthesis systems.

3.2. Noise

The noise clips were selected from Audioset [7] [20] and Freesound [8]. 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, to strictly separate speech and noise data. 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 [21]. The chosen noise types are more relevant to VoIP applications. In total, there is 181 hours of noise data. The noise files were originally fullband, which were resampled to 16 kHz for wideband use case.

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 reverberant clean speech while DNS models are expected to take noisy reverberant speech and produce clean reverberant speech. Challenge participants can do both de-reverb and denoising with their models if they prefer. These RIRs are chosen from openSLR26 [22] [9] and openSLR28 [22] [10] datasets, both released with Apache 2.0 License.

3.4. Acoustic parameters

We provide two acoustic parameters: (i) Reverberation time, T60 [23] and (ii) Clarity, C50 [24] 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. An RIR with low C50 can be described as highly reverberant and vice versa [23, 24]. These parameters are supposed to provide flexibility to researchers for choosing a sub-set of provided data for controlled studies.

4. Test set

For DNS Challenge 3, the test set included utterances in English and non-English languages recording in the presence of a variety of background noises at different SNR, target levels, and acoustic conditions. Non-English languages included tonal languages such as Punjabi, Vietnamese, Mandarin, and Cantonese. Other Non-English languages included Spanish, German, Portuguese and French. We crowd-sourced the noisy speech collection efforts to get diversity in terms of languages, acoustic conditions, noise environments, speaker's age, ethnicity and to have gender balance. Participants were instructed to collect the utterances at a distance of 1-5 meters from the microphone when they were not using a headphone to have more reverberation. The development and blind sets included utterances with emotions such as laughter, crying, yelling, and surprise in the presence of background noise. This is to measure the effects of noise suppressors on human emotions. Many noise suppressors tend to be aggressive and end up suppressing low-energy emotions and sounds. A small segment of the clips includes speech in the presence of musical instruments such as guitar, piano, violin playing in the background. This is to ensure that noise suppression methods do well in the presence of musical tones overlapping with speech. We also included speech collected in the presence of stationary noise as it is the most common use case scenario. All the clips were originally collected at a sampling rate of 48 kHz and were downsampled to 16 kHz.

5. Challenge Results and Key Takeaways

5.1. Evaluation set up and results

The final evaluation was done on the blind test set using the crowdsourced subjective evaluation framework based on ITU-T P.835 [8] to determine the DNS quality. Each clip was rated by 5 qualified raters, which gave the maximum 95% Confidence Interval (CI) of 0.05 DMOS per model. A total of 19 teams participated in track 1 and only 3 teams participated in track 2. Figures 2 and 3 shows the *Speech MOS*, *Background Noise MOS* and the *Overall MOS* results for track 1 and 2 respectively. The tables contain Differential MOS (DMOS) which is the difference between the MOS of the processed clips by a model (or the team) and the MOS of the original noisy speech. The original noisy speech MOS is shown inside the parenthesis in the row containing DMOS for "Noisy". The absolute MOS can be computed by adding the MOS of "Noisy" with the DMOS of a particular model of interest. The participants will be ranked based on *Overall MOS* given they satisfy other challenge requirements. Participants are required to submit the number of operations per second of their model. This will be used as a tie-breaker. The challenge also requires the participating teams to submit a paper to INTERSPEECH 2021 explaining their method and get the paper accepted.

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
38	(0.00)	0.05	(0.07)	0.12	0.03	(0.02)	0.03	0.04
36	0.01	0.07	(0.00)	0.16	(0.17)	(0.11)	0.01	0.04
Noisy	0 (4.02)	0 (3.83)	0 (3.93)	0 (3.8)	0 (3.97)	0 (3.87)	0 (3.89)	0.04
40	(0.04)	(0.08)	(0.17)	0.03	(0.21)	(0.23)	(0.10)	0.04
33	(0.13)	(0.15)	(0.21)	0.03	(0.13)	(0.23)	(0.12)	0.04
13	(0.15)	(0.15)	(0.17)	0.04	(0.26)	(0.24)	(0.13)	0.04
19	(0.21)	(0.18)	(0.13)	0.06	(0.26)	(0.31)	(0.15)	0.04
34	(0.16)	(0.17)	(0.12)	0.08	(0.41)	(0.36)	(0.17)	0.04
1	(0.09)	(0.24)	(0.20)	0.06	(0.38)	(0.34)	(0.17)	0.04
18	(0.35)	(0.48)	(0.47)	(0.06)	(0.73)	(0.54)	(0.39)	0.04
30	(0.44)	(0.80)	(0.23)	(0.24)	(0.53)	(0.49)	(0.43)	0.05
20	(0.48)	(0.65)	(0.38)	(0.18)	(0.62)	(0.60)	(0.45)	0.04
8	(0.47)	(0.57)	(0.37)	(0.17)	(0.97)	(0.76)	(0.52)	0.05
31	(0.45)	(0.63)	(0.54)	(0.26)	(0.84)	(0.68)	(0.53)	0.05
Baseline	(0.49)	(0.68)	(0.49)	(0.27)	(0.74)	(0.72)	(0.54)	0.04
4	(0.62)	(0.69)	(0.65)	(0.51)	(0.50)	(0.73)	(0.61)	0.05
22	(0.46)	(0.94)	(0.63)	(0.33)	(0.85)	(0.80)	(0.62)	0.05
12	(0.54)	(1.08)	(0.56)	(0.31)	(1.16)	(0.86)	(0.69)	0.05
11	(0.82)	(1.09)	(0.71)	(0.41)	(1.03)	(0.84)	(0.76)	0.05
37	(0.72)	(0.99)	(0.64)	(0.43)	(1.10)	(1.01)	(0.78)	0.05
28	(0.94)	(1.39)	(0.87)	(0.66)	(1.55)	(1.12)	(1.03)	0.05

