Received 5 August 2023; revised 13 October 2023; accepted 16 October 2023. Date of publication 12 March 2024; date of current version 18 June 2024. The review of this article was arranged by Associate Editor Atsunori Ogawa.

*Digital Object Identifier 10.1109/OJSP.2024.3376293*

# ICASSP 2023 Speech Signal Improvement Challenge

**ROSS CUTLER (Member, IEEE), ANDO SAABAS, BABAK NADERI, NICOLAE-CĂTĂLIN RISTEA, SEBASTIAN BRAUN (Senior Member, IEEE), AND SOLOMIYA BRANETS**

Microsoft Corporation, Redmond, WA 98052 USA

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

**ABSTRACT** The ICASSP 2023 Speech Signal Improvement Challenge is intended to stimulate research in the area of improving the speech signal quality in communication systems. The speech signal quality can be measured with SIG in ITU-T P.835 and is still a top issue in audio communication and conferencing systems. For example, in the ICASSP 2022 Deep Noise Suppression challenge, the improvement in the background and overall quality is impressive, but the improvement in the speech signal is not statistically significant. To improve the speech signal the following speech impairment areas must be addressed: coloration, discontinuity, loudness, reverberation, and noise. A training and test set was provided for the challenge, and the winners were determined using an extended crowdsourced implementation of ITU-T P.804's listening phase. The results show significant improvement was made across all measured dimensions of speech quality.

**INDEX TERMS** Speech enhancement, deep learning, subjective testing, speech quality assessment.

## I. INTRODUCTION

Audio telecommunication systems such as remote collaboration systems (Microsoft Teams, Skype, Zoom, etc.), smartphones, and telephones are used by nearly everyone on the planet and have become essential tools for both work and personal usage. Since the invention of the telephone in 1876 by Alexander Graham Bell, audio engineers, and researchers have innovated to improve the speech quality of telecommunication systems, with the ultimate goal of making audio telecommunication systems as good or better than face-to-face communication. After nearly 150 years of effort, there is still a long way to go toward this goal, especially with the use of mainstream devices. For example, it is still common to hear frequency response distortions, isolated and non-stationary distortions, loudness issues, reverberation, and background noise in audio calls.

The ICASSP 2023 Speech Signal Improvement Challenge is intended to stimulate research in the area of improving the send speech signal[1] quality in mainstream telecommunication systems. Subjective speech quality assessment is the gold standard for evaluating speech enhancement, processing, and telecommunication systems. The ITU-T has developed several recommendations for subjective speech quality assessment. In particular, the ITU-T Rec. P.835 [1] provides a lab-based subjective evaluation framework targeting systems that include noise suppression algorithm that gives quality scores of the speech signal (SIG), background noise (BAK), and overall quality (OVRL). In this framework, participants are asked to listen to short clips of speech in a controlled environment and rate the quality of each clip in terms of the speech signal, background noise, and overall quality on three discrete Likert scales (where 1 is Bad quality and 5 is Excellent quality). Each clip is measured by multiple raters, and the results are averaged to obtain a Mean Opinion Score (MOS). By measuring SIG, BAK, and OVRL, P.835 provides a more reliable subjective assessment [1] and allows researchers to determine which area to focus on for improving the overall quality.

The speech signal is still a top issue in audio telecommunication and conferencing systems. For example, in the ICASSP 2022 Deep Noise Suppression Challenge [2], the improvement in BAK and OVRL quality is impressive, but no improvement in SIG was observed. The same was true for the INTERSPEECH 2021 Deep Noise Suppression

[1] In telecommunication, the audio captured by a near end microphone, processed, and sent to the far end is called the send signal.

