Received 10 August 2023; revised 28 December 2023; accepted 2 January 2024. Date of publication 18 March 2024; date of current version 18 June 2024. The review of this article was arranged by Associate Editor Hung-yi Lee.

Digital Object Identifier 10.1109/OJSP.2024.3378602

ICASSP 2023 Deep Noise Suppression Challenge

HARISHCHANDRA DUBEY [1], ASHKAN AAZAMI [1], VISHAK GOPAL [1], BABAK NADERI [1], SEBASTIAN BRAUN [1] (Member, IEEE), ROSS CUTLER [1] (Member, IEEE), ALEX JU [1], MEHDI ZOHOURIAN [1], MIN TANG [1], MEHRSA GOLESTANEH [1], AND ROBERT AICHNER [1]

[1] Microsoft Corporation, Redmond, WA 98052 USA

CORRESPONDING AUTHOR: ROSS CUTLER (email: ross.cutler@microsoft.com).

ABSTRACT The ICASSP 2023 Deep Noise Suppression (DNS) Challenge marks the fifth edition of the DNS challenge series. DNS challenges were organized from 2019 to 2023 to foster research in the field of DNS. Previous DNS challenges were held at INTERSPEECH 2020, ICASSP 2021, INTERSPEECH 2021, and ICASSP 2022. This challenge aims to advance models capable of jointly addressing denoising, dereverberation, and interfering talker suppression, with separate tracks focusing on headset and speakerphone scenarios. The challenge facilitates personalized deep noise suppression by providing accompanying enrollment clips for each test clip, each containing the primary talker only, which can be used to compute a speaker identity feature and disentangle primary and interfering speech. While the majority of models submitted to the challenge were personalized, the same teams emerged as the winners in both tracks. The best models demonstrated improvements of 0.145 and 0.141 in the challenge's score, respectively, when compared to the noisy blind test set. We present additional analysis and draw comparisons to previous challenges.

INDEX TERMS Deep noise suppression, DNS challenge, perceptual speech quality, personalized deep noise suppression, personalized P.835, target speech extraction.

I. INTRODUCTION

In recent times, hybrid work has become the new normal as remote work has significantly increased following the COVID-19 pandemic. Video and audio calls are often degraded by various background noises, including munching sounds, paper shuffling, keyboard typing, mouse clicks, doors opening/closing, neighboring talkers, pets, babies crying, kitchen sounds, in-car noises, engine sounds, airport announcements, doorbells, traffic, and street noises. The presence of these noises on calls can lead to increased fatigue for participants. Furthermore, background noise can reduce participation in meetings. Therefore, achieving high speech quality in real-time video communication is crucial for inclusive and collaborative hybrid meetings. Solutions are needed to suppress these ambient noises to provide fatigue-free, highly intelligible audio during hybrid work video conferencing. By addressing this issue, we can improve meeting quality and productivity for the growing remote workforce.

Classic digital signal processing (DSP) techniques laid the foundation for noise suppression research. Common DSP-based approaches for noise suppression are mostly based on spectral suppression rules such as Wiener filtering or log-short-time spectral amplitude estimators [1], [2] and often model only stationary background noise [3]. More advanced techniques include multi-frame filtering [4]. Reverberation has to be modeled in a separate estimation module and combined suppression is not straightforward [5].

An overview of statistical model-based STFT domain noise suppression methods is presented in [6]. These approaches are appealing for real-time applications given their simplicity and low computational cost, but often struggle with non-stationary noise and speech distortions due to their simplistic model assumptions.

A good overview of early Deep Neural Network (DNN) based speech enhancement methods is given in [7]. The first main challenge of early works was that most approaches were not able to operate in real-time and if so, lost large performance gains. An important work designing a small DNN able to run on devices was RNNoise [8], which triggered a plethora of follow-up work and a series of DNS challenges, with this one being the fifth incarnation. The second main challenge is generalization to in-the-wild data and improving speech quality, i.e., removing all noise without creating additional artifacts and distorting the speech.

In recent years, deep learning-based noise suppression, further referred to as Deep Noise Suppression (DNS), has shown promising results with superior speech quality over classical approaches [9], [10], [11], [12]. The DNS challenges have been held in INTERSPEECH 2020, ICASSP 2021, INTERSPEECH 2021, ICASSP 2022, and ICASSP 2023. The DNS challenges have accelerated research progress by providing large training datasets, real recordings as test sets, a training data synthesizer, accurate objective functions [13], [14], and subjective evaluation frameworks based on ITU-T P.808 [15] and P.835 [16]. Many recent papers have leveraged these resources to develop advanced DNS models [12], [17], [18], [19], [20], [21]. From the second DNS challenge on, we introduced the task of personalized deep noise suppression (PDNS) [12], which uses speaker identity features from an independent speaker enrollment recording to focus only on the enrolled user speech and remove other interfering talkers.[1] While we feel the performance of speaker-independent speech enhancement, i.e., noise reduction and dereverberation, is slowly saturating after the large initial gains from previous challenges, PDNS has still significant room for improvement to generalize and be robust enough to disentangle primary and neighbor speakers in real-time. Therefore, PDNS is the focus of this challenge.

As part of the past four DNS challenges, we open-sourced training and test sets, and a P.835 subjective evaluation framework [16]. Our GitHub repository[2] open-sourced Personalized and Non-personalized DNSMOS P.835 [14] and word accuracy (WAcc) APIs to empower iterative model improvements for teams participating in the challenge. DNSMOS P.835 is a no-refernce deep learning model that predicts MOS (Mean Opinion Scores) for speech signal quality (SIG), background quality (BAK), and overall audio quality (OVRL) from a noisy or processed test audio clip. This reduces the barrier to entry in the field and provides standard tools for the evaluation of DNS models. Like previous challenges, each track has two test sets: (i) development (dev) test set which was released at the beginning of the challenge; (ii) blind test set released a few days before the final challenge deadline. While the dev test set enables intermediate model evaluations, the blind set is used for the final ranking of models based on challenge metrics.

A. WHAT IS NEW?

The fourth DNS challenge focused on personalized and non-personalized speech enhancement with fullband data. We have introduced the following changes in this challenge: (i) There are two tracks: Headset and Speakerphone, each containing both desktop and mobile recordings in their test sets; (ii) All test clips in both tracks include 10-30 seconds of enrollment speech (primary talker) with or without noise; (iii) The Personalized P.835 evaluation framework has been improved, now incorporating voice recognition, robust spam filtering, and more accurate evaluation of enhanced test clips with noise and neighboring talkers; (iv) The Personalized P.835 framework employs cleaned enrollment speech, enhanced using a non-causal model. We found in preliminary experiments that cleaned enrollment speech improves the consistency of subjective evaluation; (v) Both personalized and non-personalized models for a track were jointly subjected to the same subjective evaluation and ranking. In other words, personalized and non-personalized models are treated equally and compared; (vi) Separate subjective evaluations were conducted for both the Headset and Speakerphone tracks; (vii) The algorithmic plus buffering latency has been reduced from 40 ms to 20 ms to make it meet real-time communication system requirements. Achieving the same noise suppression performance with lower latency is a more challenging task.

The development test set, DNSMOS P.835 API, and WAcc API were provided at the start of the challenge to optimize models. The blind test set was released near the deadline for the final model ranking. Submitted models were evaluated on P.835 MOS scores (SIG, BAK, OVRL), and WAcc. The prediction model DNSMOS P.835, which predicts P.835 scores, was made freely available for use for development.

In this challenge, participants could use any datasets including external corpora and challenge training datasets to do model training. Participants were required to describe the datasets used for training their models in sufficient detail in their extended journal papers and provide a brief coverage in a 2-page ICASSP grand challenge paper. The challenge website[3] has details of scope and requirements; definitions of algorithmic latency, processing latency, causal model, real-time factor (RTF) and associated challenge rules, and name of winning teams, etc. Previous challenge websites are linked on the website[3] as well.

By introducing a more realistic test set with enrollment speech, improved P.835 evaluation, and joint assessment of personalized/non-personalized models, this latest challenge enables benchmarking on pertinent real-world use cases. Table 1 demonstrates the opportunity to further improve subjective quality on all dimensions based on the ICASSP 2022 challenge. Note that when given high-quality fullband audio with no distortions, our P.835 test framework achieved subjective ratings of BAK = 4.88, SIG = 4.96, OVRL = 4.74 [16]; Table 1 shows values far from these measurements.

