ICASSP 2022 DEEP NOISE SUPPRESSION CHALLENGE

Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, Robert Aichner

Microsoft Corporation, Redmond, USA
firstname.lastname@microsoft.com

ABSTRACT

The Deep Noise Suppression (DNS) challenge is designed to foster innovation in the area of noise suppression to achieve superior perceptual speech quality. This is the 4th DNS challenge, with the previous ones held at INTERSPEECH 2020, ICASSP 2021, and INTERSPEECH 2021. We open-source training and test datasets for researchers to train their deep noise suppression models, as well as a subjective evaluation framework used to evaluate and select the final winners. Many researchers from academia and industry have 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 excellent speech quality in challenging noisy real-world scenarios. In this latest challenge we make the following changes: (1) include mobile devices scenarios in the test set, (2) include a personalized noise suppression track, (3) add Word Accuracy (Wacc) as an objective metric, (4) include DNSMOS P.835, (5) the datasets and test sets are fullband (48 kHz).

Index Terms— Speech Enhancement, Perceptual Speech Quality, P.835, Deep Noise Suppressor, DNS, Machine Learning.

1. INTRODUCTION

In recent times, remote work has become the "new normal" as the number of people working remotely has increased significantly due to the pandemic. There has been a surge in the demand for reliable collaboration and real-time communication tools. Audio/Video calls with excellent speech quality are needed during these times as we try to stay connected and collaborate with people every day. We are often exposed to a variety of background noises, including a dog barking, a baby crying, kitchen noises, neighbouring talkers, etc., which may significantly degrade the quality and intelligibility of the perceived speech and lead to to increased fatigue in virtual meetings.

Real-time noise suppression for improving the perceptual quality of speech is a classical problem and researchers have proposed numerous solutions [1]. In recent years, deep learning based approaches have shown promising results for superior speech quality [2, 3, 4]. Previous DNS Challenges at INTERSPEECH 2020, ICASSP 2021, and INTERSPEECH 2021 accelerated deep noise suppression research by providing a massive training dataset, real testset, training data synthesizer and subjective evaluation frameworks based on ITU-T P.808 [5] and P.835 [6]. An increasing number of recent DNS research papers are leveraging DNS Challenge datasets for training and validating their models [7, 4].

The 4th DNS Challenge at IEEE ICASSP 2022 is intended to promote industry–academia collaboration for real-time deep noise suppression aimed to maximize the subjective (perceptual) quality of enhanced speech. This challenge focuses on personalized and non-personalized DNS for fullband (48 kHz) audio. In the era of hybrid work, personalized denoising is very important to suppress neighboring talkers and/or background noises. This challenge provides fullband datasets for training personalized and non-personalized deep noise suppressors. We focused on collecting real-world test sets and developed a framework for P.835 subjective evaluation for personalized and non-personalized deep noise suppressors. The 4th DNS Challenge has two tracks namely (1) personalized DNS; and (2) Non-personalized DNS. The development testsets for both tracks has approximately 1500 real testclips, while the personalized track has 2.5 minutes of enrollment speech for each primary talkerincluded in personalized DNS testset. All testclips in development set was collected through crowdsourcing where workers recorded the audio in different acoustic scenarios using a variety of Desktop/Laptops. The blind testset for both tracks have 3000 testclips where 1500 testclips were recorded on mobile devices (Android); and rest 1500 testclips were recorded on Desktop/Laptop platforms.

Majority of research papers in deep noise suppression are still reporting objective quality metrics such as Perceptual Evaluation of Speech Quality (PESQ), Perceptual Objective Listening Quality Analysis (POLQA), Signal to Noise Ratio (SNR), Signal to Distortion Ratio (SDR), etc. Previously, we found that these objective measures correlate poorly with crowd-sourced subjective human evaluations [4]. Hence, such objective metrics are not suitable for evaluating DNS models. The 4th DNS Challenge evaluates the submitted models based on ITU-T P.835 subjective evaluation scores namely speech quality (SIG), background noise quality (BAK), and overall audio quality (OVRL); and Word Accuracy (Wacc) from a state-of-the-art speech recognition system. In addition, we open sourced DNS-MOS P.835 [8] which is a deep learning model that predicts SIG, BAK, OVRL scores. The DNSMOS P.835 helps in intermediate evaluation of DNS models trained by DNS Challenge participants.