**TABLE 1.** Amount of Improvement (In Differential MOS (DMOS)) Remaining to Get Excellent Quality (MOS = 5) Rated Speech Based on the ICASSP 2022 DNS Challenge [2]

| Area | Headroom (DMOS) |
|---|---|
| SIG | 0.70 |
| BAK | 0.30 |
| OVRL | 0.87 |

DMOS is on a Scale of 0-4, Where 0 is No Difference and 4 is Very Annoying Compared to Excellent Quality Speech.

**TABLE 2.** Related Challenges

| Area | Related challenge |
|---|---|
| Noisiness | Deep Noise Suppression [21], [22], [3], [2], [4] |
| Coloration | None |
| Discontinuity | Packet Loss Concealment [23] |
| Loudness | None |
| Reverberation | REVERB [24] |
| Echo | AEC Challenge [25], [26], [27], [28] |

Challenge [3], and for the more recent ICASSP 2023 Deep Noise Suppression Challenge [4], which focuses on personalized noise suppression. Table 1 shows the amount of improvement in SIG, BAK, and OVRL to get excellent quality rated speech (MOS=5) for the ICASSP 2022 Deep Noise Suppression Challenge. This shows the key area of improvement is SIG, which has 2.3× more improvement opportunities than BAK. To improve SIG, the following dimensions of speech quality should be improved [5]:

- Coloration: Frequency response distortions
- Discontinuity: Isolated and non-stationary distortions
- Loudness: Important for the overall quality and intelligibility
- Reverberation: Room reverberation of speech and noise signals
- Noisiness: Background noise and circuit and coding noise

The correlation of SIG to these dimensions is given in Fig. 5. Theoretically, improving BAK is not necessary to improve SIG as they are orthogonal metrics by design. However, in practice, it is hard for subjective test participants to assess speech signal quality in the presence of strong dominant background noise.

## II. RELATED WORK

While there have been previous challenges in background noise and reverberation, there have been no challenges in coloration and loudness and a limited challenge in discontinuities (see Table 2). Moreover, there have been no previous challenges that explicitly measure and target improving SIG.

There are many previous methods to improve noisiness, coloration, discontinuity, loudness, and reverberation separately. Two new methods that target universal improvement of the speech signal are [6], [7].

The ITU-T has developed several recommendations for subjective speech quality assessment. ITU-T P.800 [8] describes lab-based methods for the subjective determination of speech quality, including the Absolute Category Rating (ACR). ITU-T P.808 [9] describes a crowdsourcing approach for conducting subjective evaluations of speech quality. It provides guidance on test material, experimental design, and a procedure for conducting listening tests in the crowd. The methods are complementary to laboratory-based evaluations described in P.800. An open-source implementation of P.808 is described in [10]. An open-source implementation of P.835 is described in [11]. More recent multidimensional speech quality assessment standards are ITU-T P.863.2 [12] and P.804 [5] (listening phase), which measure perceptual dimensions of speech quality namely noisiness, coloration, discontinuity, and loudness (see Table 3).

Intrusive objective speech quality assessment tools such as Perceptual Evaluation of Speech Quality (PESQ) [13] and Perceptual Objective Listening Quality Analysis (POLQA) [14] require a clean reference of speech. Non-intrusive objective speech quality assessment tools like ITU-T P.563 [15] do not require a reference, though it has a low correlation to subjective quality [16]. Newer neural net-based methods such as [16], [17], [18], [19] provide better correlations to subjective quality. NISQA [20] is an objective metric for P.804, though the correlation to subjective quality is not sufficient to use as a challenge metric (in the ConferencingSpeech 2022 Challenge [19] NISQA was used as a baseline model and achieved a Pearson Correlation Coefficient = 0.724 to MOS).

## III. CHALLENGE DESCRIPTION

This challenge benchmarks the performance of speech enhancement models with a real (not simulated) test set. The telecommunication scenario is the near end only send signal; it does not include echo impairments (there is no far end speech or noise). Participants evaluated their speech enhancement model (SEM) on a test set and submitted the results (clips) for subjective evaluation.