II. RELATED WORK

An important part of this challenge is how to evaluate the model performance during development and for model comparisons. The gold standard for speech quality assessment is subjective testing carried out by human test participants who

[1] This task is also referred to as target speech extraction [22]

[2] [Online]. Available: https://github.com/microsoft/DNS-Challenge

[3] [Online]. Available: https://aka.ms/dns-challenge

TABLE 1. Remaining Headroom in MOS Improvements Needed to Attain Best Speech Quality (MOS 5) is Applicable to Both Personalized and Non-Personalized DNS, as Determined From the ICASSP 2022 DNS Challenge [23]

DNS mode	Improvements area	Headroom
Personalized	SIG	0.81
Personalized	BAK	0.45
Personalized	OVRL	1.03
Non-personalized	SIG	0.70
Non-personalized	BAK	0.30
Non-personalized	OVRL	0.87

TABLE 2. Comparison of Some DNN NI-SQA Methods

Model	Data size (hours)	Data type
[37]	5.2	ACR
WAWENETS [36]	151	ACR
[32]	27.7	ACR
[34]	27.7	ACR
SESQA [40]	45.2	ACR, JND
DNSMOS [13]	300	ACR
DNSMOS P.835	75	P.835 ACR

are either instructed to hold a conversation over a telecommunication system under study (conversation test) or listen to short speech clips (listening opinion tests) and afterward rate perceived quality on one or several rating scales. Speech calls can be carried out with various devices in different environments, commonly with non-optimal acoustic surroundings. Therefore, speech enhancement algorithms are widely integrated into the communication chain to enhance the quality of the speech communication system. Those systems are typically evaluated in laboratory-based listening tests according to the ITU-T Rec. P.835 [24] in which separate rating scales are used to independently estimate the BAK, SIG, and OVRL. Separate scales are used as higher noise suppression often adversely affects the speech or the signal component, resulting in distortions or artifacts. Consequently, in a regular listening-only test, with a single-rating scale (i.e., according to the ITU-T Rec. P.800 [25]), participants can often become confused as to what they should consider in rating the overall "quality." Accordingly, each individual determines their overall quality rating by weighting the signal and background components. Such a process introduces additional errors in the overall quality ratings and reduces their reliability [24].

The ITU-T Rec, P.808 [26] details how to perform subjective speech quality test in crowdsourcing. Crowdsourcing offers a faster, cheaper, and more scalable approach than traditional laboratory tests [27]. Crowdsourcing does have its challenges: the test participants take part in the test in their working environment using their own hardware without supervision or direct quality control. Previous works showed that background noise in the participant's surroundings could mask the degradation under the test and lead to a significantly different rating [28], [29]. Different listening devices can also strongly influence the perceived quality [30]. The ITU-T Rec, P.808 [26] addresses these challenges and provides methods to collect reliable and valid data in crowdsourcing practice. The recommendation addresses different test methods including the Absolute Category Rating (ACR) test method, Comparison Category Rating (CCR), and has recently been extended to provide test methods for evaluating noise suppression algorithms in crowdsourcing (i.e., the counterpart of P.835).

As we use real test clips in the challenge, we must use a non-intrusive objective method. There are many non-intrusive objective metrics for noise suppression. ITU-T Recommendation P.563 is a non-intrusive technique and can directly operate on the degraded signal [31]. However, it was developed for narrow-band applications, and works on limited impairment types, but correlates poorly with human ratings [32]. More recently, DNN-based approaches have been proposed to estimate the speech quality scores [13], [14], [32], [33], [34], [35], [36], [37], [38]. Some of these learning-based approaches use other objective metrics as the ground truth to train their speech quality predictor. Other methods use MOS obtained using P.800 as the ground truth to train their models. In [39], the authors trained the model to identify the Just Noticeable Difference (JND). DNN-based MOS predictors learning a mapping between audio and human ratings have shown better performance than other objective metrics like PESQ or POLQA [13]. The accuracy and robustness of the learned models depend on the quality of the human labels and the quantity and diversity of the audio clips. A comparison of some common DNN-based non-intrusive speech quality assessment (NI-SQA) methods is given in Table 2.

The first DNS challenge was at INTERSPEECH 2020 [20] and had real-time and non-real-time tracks. It included a clean speech dataset of 441 hours with 2150 speakers, a noise dataset of 150 classes with 60K clips, and a synthetic data generator. Two real test sets were included, and submissions were judged using our P.808 implementation [15]. The second DNS challenge was at ICASSP 2021 [12] and expanded the datasets by adding singing, emotion, and non-English languages. It replaced the non-realtime track with a personalized deep noise suppression track. The third DNS challenge was at INTERSPEECH 2021 [21] and used P.835 [16] instead of P.808, added the objective speech quality assessment tool DNSMOS [13], and included real-time wideband and fullband tracks. The fourth DNS challenge was at ICASSP 2022 and added mobile data and word accuracy as an objective metric, included real-time fullband and real-time personalizing tracks, and added the DNSMOS P.835 objective tool. A summary of the five DNS challenges is given in Table 3.

While this challenge focuses on improving speech quality by reducing background noise (improving BAK) and reverberation which improves OVRL, other related challenges target echo cancellation [41], [42], [43], [44], packet loss concealment [45], and general speech signal improvements [46]. The ICASSP 2023 Speech Signal Improvement challenge [46] provides test sets with various types of SIG

TABLE 3. Summary of DNS Challenges

Challenge	Tracks	Training set	Algorithmic + Buffering Latency	Notes
INTERSPEECH 2020	Real-time Non-real-time	Clean: 441 hours, 2150 speakers Noise: 150 classes, 60K clips	40 ms	P.808
ICASSP 2021	Real-time Real-time personalized	Clean: 761 hours Noise: 150 classes, 60K clips	40 ms	Singing, emotion, non-English languages
INTERSPEECH 2021	Real-time wideband Real-time fullband	Clean: 761 hours Noise: 181 hours	40 ms	P.835, DNSMOS
ICASSP 2022	Real-time Real-time personalized	Clean: 761 hours Noise: 181 hours	40 ms	Mobile, WAcc, fullband PDNSMOS P.835
ICASSP 2023	Real-time personalized speakerphone Real-time personalized headset	Clean: 761 hours Noise: 181 hours	20 ms	

regressions such as poor-quality microphones and speech enhancement. In particular, the past five DNS challenges did not improve SIG, whereas the Speech Signal Improvement challenge significantly improved SIG. That challenge used a new multidimensional approach to measuring speech quality described in [47].

III. CHALLENGE TRACKS

The overarching objective of this challenge is to improve the overall signal quality while preserving the primary talker's voice and concurrently suppressing noise, reverberation and neighboring talkers. We assess this objective for two device conditions: (i) the user wearing a headset device, i.e., having a microphone close to the user's mouth, or (ii) a farfield scenario, where the device is not directly with the user, e.g., a speakerphone or laptop, therefore having potentially larger source-to-microphone distances. The idea of this division is that possibly in the headphone scenario, the acoustics give a more clear distinction between primary talkers and interfering talkers, therefore making the need for enrollment speech obsolete. The farfield case may have less clear acoustical distinction between closer primary talkers and interfering talkers.

Therefore, this challenge is divided into two tracks: Headset and Speakerphone. Each track has distinct development and blind test sets. These test sets were gathered following a similar procedure, with the key difference being that the Headset track test sets were collected using headset devices, while the Speakerphone track test sets were collected using speakerphone devices.

In both tracks, every test clip is accompanied by an enrollment speech lasting 30 seconds. This enrollment speech can exhibit variations, including being noise-free or noisy, and with or without reverberation. This setup supports the incorporation of multi-condition enrollment for primary talkers, which serves as a metric of robustness for personalized models. These personalized models utilize enrollment speech as an additional input to enhance the test clips.

Participants had the flexibility to choose whether to work on models involving speaker enrollment or models without it for one or both tracks. Each team was allowed to submit 1 to 4 models, depending on their training strategies. For instance, a participating team could submit one personalized model and one non-personalized model for the Headset track, but they could not submit two personalized or two non-personalized models for the same track. Similarly, another team could submit a total of 4 models: a personalized and a non-personalized model for the Headset track, and the same for the Speakerphone track. This rule was established to ensure a balanced representation of personalized and non-personalized models and to encourage comparable participation in both tracks.

In both tracks, all submitted models were evaluated and ranked collectively. This means that both personalized and non-personalized models for the Headset track underwent the same subjective evaluation. Similarly, for the Speakerphone track, all models were evaluated together in one subjective evaluation. While participants were encouraged to conduct experiments with both personalized and non-personalized models to uncover the advantages of personalization, it is worth noting that this was not a mandatory requirement for the challenge.