(a) *Speech MOS*

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
36	1.92	2.73	1.82	1.54	2.39	2.31	2.05	0.02
1	1.90	2.59	1.71	1.54	2.25	2.21	1.98	0.03
18	1.82	2.63	1.61	1.42	2.07	2.24	1.91	0.03
33	1.78	2.42	1.54	1.40	2.27	2.13	1.87	0.03
13	1.73	2.28	1.48	1.34	2.01	1.94	1.74	0.03
22	1.77	2.20	1.37	1.36	1.91	1.97	1.73	0.03
34	1.78	2.15	1.51	1.28	1.84	1.84	1.68	0.03
8	1.62	2.01	1.36	1.17	1.81	1.83	1.59	0.03
37	1.69	2.14	1.46	1.13	1.61	1.78	1.58	0.04
19	1.48	1.92	1.36	1.14	1.84	1.68	1.52	0.03
12	1.82	2.13	1.44	1.07	1.13	1.58	1.47	0.04
Baseline	1.35	1.68	1.32	0.92	0.97	1.64	1.28	0.04
20	1.55	1.61	1.29	1.06	1.04	1.34	1.28	0.04
11	1.39	1.52	0.95	0.86	1.43	1.30	1.20	0.04
40	1.50	1.52	0.97	0.86	1.10	1.21	1.15	0.04
31	1.24	1.70	1.08	0.72	1.21	1.21	1.12	0.04
28	1.01	1.34	0.91	0.78	0.80	1.22	1.00	0.04
30	1.53	1.23	0.87	0.64	0.62	0.59	0.85	0.05
4	0.24	0.49	0.26	0.13	0.33	0.16	0.23	0.04
Noisy	0 (2.86)	0 (1.93)	0 (2.91)	0 (3.11)	0 (2.14)	0 (2.3)	0 (2.6)	0.04
38	(0.09)	(0.04)	(0.04)	(0.03)	0.04	0.01	(0.02)	0.04

(b) *Background Noise MOS*

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
36	0.89	1.51	0.79	0.80	1.14	1.11	1.01	0.04
1	0.85	1.16	0.65	0.69	0.89	0.92	0.85	0.04
33	0.74	1.16	0.56	0.57	1.06	0.93	0.81	0.04
13	0.76	1.14	0.60	0.60	0.97	0.90	0.80	0.04
34	0.69	1.13	0.59	0.64	0.76	0.75	0.74	0.04
19	0.61	0.98	0.58	0.57	0.90	0.74	0.71	0.04
18	0.59	1.02	0.40	0.55	0.55	0.75	0.64	0.04
40	0.79	0.94	0.37	0.40	0.62	0.63	0.60	0.04
8	0.41	0.79	0.37	0.39	0.25	0.41	0.42	0.04
22	0.51	0.50	0.18	0.30	0.42	0.42	0.39	0.05
20	0.44	0.48	0.28	0.36	0.29	0.39	0.38	0.04
31	0.37	0.60	0.20	0.20	0.28	0.35	0.32	0.04
Baseline	0.25	0.47	0.31	0.21	0.21	0.41	0.30	0.04
12	0.39	0.33	0.29	0.23	(0.00)	0.28	0.25	0.04
30	0.44	0.27	0.31	0.12	0.16	0.17	0.22	0.04
37	0.18	0.41	0.15	0.13	0.13	0.20	0.19	0.04
11	0.07	0.25	(0.08)	0.09	0.16	0.25	0.14	0.04
38	(0.12)	0.04	(0.10)	0.02	0.09	0.06	0.01	0.04
Noisy	0 (3.03)	0 (2.28)	0 (3)	0 (3.04)	0 (2.57)	0 (2.52)	0 (2.77)	0.04
28	(0.12)	(0.07)	(0.10)	(0.12)	(0.44)	(0.02)	(0.13)	0.04
4	(0.22)	0.13	(0.26)	(0.27)	0.02	(0.15)	(0.15)	0.04

(c) *Overall MOS*

Figure 2: *Track 1 results*

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
Noisy	0 (3.9)	0 (3.6)	0 (3.75)	0 (3.8)	0 (3.81)	0 (3.77)	0 (3.78)	0.04
1	(0.13)	(0.23)	0.09	0.00	(0.19)	(0.34)	(0.14)	0.04
32	(0.41)	(0.49)	(0.49)	(0.40)	(0.70)	(0.73)	(0.54)	0.05
35	(0.88)	(0.75)	(1.09)	(0.57)	(1.11)	(0.94)	(0.84)	0.05