### A. CHALLENGE TRACKS

The challenge has two tracks:

1) Real-time SEM
2) Non-real-time SEM

The goal of the first track is to develop something that can be used today on a typical personal computer, while the goal of the second track is to develop something that could be run on computers much faster than a typical personal computer or be run offline.

**TABLE 3.** **Speech Quality Areas From P.804 Listening Phase (The First Four) Plus Three Additional Areas**

| Area | Description | Possible source |
|---|---|---|
| Noisiness | Background noise, circuit noise, coding noise; BAK | Coding, circuit or background noise; device |
| Coloration | Frequency response distortions | Bandwidth limitation, resonances, unbalanced freq. response |
| Discontinuity | Isolated and non-stationary distortions | Packet loss; processing; non-linearities |
| Loudness | Important for the overall quality and intelligibility | Automatic gain control; mic distance |
| Reverberation | Room reverberation of speech and noise | Rooms with high reverberation |
| Speech Signal | SIG | |
| Overall | OVRL | |

### B. LATENCY AND RUNTIME REQUIREMENTS

Algorithmic latency is defined by the offset introduced by the whole processing chain including short-time Fourier transform (STFT), inverse STFT, overlap-add, additional lookahead frames, etc., compared to just passing the signal through without modification. It does not include buffering latency. Some examples are:

- A STFT-based processing with window length = 20 ms and hop length = 10 ms introduces an algorithmic delay of window length – hop length = 10 ms.
- A STFT-based processing with window length = 32 ms and hop length = 8 ms introduces an algorithmic delay of window length – hop length = 24 ms.
- An overlap-save-based processing algorithm introduces no additional algorithmic latency.
- A time-domain convolution with a filter kernel size = 16 samples introduces an algorithmic latency of kernel size – 1 = 15 samples. Using one-sided padding, the operation can be made fully "causal", i.e., left-sided padding with kernel size - 1 samples would result in no algorithmic latency.
- A STFT-based processing with window_length = 20 ms and hop_length = 10 ms using 2 future frames information introduces an algorithmic latency of (window_length – hop_length) + 2 * hop_length = 30 ms.

Buffering latency is defined as the latency introduced by block-wise processing, often referred to as hop length, frameshift, or temporal stride. Some examples are:

- A STFT-based processing has a buffering latency corresponding to the hop size.
- A overlap-save processing has a buffering latency corresponding to the frame size.
- A time-domain convolution with stride 1 introduces a buffering latency of 1 sample.

Real-time factor (RTF) is defined as the fraction of time it takes to execute one processing step. For an STFT-based algorithm, one processing step is the hop size. For a time-domain convolution, one processing step is 1 sample. RTF = compute time / time step.

All models submitted to this challenge must meet all of the below requirements:

1) To be able to execute an algorithm in real-time, and to accommodate for variance in compute time which occurs in practice, we require RTF $\leq$ 0.5 in the challenge on an Intel Core i5 Quadcore clocked at 2.4 GHz using a single thread.
2) Algorithmic latency + buffering latency $\leq$ 20 ms.
3) No future information can be used during model inference.

More details of the challenge are available at https://aka.ms/sig-challenge.

## IV. TRAINING SET

This challenge suggested using the ICASSP 2022 Deep Noise Suppression Challenge [2] and ICASSP 2022 Acoustic Echo Cancellation Challenge [27] training and test sets for training. The AEC Challenge training set in particular includes over 10K unique environments, devices, and speakers. The near end single talk clips have been rated using P.835 and are provided, which can be used during training to improve SIG and OVRL.

However, any training set could have been used, such as [24], [29], [30].

## V. TEST SET

The test set consisted of 500 send clips, each using a unique device, environment, and person speaking. The clips were captured from both PCs and mobile devices using the same methodology as described in [27]. The recordings were stratified to have an approximately uniform distribution for the impairment areas listed in Table 3. The test set language contains English, German, Dutch, French, and Spanish languages, with the majority of files (around 80%) in English. The test set was released near the end of the competition. The distribution of subjective ratings based on P.804 (see Section VI) of the test set for all dimensions is shown in Fig. 1.