Requirements: To ensure the real-time operation of these models on typical hardware available today, the processing mode of the models must satisfy the following constraints on the overall introduced latency being equal to or lower than 20 ms. We define and give examples for algorithmic and buffering latency on the challenge website.[4] The real-time factor, measured as execution time on an Intel Core i5 Quadcore clocked at 2.4 GHz using single threading, must be less than 0.5. Additionally, participants are asked to report the number of multiply-accumulate operations of their models.

IV. CHALLENGE DATASETS

All datasets utilized in this challenge are full bandwidth (48 kHz). In this section, we discuss details of the training, development, and test sets. To allow supervised training, the training data is synthesized from clean speech, room impulse responses, and noise, while the development and test sets are real recordings to ensure real-world generalization.

[4][Online]. Available: https://aka.ms/dns-challenge

A. TRAINING DATA

The clean speech training set is a total of 760.53 hours of data. The speech data encompasses various languages, showcasing a diversity of talkers and devices. The clean speech data is further categorized into four subsets:

- Read speech recorded under clean conditions (562.72 hours)
- Singing voice (8.80 hours)
- Emotional clean speech (3.6 hours)
- Non-English (German, French, Italian, Mandarin, Russian, Spanish) clean speech (185.41 hours)

The clean speech data for the Headset track is derived as a subset of only non-reverberant speech, as we are targeting headset scenarios, where there is little to no reverberation present. We employed a DNN-based predictor, a modified version of [48], to detect clips as having a high direct-to-reverberation ratio above 40 dB, i.e., assumed to be near-field recorded or non-reverberant speech. This subset is subsequently released as the clean speech dataset for the headset track.

For the Speakerphone track, we utilized the entire clean speech dataset from the fourth DNS Challenge.

To assist in the development of personalized models for both tracks, we provided speaker ID information for all clean speech clips in the training set. Furthermore, we released code for extracting speaker embeddings based on state-of-the-art ECAPA-TDNN embeddings, trained on the VoxCeleb dataset [49], [50], [51].

The noise dataset and impulse responses utilized in this challenge remain consistent with those used in the fourth DNS Challenge [23] and is described in the following. The noise data integrated into the training set is selected from Audio Set [52], and it mirrors the noise set used in the fourth DNS Challenge [23]. Audio Set comprises roughly 2 million human-labeled 10-second sound clips extracted from YouTube videos. Within Audio Set, more than a million clips encompass audio classes like music and speech, while classes such as toothbrush or creak are represented by fewer than 200 clips. Around 42% of the clips are associated with a single class, but the remainder might carry 2 to 15 labels. To rectify this imbalance, we devised a sampling strategy to ensure that each class includes a minimum of 500 clips.

To remove any speech from Audio Set noise data, we utilized a speech activity detector to eliminate clips with any form of speech activity. These clips were sourced from Audio Set and were initially available at a 44.1 kHz sample rate, which we subsequently upsampled to fullband (48 kHz). Consequently, the resulting noise dataset encompasses 152 audio classes and 60,000 clips [23]. Altogether, the noise training data contains a cumulative noise data duration of 181 hours.

As in previous challenges, the room impulse responses (RIRs) are from several data sets; 248 are real and about 60,000 are simulated RIRs from the openSLR26 and openSLR28 [53] datasets. These RIRs can be used to generate reverberant speech data. We provide a training data synthesizer that convolves speech with RIRs and adds noise depending on the chosen configuration.

B. DEVELOPMENT TEST SET

Both test sets consist of fullband audio clips recorded in real-world scenarios, obtained through crowdsourcing. Workers read provided text prompts and record their voices using desktops, laptops, or mobile devices while contending with ambient noise and/or neighboring talkers. It should be noted that no ground truth clean speech data is available for the test sets.

The development test set for the Headset track contained 641 real test clips recorded using a variety of headset devices. Similarly, the Speakerphone track development set has 600 real test clips recorded on speakerphone devices. This helped challenge participants to conduct an intermediate evaluation of their models.

In this challenge, the final ranking of submitted models was solely done based on subjective and WAcc evaluation of the blind set. Thus, the development test set and DNSMOS score were only to be used for aiding the model development.

C. BLIND TEST SET

We have introduced new noise types into the test set, covering a range of pertinent real-world scenarios, device variations, and the addition of a paralinguistic test set as a novel category. The blind test set encompasses genuine test clips that have not been previously utilized in any challenge and are not otherwise available to the public. Our test set comprises real-world test clips recorded by crowdsourced workers.

We executed a stringent quality assurance process to ensure that the blind set accurately mirrors real-world scenarios. This encompassed a diversity of speaker profiles, device types, various acoustic situations, different direct-to-reverberation ratios (DRR), varying T60 times achieved through stratifying the collected samples which varied the relative and absolute positions of primary and interfering talkers, the presence of noise sources, and the inclusion of reflecting surfaces.

The paralinguistic test clips encompass standard forms of paralanguage,[5] including but not limited to the throat-clear, “hmm” or “mhm”, “Huh?” or “what?”, gasps, sighs, moans, groans, deceptive speech, sincere speech, bass-heavy speech, speech with high and low pitch, confident, tired, persuasive speech, and voice change mid-clip (i.e., imitating someone else’s voice in the last 50% of the clip). Emotional speech includes but is not limited to happiness, sadness, anger, yelling, crying, and laughter. Furthermore, the blind test set comprises acoustic conditions characterized by:

- High reverberation
- High reverberation with noise
- Noise in the presence of interfering talkers

There was a variety of noise types in the test set. Fig. 2 shows the distribution of noise types.

In total, the Headset track has 389 real test clips out of which 220 have interfering talkers, and 51 are leakage clips.

[5] [Online]. Available: https://en.wikipedia.org/wiki/Paralanguage

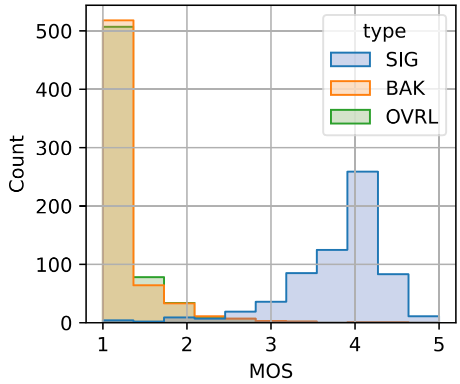

FIGURE. 1. Visualization depicting the distribution of subjective scores (P.835 MOS) for unprocessed clips within the blind test set.

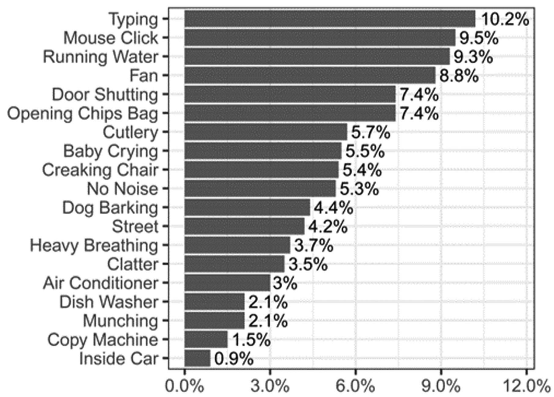

FIGURE. 2. Distribution of noise types within the blind test set.

Leakage clips have a duration of 6 minutes and were designed to identify personalized models that may forget the primary talker when the primary talker goes on a long pause. The test set was recorded using a variety of wired and wireless devices. The blind set was collected by four crowdsourcing data vendors with the following distributions: Vendor_1 (222 test clips), Vendor_2 (77 test clips), Vendor_3(39 test clips), Vendor_4 (51 test clips).

The blind set for the Speakerphone track has 331 real test clips out of which 220 have interfering talkers, and 51 are leakage clips. The leakage clips were common and identical in both tracks. These clips were recorded using speakerphone devices. The Speakerphone track blind set was collected by four vendors with the following distributions: Vendor_1 (221 test clips); Vendor_2 (59 test clips); and Vendor_4 (51 clips).

Fig. 1 shows the distribution of MOS obtained from subjective evaluation. It shows more variety in SIG values and fewer variations in BAK and OVRL for blind sets. Most noisy clips have low values (<2) for BAK and OVRL, thus confirming a challenging test set.

V. EVALUATION SETUP

A. BASELINE MODELS

Instead of providing a baseline model, we provided the enhanced blind set clips for both tracks. The personalized and non-personalized baseline models were derived from the architecture proposed in [54], adhering to the RTF and lookahead constraint challenge rules.