(a) *Speech MOS*

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
1	2.02	2.49	1.62	1.54	2.23	2.26	1.99	0.03
32	1.36	1.70	1.00	0.75	1.11	1.34	1.16	0.04
35	1.23	0.86	0.76	0.70	0.65	0.72	0.80	0.05
Noisy	0 (2.7)	0 (1.87)	0 (2.97)	0 (3.08)	0 (2.23)	0 (2.2)	0 (2.56)	0.05

(b) *Background Noise MOS*

Team #	Stationary DMOS	Emotional DMOS	Tonal DMOS	Non-English DMOS	Musical DMOS	English DMOS	Overall DMOS	CI
1	0.81	1.01	0.72	0.71	0.91	0.87	0.82	0.04
32	0.42	0.61	0.08	0.12	0.20	0.32	0.27	0.04
Noisy	0 (2.92)	0 (2.25)	0 (2.96)	0 (3.0)	0 (2.6)	0 (2.43)	0 (2.71)	0.04
35	(0.02)	0.07	(0.44)	(0.11)	(0.27)	(0.12)	(0.13)	0.04

(c) *Overall MOS*

Figure 3: *Track 2 results*

ponents. This is due to inaccurate estimation of suppression mask as a result of poor learning capability of the model.

2. The *Background Noise MOS* in 2b shows that almost all the models are trained well to suppress the background noise. The top 3 models achieved a DMOS of almost 2 (MOS of 4.6), which is a significant improvement over noisy speech.

3. According to figure 2c, the best model achieved an *Overall MOS* of about 3.78, which is about 1.2 DMOS less than the perfect quality of MOS 5. This shows that achieving superior noise suppression without distorting speech is a challenging problem. Analysis in [8] shows that the best theoretical *Overall MOS* that can be achieved is 3.92 assuming *Background Noise MOS* of 5 and *Speech MOS* same as that of noisy speech. Hence, the field of speech enhancement optimized for human perception is still in its nascent phase.

4. Researchers can get the biggest bang for the buck in terms of *Overall MOS* by improving the *Speech MOS* by maintaining excellent *Background Noise MOS*.

6. Conclusions

The INTERSPEECH 2021 DNS Challenge was organized to help researchers from academia and industry to come together and tackle this challenging problem in speech enhancement. Large inclusive and diverse training and test datasets with supporting scripts were open sourced along with other tools such as perceptual objective metrics. Many participants from both industry and academia found the datasets very useful.

7. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE TASLP*, 1984.

[2] 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

5.2. Key takeaways

1. We can see from figure 2a that all the teams except for the top 2 did worse than noisy in terms of speech quality measured by *Speech MOS*. Most of the noise suppressors introduce speech distortion when they get aggressive in suppressing noise and end up suppressing speech com-

smartphone as an assistive device," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1601–1605, 2017.

[3] 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)*.

[4] H.-S. Choi, H. Heo, J. H. Lee, and K. Lee, "Phase-aware single-stage speech denoising and dereverberation with U-net," *arXiv preprint arXiv:2006.00687*, 2020.

[5] Y. Koyama, T. Vuong, S. Uhlich, and B. Raj, "Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks," *arXiv preprint arXiv:2005.11611*, 2020.

[6] C. K. Reddy *et al.*, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *ISCA INTERSPEECH*, 2020.

[7] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6623–6627.

[8] B. Naderi and R. Cutler, "A crowdsourcing extension of the itu-t recommendation p. 835 with validation," *arXiv preprint arXiv:2010.13200*, 2020.

[9] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[10] C. K. Reddy *et al.*, "A scalable noisy speech dataset and online subjective test framework," *arXiv preprint arXiv:1909.08050*, 2019.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE ICASSP*, 2015.

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

[13] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset." in *ISMIR*, 2018.

[14] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[15] Z. Z. Dong Wang, Xuewei Zhang, "THCHS-30 : A free chinese speech corpus," 2015. [Online]. Available: http://arxiv.org/abs/1512.01882

[16] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE.

[17] "The Spoken Wikipedia Corpora," https://nats.gitlab.io/swc/, [Online; accessed 2020-09-01].

[18] "Telecooperation German Corpus for Kinect," http://www.repository.voxforge1.org/downloads/de/german-speechdata-TUDa-2015.tar.gz, [Online; accessed 2020-09-01].

[19] "M-AILABS Speech Multi-lingual Dataset," https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/, [Online; accessed 2020-09-01].

[20] 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.

[21] J. Thiemann, N. Ito, and E. 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*, p. 3591, 05 2013.

[22] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE ICASSP*, 2017.

[23] P. Antsalo *et al.*, "Estimation of modal decay parameters from noisy response measurements," in *Audio Engineering Society Convention 110*, 2001.

[24] H. Gamper, "Blind C50 estimation from single-channel speech using a convolutional neural network," in *Proc. IEEE MMSP*, 2020, pp. 136–140.