2. CHALLENGE TRACKS

The 4th DNS Challenge at ICASSP 2022 has two tracks namely, (1) non-personalized DNS and (2) Personalized DNS (PDNS) for fullband (48 kHz) audio. Unlike previous DNS challenges, we do not have wideband (16 kHz) data in our training and testset. Personalized deep noise suppression (PDNS) leverages speaker embeddings (features) for preserving only the primary talker in a noisy environment and suppresses neighboring talkers and noise. PDNS testset consists of real recordings for three scenarios: (i) primary talker in the presence of noise; (ii) primary talker in presence of neighboring talkers; (iii) primary talker in presence of neighboring talker and noise. In a given PDNS testclip, there can be a maximum of one neighboring talker. We provide development testset (dev testset) which is collected through crowd-sourcing where workers read provided text prompts and record their voice using Desktop/Laptop

in presence of noise and/or neighboring talkers. The Blind testset is planned to include testclips recorded on desktop as well as mobile in realistic noisy environments. Similar to previous DNS Challenge, we provide training data synthesizer which can be used with other datasets participants may choose to use. Data synthesizer, configuration and data download scripts are provide in Github repo for 4th DNS Challenge [1].

We provide baseline models for both track in terms of enhanced testclips for respective tracks. We briefly describe baseline models for non-personalized DNS and PDNS tracks in later section of this paper. We also provide P.835 subjective evaluation framework and its modified version for PDNS evaluation. We introduced Word Accuracy (Wacc) as an objective metric for impact on DNS on performance of speech recognition systems. Higher Wacc shows superior denoising performance of DNS models. The subjective test ITU-T P.835 [6] provides three scores for each audio clip, namely speech quality (SIG), background noise quality (BAK), and overall quality (OVRL). Participants submitted enhanced clips for one or both tracks. We conduct ITU-T P.835 and Wacc computation on submitted enhanced clips. Models which do well in all four metrics namely SIG, BAK, OVRL, and Wacc are ranked higher. The motivation to add Wacc as additional evaluation metric stem from the fact that several models from past DNS Challenges had noticeable Wacc degradation resulting from over suppression of noise and speech distortions. We provide an Azure service for estimating Wacc and another Azure service for DNSMOS P.835 [8] which is a deep neural network (DNN) model for prediction of speech, background noise, and overall quality.

In this challenge, we cleaned the clean speech in the training dataset. We provide the DNS-MOS P.835 score for all training to enable participants choose different threshold for SIG, BAK and OVRL for choosing clean speech consumer by training data synthesizer. We summarize below the requirements for two tracks in this challenge:

1. Track 1: requirements for real-time denoising
    - 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 $\leq$ 40ms. For example, for a real-time system that receives 20ms audio chunks, if a frame length of 20ms with a stride of 10ms is used that results in an algorithmic latency of 30ms, then the latency requirements are satisfied. If a frame size of 32ms with a stride of 16ms is used, resulting in an algorithmic latency of 48ms, then the latency requirements are not met as the total algorithmic latency exceeds 40ms. If the frame size plus stride $T_1 = T+T_s$ is less than 40ms, then up to $(40 - T_1)$ms future information can be used.
2. Track 2: requirements for personalized real-time denoising
    - Satisfy Track 1 requirements.
    - 2.5 minutes of clean speech for each primary talker in the test set is provided for adopting PDNS model for primary talker. PDNS track has a separate dev test set and blind test set.