B. SUBJECTIVE EVALUATION

This section describes the preparation of enrollment clips, the conduction of the subjective listening tests, and the reproducibility study. The subjective evaluation utilized only a 5-second segment of enrollment speech.

1) CLEANING ENROLLMENT CLIPS

We manually selected the 5-second segment from the enhanced enrollment clip, ensuring the removal of all long pauses (>0.2 s). The resulting 5-second enrollment audio was loudness normalized to facilitate easier recognition of the primary talker by human raters. Additionally, enrollment clips were enhanced using a non-causal non-personalized DNS model based on the end-to-end enhancement network (E3Net) architecture [55]. Instead of using the Short-Time Fourier Transform (STFT) and its inverse (iSTFT), this enhancement model employs learnable encoders and decoders, which helps mitigate the issue of imperfect phase reconstruction commonly encountered in most time-frequency-based speech enhancement methods. The performance was improved by converting the causal model to non-causal processing by using a bidirectional Long Short-Term Memory (LSTM) block.

2) PERSONALIZED P.835 FRAMEWORK

This challenge relies on the P.808 Toolkit [16], which is an implementation of the ITU-T Rec. P.808 [26], and its test method for subjective evaluation of noise suppression algorithms (i.e. crowdsourcing counterpart of P.835). In this challenge, we designed a novel personalized version for P.835 test method, referred to as personalized P.835. The personalized P.835 subjective framework collects three MOS scores for each clip: SIG, BAK, and OVRL. In this approach, for each test clip, 5 seconds of clean enrollment speech are presented to test participants to identify the *target speaker* (see Fig. 4). This facilitated the human raters to identify the primary talker's voice, aiding in the assignment of subjective scores for the *judgment segment* of length 7 seconds accordingly.

In a prior training session, raters were instructed to concentrate on the quality of the primary speaker's voice when assessing the speech quality in the judgment segment (which may contain neighbor speakers). In addition, the following specific instructions were provided to workers: If the target speaker is completely removed from the judgment section and another person is present, rate BAK = 1 (presence of high background noise), SIG = 1 (removal of target speaker), and OVRL = 1 (poor performance of the model). During the training session a feedback with expected answer and explanations was provided to the participants. We also added two new qualification tests to ensure 1) participants are able to identify different talkers and 2) their device can playback audio in fullband (see Fig. 3). Additionally, we improved the reliability checks by incorporating gold clips designed for personalized

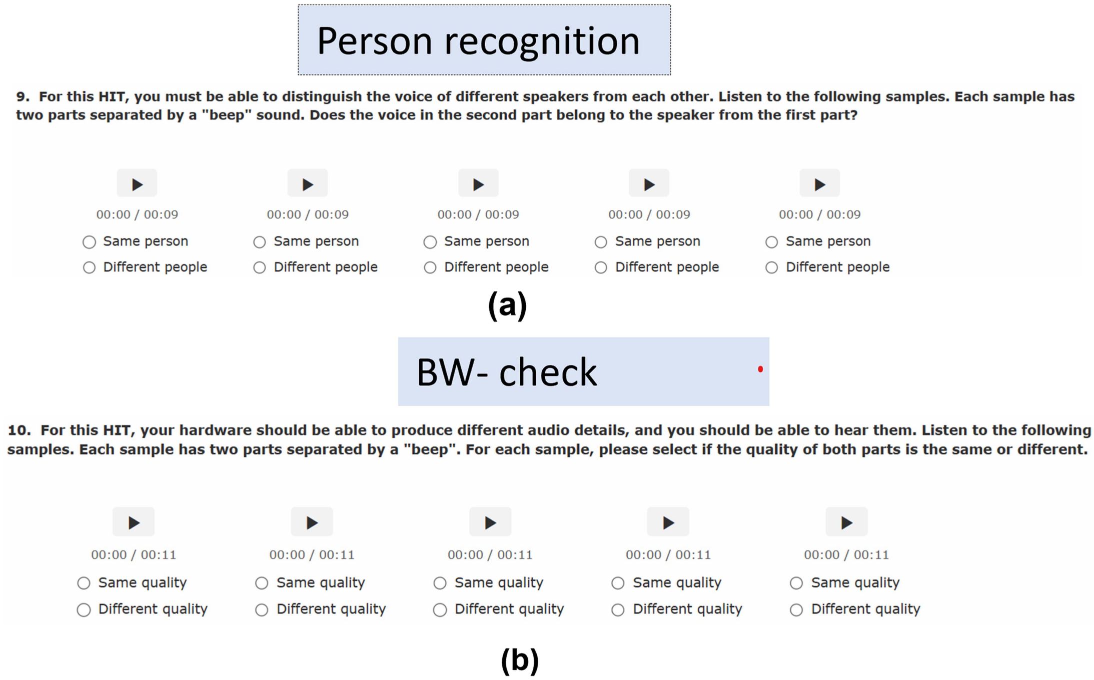

FIGURE. 3. User interface showing extended training and qualification tests in Personalized P.835 to ensure (a) person recognition, (b) Bandwidth check.

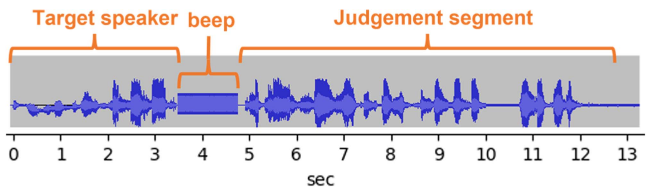

FIGURE. 4. Personalized P.835 test sound clip structure.

TABLE 4. PCC Between 4 Reproducibility Runs in Amazon Mechanical Turk.

	PCC
SIG	0.993
BAK	0.917
OVRL	0.989

Each run has 10 models including noisy (approximately 300 clips).

CMP	PCC	SRCC	Tau-B	Tau-B_95
MOS_BAK - run0_run1	0.993	0.976	0.929	0.783
MOS_BAK - run0_run2	0.997	0.929	0.857	0.837
MOS_BAK - run0_run3	0.997	0.952	0.857	0.74
MOS_BAK - run1_run2	0.989	0.818	0.689	0.558
MOS_BAK - run1_run3	0.989	0.709	0.556	0.788
MOS_BAK - run2_run3	0.995	0.903	0.778	0.796
MOS_SIG - run0_run1	0.961	0.952	0.857	0.917
MOS_SIG - run0_run2	0.917	0.786	0.643	0.81
MOS_SIG - run0_run3	0.907	0.881	0.714	0.816
MOS_SIG - run1_run2	0.898	0.818	0.644	0.87
MOS_SIG - run1_run3	0.861	0.758	0.556	0.795
MOS_SIG - run2_run3	0.959	0.806	0.644	0.884
MOS_OVRL - run0_run1	0.997	1	1	0.874
MOS_OVRL - run0_run2	0.991	0.833	0.714	0.804
MOS_OVRL - run0_run3	0.995	0.881	0.714	0.772
MOS_OVRL - run1_run2	0.98	0.673	0.556	0.637
MOS_OVRL - run1_run3	0.978	0.709	0.556	0.758
MOS_OVRL - run2_run3	0.994	0.782	0.644	0.812

FIGURE. 5. Comparison of Pearson Correlation Coefficient (PCC), SRR, Tau-B and Tau-B_95 between 4 runs of Amazon Mechanical Turk reproducibility study.

scenario. We also use other reliability check methods provided in the P.808 Toolkit [15].

3) PERSONALIZED P.835 REPRODUCIBILITY STUDY

Fig. 5 shows the comparison of Pearson correlation coefficient (PCC), Spearman's rank correlation coefficient (SRCC), Kendall's Tau-B and Tau-B_95 [56] between 4 separate runs for a reproducibility study on Amazon Mechanical Turk in model level. Specifically, we did a different run on four different days, each with different raters. The dataset used in the test included 10 models applied on 300 clips. On average we have collected 5 to 6 ratings for each clip in each run. Table 4 shows the PCC between MOS scores from these reproducibility runs. The high PCC and SRCC show that the personalized P.835 framework is reproducible.