## VI. EVALUATION METHODOLOGY

The challenge evaluation is based on a subjective listening test. We have developed an extension of P.804 (listening phase) / P.863.2 (Annex A) based on crowdsourcing and the P.808 toolkit [10] for subjective evaluation. In particular, we added reverberation, speech quality, and overall quality to P.804's listening phase (see Table 3). Details of this P.804 extension are given in Section VI-A and [31].

The challenge metric $M \in [0, 1]$ is:

$$M = \frac{(\text{SIG} - 1)/4 + (\text{OVRL} - 1)/4}{2} \tag{1}$$

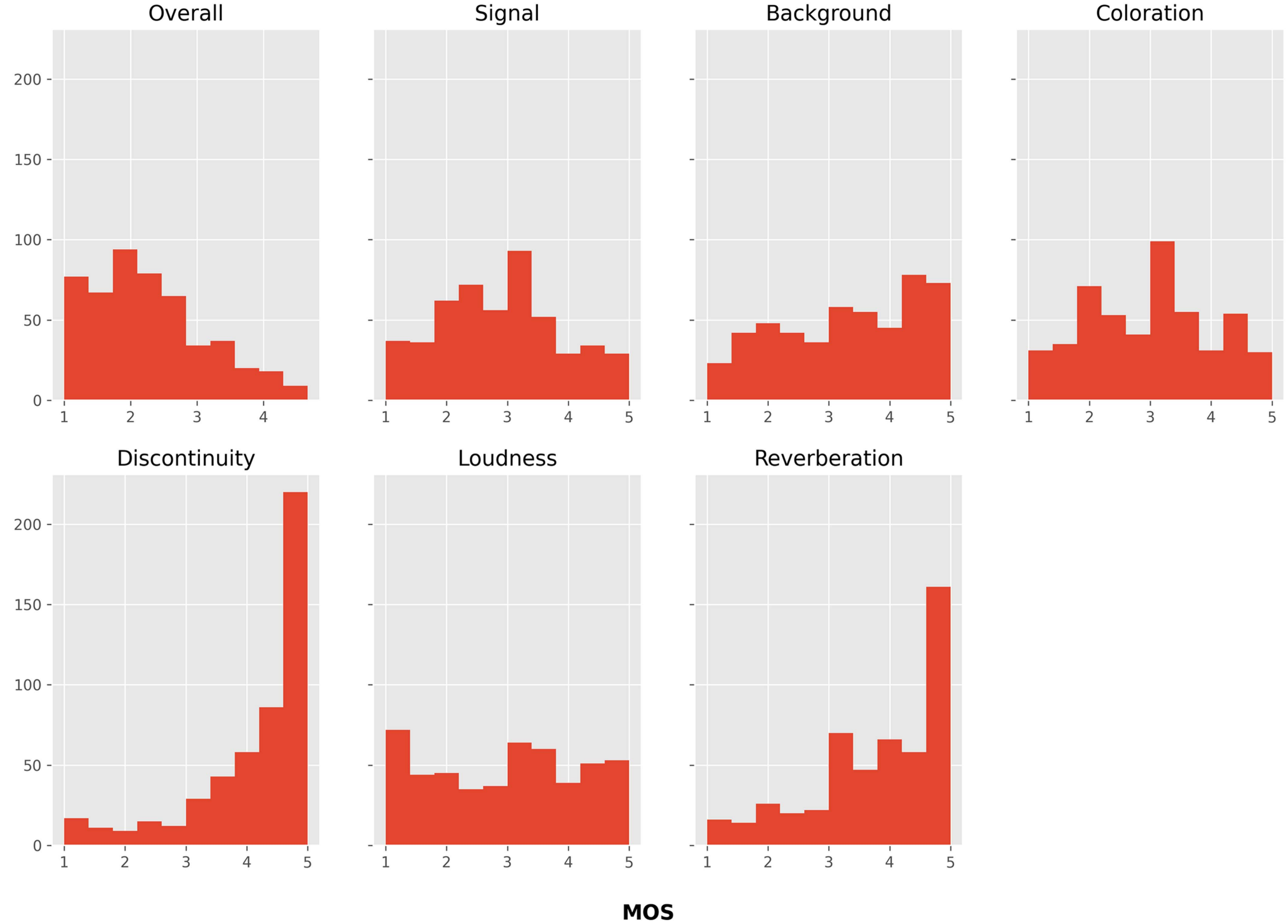


**FIGURE 1.** **Distribution of subjective scores of clips in the blind test set. Ideally, the distribution would be uniform for each dimension, but it is skewed for discontinuity and reverberation.**

In addition, differential SIG (DSIG) must be $> 0$. Since OVRL $\sim$ BAK + SIG [11], BAK should also be improved to increase OVRL performance. The other metrics measured in Table 3 are informational only. However, in Section VIII-E we show that overall is influenced by the speech signal, reverberation, loudness, discontinuity, coloration, and noisiness, so optimizing each is a good strategy. The clips are evaluated as fullband (48 kHz) in P.804, so frequency extension can help.

### *A. ONLINE SUBJECTIVE EVALUATION FRAMEWORK*

We extended the P.808 Toolkit [10] to include a test template for a multi-dimensional quality assessment. The toolkit provides scripts for preparing the test, including packing the test clips in small test packages, preparing the reliability check questions, and analyzing the results. We ask participants to rate the perceptual quality dimensions of speech namely coloration, discontinuity, noisiness, loudness, reverberation, signal quality, and overall quality of each audio clip. In the following, each section of the test template, as seen by participants, is described. These sections are predefined and only the audio clips under the test will be changed from one study to another.