[1] https://github.com/microsoft/DNS-Challenge/

3. TRAINING DATASETS

We provide raw clean speech, noise. impulse responses and training data synthesizer for both tracks. Same noise and impulses responses are being provided for both tracks. Each track has it training data synthesizer. PDNS track has clean speech where each audio clip is concatenations of all audio clips belonging to a talker. We provide 2.5 minutes enrollment speech for each talker in training clean speech. We also provide baseline speaker embeddings for each talker in PDNS training set. We cleaned the clean speech portion of training set by choosing the clips with more than 4.25 DNS-MOS P.835 OVRL score. We verified a random sample of clean data with crowd-sourced ITU-T P.835 OVRL to have score more than 4. PDNS tracks leverages cleaned clean speech with score greater than 4.25 DNS-MOS P.835. Next, the audio files for each talker is combined into a single clip. We randomly sample 2.5 minutes of clean speech for each speaker and provide it as enrollment speech. We extract speaker embedding for each enrollment audio using baseline RawNet2 speaker model [9]. Our training data consists of english read-speech, english singing voice, french, german, italian, russian and spanish languages. Next, we describe the clean and noise dataset in the following sections.

3.1. Clean Speech

Clean speech consists of six languages namely English, French, German, Italian, Russian and Spanish. English clean speech consists of read speech and singing voice while rest of the languages only have read speech. Non-personalized training set consist of original testclips from various corpora. We provide DNS-MOS P.835 scores to help participants filter the data based on DNS-MOS scores. Personalized track consists of clips with DNS-MOS P.835 OVRL score greater than 4.25. We combine all audio clips from each unique talker into a single file. PDNS training data has total of 3230 talkers out of which 60% talkers are randomly chosen to be primary talker rest are neighboring talker. We provide the filelist for PDNS cleanspeech with 'primary' and 'secondary' tags. Challenge participants can use the provided primary/secondary tags or generate their own. We smapled 60% speakers as primary ensuring uniform distribution among all langauges, read speech or singing voice.

English clean speech is derived from Librivox [2] where we include audio clips chosen using a subjective ITU-T P.808 framework [5]. English singing voice data consists of high-quality audio recordings from professional singers contained in *VocalSet* corpus [10]. 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 standard and extended vocal techniques, and sung in contexts of scales, arpeggios, long tones, and excerpts. PDNS English clean speech contained 1934 talkers from Librivox, 110 talkers from VCTK corpus, 20 talkers from Vocalset. PDNS clean speech consists of 47 talkers for French, 874 talkers from German, 14 talkers from Italian, 7 talkers from Russian, and 224 talkers from Spanish languages.

3.2. Noise

We have same noise clips for both tracks. Noise data consists of 62000 clips belong to 150 noise classes. The noise clips were cho-

[2] https://librivox.org/

sen from Audioset [3] [11] and Freesound [4]. Audioset is a collection of about 2 million human labeled 10s sound clips drawn from YouTube videos belonging to about 600 audio events. Certain audio event classes are over-represented in Audioset. 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. We developed a sampling approach to balance the noise classes in our dataset such that each class has at least 500 clips. We used a speech activity detector to remove the clips with any kind of speech activity (voice content). This enabled us to get noise clips with no presence of speech. We augmented chosen Audioset noise clips with 10,000 noise clips downloaded from Freesound and DEMAND databases [12]. Our noise data constitute 181 hours.

3.3. Impulse Responses

We provide 248 real and 60,000 synthetic rooms impulse responses which can be leveraged for generating reverberant noisy training data. Training data synthesizer adds noise to reverberant clean speech. Participants may chose to use clean speech or reverberant speech as training targets for their DNS models. We chose impulse responses from openSLR26 [13] [5] and openSLR28 [13] [6] datasets.

4. TEST SET

We have two testset for each track namely development testset (dev testset) and blind testset. Dev test is intended for model development and optimization. Blind testset is used for ranking the challenge model in terms of evaluation metrics. Good performance of a model on blind testset will shows its generalizability for unseen conditions. Our testset consists of fullband (48 kHz) audio clips recorded in real-world scenarios.