C. WORD ACCURACY

We estimated WAcc using the Microsoft Teams endpoint speech recognition system. This WAcc computation process was carried out by the organizers during the final week of the challenge, ensuring that all models were assessed using the same methodology. WAcc serves as an objective metric for evaluating the impact of speech enhancement on speech transcription services. The formula defining WAcc is as follows:

$$\mathrm{WAcc} = 1 - \mathrm{WER}, \tag{1}$$

TABLE 5. PCC Between WAcc and BAK, SIG, and OVRL for [23] Track 1

	PCC
BAK	0.150
SIG	0.647
OVRL	0.526

where WER represents the word error rate of the speech recognition system compared to the transcribed speech. As the ground truth transcripts for WAcc include only words spoken by the primary talker, recognized words from interfering talkers degrade WAcc, which therefore acts as a very sensitive metric to neighbor talker leakage.

To derive the ground truth, we transcribed the complete blind set for both tracks. The development test set was not transcribed. Distinct from the subjective P.835 framework that employs a manually selected 7-second segment from either noisy or enhanced clips, the WAcc is computed for the entire length of the test clips. As mentioned in Section IV-B, the test clips within the blind set vary in duration, ranging from 10 seconds to over 6 minutes.

The transcriptions of the blind test set were gathered through crowdsourced data collection. Workers were provided with text prompts to read from, though these prompts did not necessarily reflect the exact transcriptions due to reading errors, word omissions, and other variations. To establish accurate ground truth transcriptions for the blind test set, we followed a five-step methodology:

1) *Prompt Collection:* In the initial step, we gathered the text prompts corresponding to each test clip within the blind set.
2) *Speech Recognition Transcription:* Using a state-of-the-art speech recognition engine Whisper [57], we obtained transcriptions for each test clip in the blind set.
3) *Human Listener Transcription:* Expert human listeners then carefully listened to each test clip and generated corresponding human-generated transcripts. These expert listeners were instructed to listen to the audio clips multiple times until they were confident in their transcription.
4) *Word Error Rate Computation:* We calculated the word error rate (WER) for each test clip in the blind set. Clips with WER > 0.5 were identified for further review.
5) *Correction and Validation:* For the test clips with a WER > 0.5, a fifth round of listening took place. Human listeners re-listened to these clips and validated or corrected the human-generated transcriptions. It is noteworthy that only a small number of clips required correction during this stage, underscoring the robustness of our transcription process.

Clips that were untranscribable even after this five-step approach were consequently discarded.

The correlation of WAcc and BAK, SIG, and OVRL for the ICASPP 2022 DNS Challenge Track 1 [23] is given in Table 5, which shows that WAcc is most correlated with SIG and least with BAK. Therefore, to improve WAcc it would be most effective to improve SIG.

D. CHALLENGE METRIC

In alignment with previous challenges, the models in both tracks were ranked using a final score derived from an average of personalized ITU-T P.835 OVRL and WAcc. The inclusion of WAcc serves as an objective metric quantifying the impact of speech enhancement on automatic speech recognition-based transcription services. OVRL and WAcc were used on the blind test set to rank the models using the below formula:

$$\text{Score} = 0.5[\text{WAcc} + 0.25(\text{OVRL} - 1)] \tag{2}$$

E. CHALLENGE MODE

At the challenge start, the training and development sets were released. During the development phase, participants could submit enhanced clips generated by their models on the development set, which were evaluated by the organizers and shared with the participating teams. This comprehensive approach ensured a thorough assessment of the submitted models.

VI. RESULTS & ANALYSIS

A. SUBMISSIONS AND RESULTS

Both tracks attracted significant participation, each drawing in 11 submissions. 10 teams participated in both tracks. The models submitted across both tracks predominantly consisted of personalized models.

In the Headset track, there were two baseline models, one of which was non-personalized (*Baseline_nonp*). In contrast, the Speakerphone track's baseline model was personalized (*Baseline_p*). The non-personalized baseline in the Headset track was trained on extracting only nearfield speech utilizing the headset scenario acoustic conditions to blend out farfield speakers, which are more reverberant. This removes the need for using enrollment data. For reference, we also include two newer internal personalized models, denoted by *MSFT-1* and *MSFT-2*, which also conform with the challenge rules, but do not participate in the challenge. Details about those models may be released in future papers.

Fig. 6 shows the main results of the challenge in terms of subjective personalized P.835 scores, WAcc, and the overall challenge score (2) for all participating teams. The order of teams is arranged in descending order of the score. In this context, *dMOS* indicates the difference in SIG, BAK, and OVRL between the enhanced clip and the corresponding noisy clip. Similarly, *dWAcc* represents the difference in WAcc between the enhanced clip and the noisy clip.

B. COMPARISON OF CHALLENGE MODELS

Table 6 shows the descriptions and additional information on the models we have obtained from the top five teams. The table shows RTF, training data and its size, number of training stages, input type, speaker embedding model and

TABLE 6. Comparison of the Top Five Teams in ICASSP 2023 DNS Challenge. N/A Means Data Not Available Yet From Participants

Team Name	UnbeatableTencent [58]	NAPSE [59]	TSpeech-AI [60], TencentASSP, TencentVPPaaS	NJUAALab2 [61]	SZAudio [62]
Rank	HS: 1 SP: 1	HS: 1 SP: 2	HS: 3 SP: 3	HS: 4 SP: 4	HS: 4 SP: 4
Params	22.2 M	12.49 M	N/A	1.97 M (generator size)	5.97 M
Real-time factor	0.46	0.48	0.42	0.49	0.41
Training data	DNS5	DNS5 + DNS4 track2	DNS5	DNS5	DNS5 + Didi-speech [63]
Training data size	generated on-the-fly	2,000 hours	generated on-the-fly	800 hours of speech and 200 hours of noise	100,000 4 s clips generated on-the-fly noisy
Training Stages	3	3	3	1	1
Domain	STFT	time-domain	time-domain	STFT	STFT
Speaker Embedding	ECAPA-TDNN [50], [64]	RestNet34 [65]	RestNet34 [65]	None	ECAPA-TDNN [50], [64]
Description	Two stage TEA-PSE+ LSTM+LGR+Multi-STFT→re-train.	two stage TEA-PSE 2.0→MetricGAN.	Band-split RNN [63] + multi-resolution STFT discriminator.	Spectral dimension compression + CRNN encoder + MetricGAN.	Speaker attentive module + band-split RNN.
Challenge Score Delta	HS: 0.60 SP: 0.60	HS: 0.59 SP: 0.58	HS: 0.57 SP: 0.57	HS: 0.55 SP: 0.55	HS: 0.53 SP: 0.55

Headset track

Model	Signal MOS	Signal dMOS	Background MOS	Background dMOS	Overall MOS	Overall dMOS	CI	Wacc	dWAcc	Score
MSFT-1*	3.56	-0.19	3.13	1.91	2.90	1.68	0.08	0.770	-0.073	0.623
MSFT-2*	3.57	-0.19	3.14	1.92	2.92	1.70	0.08	0.762	-0.081	0.621
UnbeatableTencent	**3.52**	**-0.24**	**2.88**	**1.66**	**2.71**	**1.48**	**0.08**	**0.761**	**-0.082**	**0.594**
NAPSE	**3.58**	**-0.17**	**2.87**	**1.66**	**2.69**	**1.47**	**0.08**	**0.758**	**-0.085**	**0.590**
Tspeech	**3.58**	**-0.18**	**2.82**	**1.61**	**2.65**	**1.43**	**0.08**	**0.725**	**-0.118**	**0.569**
TencentASSP	**3.52**	**-0.23**	**2.81**	**1.60**	**2.60**	**1.38**	**0.08**	**0.734**	**-0.109**	**0.567**
NJUAALab2	**3.48**	**-0.28**	**2.67**	**1.45**	**2.48**	**1.26**	**0.08**	**0.725**	**-0.118**	**0.548**
SZAudio	**3.51**	**-0.25**	**2.48**	**1.26**	**2.30**	**1.08**	**0.07**	**0.733**	**-0.110**	**0.529**
TencentVPPaaS	**3.44**	**-0.32**	**2.74**	**1.53**	**2.48**	**1.26**	**0.08**	**0.683**	**-0.160**	**0.527**
HITIoT	3.14	-0.62	2.68	1.46	2.34	1.12	0.08	0.713	-0.130	0.524
Baseline_nonp	3.14	-0.61	2.60	1.38	2.34	1.12	0.07	0.707	-0.136	0.521
Baseline_p	3.20	-0.56	2.67	1.46	2.34	1.12	0.08	0.687	-0.156	0.511
NJUAALab1	3.44	-0.32	2.13	0.92	2.00	0.77	0.07	0.716	-0.127	0.483
Doreso	3.29	-0.47	2.05	0.83	1.90	0.68	0.07	0.702	-0.141	0.464
noisy	3.76	0.00	1.22	0.00	1.22	0.00	0.03	0.843	0.000	0.449
CQUPT_Liu	3.16	-0.60	1.67	0.45	1.49	0.26	0.05	0.676	-0.167	0.399

*Not included in challenge ranking. Microsoft internal model which satisfies challenge requirements.