In the first section, the participant's eligibility and device suitability are tested and a qualification is assigned to those that pass which remains valid for the entire experiment. The participant's hearing ability is evaluated through digit-triplet-test [32]. Moreover, we test if their listening device supports the required bandwidths (i.e., fullband, wideband, and narrowband); details are in Section VI-A1).

Next, the participant's environment and device are tested using a modified-JND test [33] in which they should select which stimulus from a pair has a better quality in four questions. A temporal certificate will be issued for participants after passing this section which expires after two hours and consequently repeating this section will be required. Detailed instructions are given in the next section including introducing the rating scales and providing multiple samples for each perceptual dimension. Participants are required to listen to all samples for the first time. Fig. 2 illustrates how the rating scale for quality dimensions is presented to participants. In addition, we used a Likert 5-point scale for signal quality and overall quality as specified by ITU-T Rec. P.835. In the Training section participants should first adjust the playback loudness to a comfortable level by listening to a provided sample and

**TABLE 4. Labels on Each Scale's Pole and Descriptive Adjectives Provided to Participants**

| Scale | Positive Pole | Negative Pole |
|---|---|---|
| Discontinuity | **Continuous**<br>Steady, Smooth, Clean | **Discontinuous**<br>Shaky, Choppy, Uneven |
| Loudness | **Optimal loudness**<br>Easy to hear, Pleasant, Level | **Sub-optimal loudness**<br>Too quiet, Varying volume, Too loud |
| Noisiness | **Not noisy**<br>Clean/Clear, Noiseless, Not hissing | **Noisy**<br>Buzzy, Hissing, Clanging |
| Coloration | **Uncolored**<br>Normal, Natural, Direct | **Colored**<br>Distant/Far, Thin, Muffled |
| Reverberation | **No reverb**<br>Clear, Clean, No echo | **High reverb**<br>Echo, Sound reflection, Tunnel sound |

Terms Used in ITU-T Rec. P.804 are Marked in Red.