4.1. Non-personalized Dev Testset

Dev testset for non-personalized DNS consists of 930 real recordings. All clips contains noisy speech in English language. Among these, 193 testclips have emotional speech in presence of noise. There are six emotion types namely happy, sad, angry, yelling, crying and laughter. Crowd-sourced workers were asked to read a provided text file and they were supposed to create emotional events in each testclip. Rest of the clips contain the voice of a talker reading text in presence of following noise types: Fan, Air conditioner, Typing, Door shutting, Clatter Noise, Car noise (i.e. standing near a car on a busy street or are standing outside the car), Kitchen noise (noise from kitchen utensils, dish scrubbing etc.), Dish Washer, Running Water, Opening chips packet, Munching or eating with noise, Creaking chair, Heavy Breathing, Copy machine, Baby crying, Dog barking, Inside-car noise (e.g., sitting on a passenger seat in a car which is being drive by someone else), Mouse clicks, mouse scroll wheel, touchpad clicks etc. Each testclip was recorded at 48 kHz with 10-20 second duration. Workers were asked to record in near-field (close talk) and far-filed with distances of 1 metre, 2 metres and 3 metres. All testclips in non-personalized Dev Testset were recorded using Laptop/Desktop.

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

4.2. Personalized Dev Testset

Dev testset for Personalized track consists of 1443 real recordings. All clips contains noisy speech in English language. Among these, 193 testclips have emotional speech in presence of noise where only one talker is reading. There are six emotion types namely happy, sad, angry, yelling, crying and laughter. Crowd-sourced workers were asked to read a provided text file and they were supposed to create emotional events in each testclip. There are 737 testclips where primary talker reads the provided text in presence of following noise types: Fan, Air conditioner, Typing, Door shutting, Clatter Noise, Car noise (i.e. standing near a car on a busy street or are standing outside the car), Kitchen noise (noise from kitchen utensils, dish scrubbing etc.), Dish Washer, Running Water, Opening chips packet, Munching or eating with noise, Creaking chair, Heavy Breathing, Copy machine, Baby crying, Dog barking, Inside-car noise (e.g., sitting on a passenger seat in a car which is being drive by someone else), Mouse clicks, mouse scroll wheel, touchpad clicks etc. Each testclip was recorded at 48 kHz with 10-20 second duration. Workers were asked to record in near-field (close talk) and far-filed with distances of 1 metre, 2 metres and 3 metres. There are 166 testclips with primary talker speaking in presence of a neighboring talker and noise where both the noise and neighboring talker is simultaneously active in primary talker's background. There are 347 testclips where primary talker is speaking in presence of a neighboring speaker with no background noise. Thus, we have simulated three scenarios for PDNS: (i) primary talker in presence of noise; (ii) primary talker in presence of neighboring talker; and (iii) primary talker in presence of simultaneously active neighboring talker and noise. All testclips in personalized Dev Testset were recorded using Laptop/Desktop.

5. BASELINE MODELS

5.1. RawNet2: Baseline Speaker Embedding Extractor

RawNet2 [14] was trained on wideband (16 kHz) audio from vox-Celeb data. The participants can retrain the RawNet2 model with fullband data if they prefer to do so.

5.2. Baseline for Non-personalized DNS

We trained NSNet2 [15] on fullband non-personalized training dataset to obtain the baseline. We denoise the dev test with baseline model and release only the enhanced clips to challenge participants. In 4th DNS Challenge, we do not release the trained model or training code of baseline model.