(a)

Speakerphone track

Model	Signal MOS	Signal dMOS	Background MOS	Background dMOS	Overall MOS	Overall dMOS	CI	Wacc	dWAcc	Score
MSFT-1*	3.70	-0.13	3.20	1.98	2.97	1.74	0.08	0.778	-0.079	0.636
MSFT-2*	3.71	-0.12	3.24	2.02	3.00	1.77	0.08	0.767	-0.090	0.634
UnbeatableTencent	**3.64**	-0.19	**2.92**	1.70	**2.72**	1.49	**0.08**	**0.768**	**-0.089**	**0.600**
NAPSE	**3.60**	-0.23	**2.78**	1.55	**2.58**	1.34	**0.08**	**0.769**	**-0.088**	**0.581**
TencentASSP	**3.60**	-0.24	**2.84**	1.62	**2.65**	1.41	**0.08**	**0.735**	**-0.122**	**0.573**
Tspeech	**3.64**	-0.19	**2.88**	1.65	**2.66**	1.42	**0.08**	**0.724**	**-0.133**	**0.570**
TencentVPPaaS	**3.61**	-0.22	**2.86**	1.64	**2.64**	1.40	**0.08**	**0.728**	**-0.129**	**0.569**
NJUAALab2	**3.54**	-0.30	**2.77**	1.55	**2.54**	1.30	**0.08**	**0.734**	**-0.123**	**0.559**
SZAudio	**3.53**	-0.30	**2.58**	1.36	**2.39**	1.15	**0.08**	**0.749**	**-0.108**	**0.548**
Baseline_p	3.22	-0.62	2.68	1.46	2.38	1.14	0.07	0.727	-0.130	0.536
Team29_HITIoT_p_Pref1	3.07	-0.76	2.69	1.47	2.26	1.02	0.07	0.731	-0.126	0.523
Team8_NJUAALab1_p_Pref1	3.44	-0.39	2.28	1.06	2.10	0.86	0.07	0.731	-0.126	0.503
Team2_Doreso_p_Pref1	3.26	-0.58	2.11	0.89	1.93	0.69	0.07	0.727	-0.130	0.480
Team33_CZUR_nonp_Pref1	3.59	-0.24	1.88	0.66	1.76	0.53	0.06	0.735	-0.122	0.463
noisy	3.83	0.00	1.22	0.00	1.24	0.00	0.03	0.857	0	0.458

*Not included in challenge ranking. Microsoft internal model which satisfies challenge requirements.

(b)

FIGURE. 6. Results from the personalized P.835 subjective evaluation, WAcc, and the Challenge metric (Score) were computed on the blind test set for all teams in both tracks: (a) Headset, (b) Speakerphone.

Headset Track	UnbeatableTencent	NAPSE	Tspeech	TencentASSP	NJUAALab2	SZAudio	TencentVPPaaS
UnbeatableTencent		0.644	0.128	0.015	0	0	0
NAPSE	0.644		0.288	0.048	0	0	0
Tspeech	0.128	0.288		0.36	0	0	0
TencentASSP	0.015	0.048	0.36		0.008	0	0.005
NJUAALab2	0	0	0	0.008		0	0.858
SZAudio	0	0	0	0	0		0
TencentVPPaaS	0	0	0	0.005	0.858	0	

(a)

Speakerphone Track	UnbeatableTencent	NAPSE	TencentASSP	Tspeech	TencentVPPaaS	NJUAALab2	SZAudio
UnbeatableTencent		0.001	0.07	0.122	0.056	0	0
NAPSE	0.001		0.114	0.064	0.134	0.439	0
TencentASSP	0.07	0.114		0.788	0.93	0.019	0
Tspeech	0.122	0.064	0.788		0.719	0.009	0
TencentVPPaaS	0.056	0.134	0.93	0.719		0.023	0
NJUAALab2	0	0.439	0.019	0.009	0.023		0.001
SZAudio	0	0	0	0	0	0.001	

(b)

FIGURE. 7. Paired t-test for (a) Headset track; (b) Speakerphone track to test the statistical difference between the top 5 models.

overall challenge score. All models have similar RTF and most of them used the challenge dataset. The top two models have a significantly higher number of parameters than others, which may indicate some correlation. We however do not see a strong correlation between RTF and ranking. The top three models trained in several stages which suggests that this may help achieve better performance. We have a mixed bag of STFT and time-domain models. Four out of five models are personalized models. The best and worst models are based on ECAPA-TDNN speaker embeddings which were provided as baseline speaker embedding for this challenge while the second and third-rank teams used ResNet34 for speaker modeling.

Fig. 7 shows paired t-tests for both tracks. It shows that in the Headset track, no statistical difference was observed between performance of the top three models (orange). For the Speakerphone track, the best two models are statistically

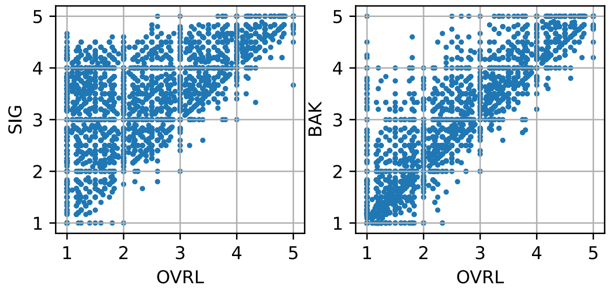

FIGURE. 8. Visualized distribution of clip-level subjective scores for SIG, BAK, and OVRL for all models including challenge participants and Microsoft.

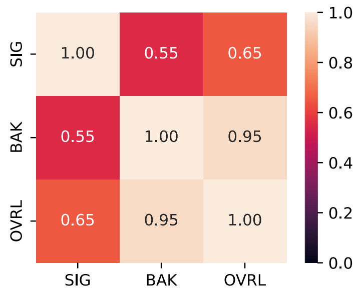

FIGURE. 9. Heatmap of Pearson correlation between SIG, BAK, and OVRL from Personalized P.835 subjective evaluation. This includes all challenge models and internal Microsoft models.

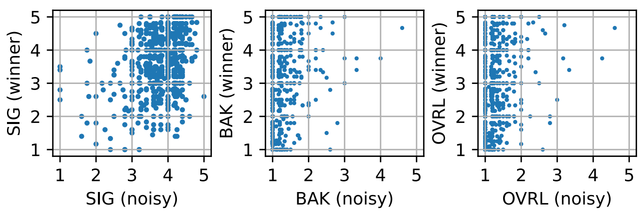

FIGURE. 10. Visualization of subjective ratings (P.835 MOS) distribution for the noisy blind test set (*X*-axis) and the challenge winner model (*Y*-axis).

different while teams TencentASSP, TSpeech-AI, and TencentVPPaaS are statistically similar.

C. SCORE DISTRIBUTIONS

Fig. 8 shows the distribution of SIG vs OVRL and BAK vs OVRL for all models including challenge entries and Microsoft internal models. Each point in Fig. 8 corresponds to one MOS-rated clip. Both graphs show a positive correlation between SIG and OVRL, and BAK and OVRL. Interestingly, we observe a strong trend that almost always we obtain ratings with SIG $\geq$ OVRL. BAK and OVRL seem to have a more linear relation than SIG and OVRL. This is further analyzed in Fig. 9, showing the Pearson correlation between SIG, BAK, and OVRL scores for all models including challenge entries and Microsoft internal models. We can observe that BAK and OVRL have a very strong correlation of 0.95, whereas other pairs have only a moderate correlation.

Fig. 10 shows the subjective score data points of the top model on the Y axis compared to the corresponding score of noisy on the X axis. We can see that the noisy BAK and OVRL values, which reside largely in the lower MOS regions (see also Fig. 1) are getting shifted up by the winner's processing system and become distributed over the whole MOS range. This means that BAK and OVRL are largely improved on average. However, for SIG, the already higher SIG distribution is not significantly shifted up. On the contrary, we can observe severe SIG degradations for a significant portion of the clips, which results in minor degradation on the mean score as shown in Fig. 6.