5.3. Baseline for Personalized DNS

PDNS refers to DNN models which suppress neighboring talker and background noise and only preserve the enhanced speech from the primary talker. To achieve this, PDNS models leverage speaker embedding features extracted from the enrollment speech (or noisy audio) along with spectral features (or raw-waveform) of the noisy input audio. Participants can leverage state-of-the-art speaker models [16, 17, 14] and may choose a speaker embedding extractor which is publicly available [16, 17]. We use the personalized DCCRN (pDCCRN) model described in [18] as the baseline PDNS model for 4th DNS Challenge. Our results on dev testset is shown in Table 1. The speaker embedding extractor was trained on wideband (16 kHz) data and hence the input audio is first downsampled to 16 kHz for extracting the speaker embedding. The participants can

Table 1. Performance of baseline models on Dev Testset.

Model	Testset	DSIG	DBAK	DOVRL
NSNet2	Non-personalized dev testset			
pDCCRN	Personalized dev testset			

retrain the model with fullband data if they prefer to do so and then they can use fullband enrollment data for each talker in the testset.

Each unique talker in personalized training dataset and PDNS dev testset has 2.5 minutes of enrollment speech. PDNS models are expected to leverage talker-aware training and talker-adapted inference. There are two motivations to provide clean speech for enrollment of primary talker: (1) speaker models are sensitive to false-alarms in speech activity detection (SAD) [19]; clean speech can be used for obtaining accurate SAD labels which can help improve speaker embeddings. (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.

6. CHALLENGE RESULTS

6.1. Evaluation Methodology

The objective measures of speech quality, such as PESQ [20], SDR, and POLQA [21] do not correlate well with subjective speech quality [22]. The 4th DNS Challenge relies on ITU-T P.835 [6] subjective evaluation for ranking the challenge entries. We modified the ITU-T P.835 for measuring the performance personalized DNS models. The ITU-T P.835 subjective evaluation scores namely speech quality (SIG), background noise quality (BAK), and overall audio quality (OVRL); and Word Accuracy (Wacc) from Speech Recognition System are the gold standard for this task. The final evaluation for ranking the challenge entries will be conducted on the blind testset

The P.835 for personalized DNS has modified clips where we appended each enhanced (or noisy) testclip after a 5 second speech for the primary talker followed by one second silence. We modified P.835 to instruct the human raters to focus on the quality of the voice of the primary talker in the remainder of the enhanced (or noisy) testclip. The 6 second of each appended clip is to make workers aware of primary talker's speech. We conducted a reproducibility test for modified P.835 and found the average Spearman Rank Correlation Coefficient (SRCC) between the 5 runs was . Hence, we can conclude that the modified P.835 is working as intended. For reproducibility study, we used 5 raters per clip, resulting in a 95% confidence interval (CI) of 0.03.

7. SUMMARY & CONCLUSIONS

The ICASSP 2022 DNS Challenge is designed to advance the field of real-time deep noise suppression optimized for human perception in challenging noisy conditions. We hope the participants enjoy the challenge as much as we had in creating it. The challenge dataset is opensourced to help researchers leverage that for pushing the field forward even after the conclusion of DNS Grand Challenge at ICASSP 2022.

8. REFERENCES

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

[2] 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.

[3] 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.

[4] Chandan K. A. Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram 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.

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

[6] Babak Naderi and Ross Cutler, "A crowdsourcing extension of the itu-t recommendation p. 835 with validation," *arXiv e-prints*, pp. arXiv–2010, 2020.

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

[8] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv preprint arXiv:2110.01763*, 2021.

[9] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

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

[11] 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.

[12] 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*, p. 3591, 05 2013.

[13] 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 *IEEE ICASSP*, 2017.

[14] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *arXiv preprint arXiv:2004.00526*, 2020.

[15] Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian

Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," *ISCA INTERSPEECH*, 2021.

[16] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[17] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[18] Sefik Emre Eskimez, Takuya Yoshioka, Huaming Wang, Xiaofei Wang, Zhuo Chen, and Xuedong Huang, "Personalized speech enhancement: New models and comprehensive evaluation," *arXiv preprint arXiv:2110.09625*, 2021.

[19] 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.

[20] "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.

[21] John Beerends et al., "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.

[22] 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.