TABLE 7. Remaining Headroom in MOS Improvements Needed to Attain Excellent Speech Quality (MOS 5) for Both Tracks, as Determined From the ICASSP 2023 DNS Challenge

DNS mode	Improvements area	Headroom
Headset	SIG	1.48
Headset	BAK	2.12
Headset	OVRL	2.29
Speakerphone	SIG	1.36
Speakerphone	BAK	2.08
Speakerphone	OVRL	2.28

D. COMPARISON TO PREVIOUS CHALLENGES

Table 7 shows the remaining headroom in SIG, BAK, and OVRL for headsets and speakerphones. The headroom for all metrics here are significantly larger than those in Table 1 for the previous DNS challenge at ICASSP 2022. Possible reasons for this are: (i) the test set is more challenging, and (ii) the addition of the personalization task, i.e., speaker identity-informed target speaker extraction makes the problem more challenging.

VII. LIMITATIONS

The primary limitation of this challenge is the potential lack of representative samples in the test set. An ideal methodology would be to sample audio clips from a real-time communication system in production and stratify these clips to cover all significant scenarios. However, doing so would have many privacy issues both in content discussed as well as biometric identity. The scenarios included in this challenge are the top ones we see in Microsoft Teams and Microsoft Skype, which may not be representative to all real-time communication systems.

VIII. CONCLUSION

This paper describes the fifth incarnation of the Deep Noise Suppression Challenge, evaluating state-of-the-art DNN systems on the task of personalized speech enhancement. The challenge rules set constraints on model runtime and look-ahead, enforcing models that can practically be used for on-device real-time communication pipelines. The test sets and evaluation metrics are designed to generalize in the best possible way to realistic performance by evaluating real recordings, collected from a variety of devices and acoustic settings, including paralinguistic and leakage test clips, and using directly relevant metrics such as subjective human MOS ratings and automatic speech recognition performance.

The top models improve overall quality and suppress background noise and interfering talkers impressively, however at the cost of degrading SIG compared to the unprocessed signal. Additionally, personalized models are more prone to inadvertently suppress the primary talker's speech due to confusion with interfering speech, which creates more speech distortions and degrades robustness significantly for practical use.

Looking forward, there are exciting emerging research areas in speech enhancement and processing. One is self-supervised training for DNS models [18] which enables the use of real-world data for training. Another is to have a unified model for both personalized and non-personalized speech enhancement [17]. These two approaches could potentially be combined into a single self-supervised DNS model which can perform personalized and non-personalized DNS.

Future challenges could also relax the real-time requirements, which would allow much more complex models to be applied, e.g., multi-modal large language models [66], [67], [68]. In addition, we could also relax the latency requirements, which would be useful for non-real-time scenarios such as offline speech enhancement of recorded meetings.

REFERENCES

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

[2] J. Benesty, J. Chen, and A. P. E. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer, Sep. 2011.

[3] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[4] D. Fischer, K. Brumann, and S. Doclo, "Comparison of parameter estimation methods for single-microphone multi-frame wiener filtering," in *Proc. IEEE 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[5] S. Braun and A. P. E. Habets, "Linear prediction-based online dereverberation and noise reduction using alternating Kalman filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1119–1129, Jun. 2018.

[6] M. Parchami, W.-P. Zhu, B. Champagne, and E. Plourde, "Recent developments in speech enhancement in the short-time fourier transform domain," *IEEE Circuits Syst. Mag.*, vol. 16, no. 3, pp. 45–77, Third Quarter, 2016.

[7] DeLiang Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[8] J. M. Valin, "A hybrid DSP/Deep learning approach to real-time full-band speech enhancement," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process.*, 2018, pp. 1–5.

[9] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," 2019, *arXiv:1903.03107*.

[10] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 656–660.

[11] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Appl. Acoust.*, vol. 187, Feb. 2022, Art. no. 108499.

[12] C. K. A. Reddy et al., "ICASSP 2021 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6623–6627.

[13] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6493–6497.

[14] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 886–890.

[15] B. Naderi and R. Cutler, "An open source implementation of ITU-T recommendation P.808 with validation," in *Proc. INTERSPEECH Conf.*, 2020, pp. 2862–2866.

[16] B. Naderi and R. Cutler, "Subjective evaluation of noise suppression algorithms in crowdsourcing," in *Proc. INTERSPEECH Conf.*, 2021, pp. 2132–2136.

[17] Z. Wang, R. Giri, D. Shah, J.-M. Valin, M. M. Goodwin, and P. Smaragdis, "A framework for unified real-time personalized and non-personalized speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[18] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[19] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 9206–9210.

[20] C. K. A. Reddy et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. INTERSPEECH Conf.*, 2020, pp. 2492–2496.

[21] C. K. A. Reddy et al., "INTERSPEECH 2021 deep noise suppression challenge," in *Proc. INTERSPEECH Conf.*, 2021, pp. 2796–2800.

[22] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Process. Mag.*, vol. 40, no. 3, pp. 8–29, May 2023.

[23] H. Dubey et al., "ICASSP 2022 deep noise suppression challenge," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 9271–9275.

[24] *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm*, Rec. ITU-T P.835, International Telecommunications Union, Geneva, Switzerland, 2003.

[25] *Methods for Subjective Determination of Transmission Quality*, Rec. ITU-T P.800, International Telecommunications Union, Geneva, Switzerland, 1996.

[26] *Subjective Evaluation of Speech Quality with a Crowdsourcing Approach*, Rec. ITU-T P.808, International Telecommunications Union, Geneva, Switzerland, 2018.

[27] T. Hossfeld et al., "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, pp. 541–558, 2014.

[28] R. Z. Jiménez, B. Naderi, and S. Möller, "Effect of environmental noise in speech quality assessment studies using crowdsourcing," in *Proc. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.

[29] B. Naderi, M. Sebastian, and G. Mittag, "Speech quality assessment in crowdsourcing: Influence of environmental noise," *Deutsche Jahrestagung für Akustik*, vol. 44, pp. 229–302, 2018.

[30] F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 2416–2419.

[31] "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," Rec. ITU-T P.563, International Telecommunications Union, Geneva, Switzerland, 2011.

[32] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Brighton, U.K., May 2019, pp. 631–635.

[33] X. Dong and D. S. Williamson, "An attention enhanced multi-task model for objective speech assessment in real-world environments," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Barcelona, Spain, 2020, pp. 911–915.

[34] H. Gamper, C. K. A. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 85–89.

[35] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, “Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM,” in *Proc. INTERSPEECH Conf.*, 2018, pp. 1873–1877.

[36] A. A. Catellier and S. D. Voran, “Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 331–335.

[37] B. Cauchi et al., “Non-intrusive speech quality prediction using modulation energies and LSTM-Network,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1151–1163, Jul. 2019.

[38] G. Yi et al., “ConferencingSpeech 2022 challenge: Non-intrusive objective speech quality assessment (NISQA) challenge for online conferencing applications,” in *Proc. INTERSPEECH Conf.*, 2022, pp. 3308–3312.

[39] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, “A differentiable perceptual audio metric learned from just noticeable differences,” in *Proc. INTERSPEECH*, 2020, pp. 2852–2856.

[40] J. Serrá, J. Pons, and S. Pascual, “SESQA: Semi-supervised learning for speech quality assessment,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 381–385.

[41] K. Sridhar et al., “ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 151–155.

[42] R. Cutler et al., “INTERSPEECH 2021 acoustic echo cancellation challenge,” in *Proc. INTERSPEECH Conf.*, 2021, pp. 4748–4752.

[43] R. Cutler et al., “ICASSP 2022 acoustic echo cancellation challenge,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 9107–9111.

[44] R. Cutler et al., “ICASSP 2023 acoustic echo cancellation challenge,” *IEEE Open J. Signal Process.*, early access, doi: 10.1109/OJSP.2024.3376289.

[45] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, “INTERSPEECH 2022 audio deep packet loss concealment challenge,” in *Proc. INTERSPEECH Conf.*, 2022, pp. 580–584.

[46] R. Cutler, A. Saabas, B. Naderi, N.-C. Ristea, S. Braun, and S. Branets, “ICASSP 2023 speech signal improvement challenge,” Apr. 2023, *arXiv:2303.06566*.

[47] B. Naderi, R. Cutler, and N.-C. Ristea, “Multi-dimensional speech quality assessment in crowdsourcing,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 696–700.

[48] S. Braun and I. Tashev, “On training targets for noise-robust voice activity detection,” in *Proc. 29th Eur. Signal Process. Conf.*, 2021, pp. 421–425.

[49] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, “ECAPA-TDNN embeddings for speaker diarization,” in *Proc. Interspeech Conf.*, 2021, pp. 3560–3564.

[50] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification,” in *Proc. INTERSPEECH Conf.*, 2020, pp. 3830–3834.

[51] “DNSMOS Git Repo,” [Online]. Available: https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[52] F. Jort et al., “Audio set: An ontology and human-labeled dataset for audio events,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 776–780.

[53] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 5220–5224.

[54] S. E. Eskimez, T. H. Yoshioka, X. Wang, Z. W. Chen, and X. Huang, “Personalized speech enhancement: New models and comprehensive evaluation,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 356–360.

[55] M. Thakker, S. E. Eskimez, T. Yoshioka, and H. Wang, “Fast real-time personalized speech enhancement: End-to-end enhancement network (E3Net) and knowledge distillation,” in *Proc. INTERSPEECH Conf.*, 2022, pp. 991–995.

[56] B. Naderi and R. Cutler, “A crowdsourcing approach to video quality assessment,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 2810–2814.

[57] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. 40th Int. Conf. Mach. Learn.*. 2023, pp. 28492–28518.

[58] Y. Ju et al., “TEA-PSE 3.0: Tencent-ethereal-audio-lab personalized speech enhancement system for ICASSP 2023 DNS-challenge,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–2.

[59] X. Yan, Y. Yang, Z. Guo, L. Peng, and L. Xie, “The NPU-Elevoc personalized speech enhancement system for ICASSP 2023 DNS challenge,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–2.

[60] J. Yu, H. Chen, Y. Luo, R. Gu, W. Li, and C. Weng, “TSpeech-AI system description to the 5th deep noise suppression (DNS) challenge,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–2.

[61] Z. Hou, Q. Hu, T. Sun, Y. Hu, C. Zhu, and K. Chen, “Convolutional recurrent MetriCGAN with spectral dimension compression for full-band speech enhancement,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–2.

[62] X. Le et al., “Personalized speech enhancement combining band-split RNN and speaker attentive module,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–2.

[63] J. Yu, Y. Luo, H. Chen, R. Gu, and C. Weng, “High fidelity speech enhancement with band-split RNN,” in *Proc. INTERSPEECH*, 2023, pp. 2483–2487.

[64] M. Ravanelli et al., “SpeechBrain: A general-purpose speech toolkit,” Jun. 2021, *arXiv:2106.04624*.

[65] H. Wang et al., “Wespeaker: A research and production oriented speaker embedding learning toolkit,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Rhodes Island, Greece, 2023, pp. 1–5.

[66] Z. Borsos et al., “AudioLM: A language modeling approach to audio generation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2523–2533, 2023.

[67] P. K. Rubenstein et al., “AudioPaLM: A large language model that can speak and listen,” Jun. 2023, *arXiv:2306.12925*.

[68] Y. Shu et al., “LLaSM: Large language and speech model,” Sep. 2023, *arXiv:2308.15930*.

HARISHCHANDRA DUBEY received the B.Tech. degree in electronics and communication engineering from the Motilal Nehru National Institute of Technology, Allahabad, India, in 2012, the M.Sc. degree from the FAU University of Erlangen-Nuremberg, Erlangen, Germany, in 2015, and the Ph.D. degree in electrical engineering from the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX, USA. He was an Applied Scientist with Microsoft.

ASHKAN AAZAMI received the B.S. degree in computer hardware in 2000, the M.S. degree in theoretical computer science from Sharif University, Tehran, Iran, in 2002, the second M.S. degree in software engineering from the University of Southern California, Los Angeles, CA, USA, in 2004, and the Ph.D. degree in combinatorics & optimization from the University of Waterloo, Waterloo, ON, Canada, in 2008, for his research on approximation algorithms and limits of approximation in network and graph theory problems. He is currently a Senior Data Scientist with IC3-AI Group, Microsoft, where he applies AI to improve Teams/Skype audio and video quality and reliability. His research interests include speech enhancement and bandwidth control using deep learning, subjective evaluation, online experimentation (A/B testing), and optimization problems in general.

VISHAK GOPAL received the bachelor's degree in electronics and communication engineering from Visvesvaraya Technological University, Belagavi, Karnataka, India, in 2005. Since 2008, he has been with Microsoft, starting off as a Software Developer in Dynamics in Copenhagen, Denmark to Skype/Teams in Redmond, USA. He is currently an Engineering Manager with IC3-AI Group, Microsoft, where he manages a team of applied scientists and software engineers focusing on improvements into Teams/Skype audio/video stack in areas like noise suppression and bandwidth control. His research interests include communication stack from signaling services, bot framework, telemetry pipeline to ML models in the media stack.

BABAK NADERI received the Ph.D. degree from the Technical University of Berlin, Berlin, Germany, in 2017, for work on crowdsourcing. He is currently a Senior Applied Scientist with IC3-AI Group, Microsoft, where he is working on the application of AI for improving Teams/Skype audio and video quality and reliability. His research interest includes media quality assessment, application of deep learning in quality of experience, machine learning, and enhanced crowdsourcing.

SEBASTIAN BRAUN (IEEE Member, IEEE SPS Member) received the M.Sc. degree in electrical and sound engineering from the Technical University (TU), Berlin, Germany, and University of Arts (KU) Graz, Graz, Austria, respectively, in 2012, and the Ph.D. degree from International Audio Laboratories, a joint institution of the University of Erlangen-Nuremberg, Erlangen, Germany, and the Fraunhofer IIS, Erlangen, Germany, in 2018. He is currently a Researcher with Audio and Acoustics Research Group, Microsoft Research. He has authored or coauthored more than 20 peer-reviewed conference and journal papers. His research interests include audio and acoustic signal processing, speech enhancement and separation, microphone array processing, and spatial and binaural audio processing. He was the recipient of the Best Student Paper Award at IWAENC 2014.

ROSS CUTLER (IEEE Member) received the B.S. degree in mathematics, computer science, and physics in 1992, and the Ph.D. degree in computer science in the area of computer vision from the University of Maryland, College Park, MD, USA, 2000. He is currently a Distinguished Engineer with IC3 Group, Microsoft, where he manages the Team of Applied Scientists and Software Engineers with a focus on improving Teams/Skype audio/video quality and reliability and enabling new functionality with AI. Since 2000, he has been with Microsoft, starting as a Researcher with Microsoft Research. He has authored or coauthored more than 80 peer-reviewed conference and journal papers and has more than 100 granted patents in his research interests which include computer vision, machine learning, signal processing, acoustics, optics, and VoIP.

ALEX JU received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 2020. He is currently an Applied Scientist II with Azure AI Group, Microsoft. Since 2000, he has been Microsoft. His research interests include machine learning, audio processing, and speech enhancement and separation.

MEHDI ZOHOURIAN received the Ph.D. degree in electrical engineering and information technology from the Ruhr-University, Bochum, Germany, in 2019. He was an Audio Signal Processing Algorithm Developer with the hearing-aid manufacturer WSAudiology (formerly known as Siemens Audiology Technique), Erlangen, Germany. Since 2021, he has been with Microsoft. He is currently an Applied Scientist with Azure Cognitive Services Team, Microsoft, where he is working on AI-based techniques for speech enhancement. His research interests include audio signal processing, microphone array, and binaural hearing aids.

MIN TANG received and the M.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1996, and the M.S. degree in computer science with Colorado University, Boulder, CO, USA, in 2005. He was a Speech Scientist with Nuance and VoiceBox, working on in-car speech recognition. Since 2019, he has been with Microsoft. He is currently a Principal Applied Scientist Manager with Azure Cognitive Services Team, Microsoft, where he is working on AI-based techniques for speech enhancement, speech recognition, and text to speech.

MEHRSA GOLESTANEH received the B.S. degree in computer science from the Amirkabir University of Technology - Tehran Polytechnic, Tehran, Iran, in 2012, and the M.S. degree in computer science from Western University, London, ON, Canada, in 2014. Since 2021, she has been with Microsoft. She is currently a Senior Product Manager with Intelligent Communication and Conversation Cloud division, where she helps develop the calling and messaging infrastructure technology for Skype and Teams by applying advanced ML algorithms.

ROBERT AICHNER received the Ph.D. degree in electrical engineering in the area of audio signal processing from the University of Erlangen-Nuremberg, Erlangen, Germany. Since 2007, he has been with Microsoft starting as an Audio Signal Processing Developer and then held multiple roles in program management ranging from a focus on the algorithms in the real-time communications stack to also working in the user interface/user experience area. He is currently a Partner Group Program Manager with Intelligent Communication and Conversation Cloud Group, Microsoft, responsible for the areas of machine learning and call quality. He has authored or coauthored more than 30 conference papers, journal papers, and book chapters.