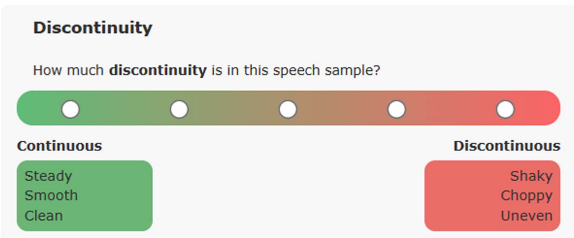


**FIGURE 2. Sub-dimensions are rated on a 5-point discrete scale with descriptive adjectives on poles.**

then rate 7 audio clips. This section is similar to the ratings section, but the platform provides live feedback based on their ratings. By completing this section a temporal certificate is assigned to the participants which is valid for one hour. Last is the Ratings section, where participants listen to ten audio clips and two gold standard and trapping questions and cast their votes on each scale. The gold standard questions are the ones that the experimenter already knows their answers (being excellent or bad) and participants are expected to vote on each scale with a minor deviation from known the answer [32]. Trapping questions are questions in which a synthetic voice is overlaid to a normal clip and asks participants to provide a specific vote to show their attention [34]. For this test, we provide scripts for creating the trapping clips, which ask participants to select answers reflecting the best or worst quality in all scales. For rating an audio clip, the participant should first listen to the end of the clip, and then they start casting their votes. During that time, the audio will be played back in a loop. After participants finish with a test set, they can continue with the next one where only the rating section will be shown until other temporal certificates are valid. By the expiration of any certificate, the corresponding section will be shown when they start the next test set.

#### 1) SURVEY OPTIMIZATION

We utilized the multi-scale template in various research studies and improved it through the incorporation of experts and test participant feedback.

**Descriptive adjectives:** The understanding of perceptual dimensions might not be intuitive for naive test participants, therefore the P.804 recommendation includes a set of descriptive adjectives to describe the presence or absence of each quality dimension. We expanded this list through multiple preliminary studies, where participants were asked to listen to samples from each perceptual dimension and name three adjectives that best describe them. For each dimension, we selected the top three most frequently selected terms and presented them below each pole of the scale, as shown in Fig. 2. The list of selected terms is reported in Table 4. We used discrete scales for dimensions to be consistent with signal and overall scales.

*Bandwidth check:* This test ensures the participant devices support the expected bandwidth. The test consists of five samples, and each has two parts separated by a beep tone. The second part is the same as the first part but in three samples superimposed by additive noise. Participants should listen to each sample and select if both parts have the same or different quality. We filtered the white noise with the following bandpass filters: 3.5–22K (all devices should play the noise), 9.5–22k (super-wideband or fullband is supported), and 15–22K (fullband is supported).

*Gold questions:* Gold questions are widely used in crowdsourcing [32]. Here we observed gold questions that represent the strong presence of an impairment on one dimension and the clear absence of impairment on all dimensions can best reveal an inattentive participant.

*Randomization:* We randomize the presentation order of scales for each participant. However, the signal and overall quality are always presented at the end. The randomized order is kept for each participant until a new round of training is required.

## VII. RESULTS

There were 7 entries for the real-time track and 5 for the non-real-time track, though the top 3 for non-real-time track were identical submissions to the real-time track and therefore were only considered for the real-time track. Team Cvt-tencent was statistically tied with Legends-tencent and withdrew.

**TABLE 5. Real-Time Track Challenge Results**

| Model | Final | Overall | | Signal | | Background | | Coloration | | Discontinuity | | Loudness | | Reverberation | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Score | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | 95%CI |
| Legends-tencent | 0.610 | 3.271 | 0.911 | 3.612 | 0.684 | 4.636 | 1.333 | 3.598 | 0.569 | 4.201 | 0.140 | 4.109 | 1.117 | 4.316 | 0.465 | 0.043 |
| Ctv-tencent | 0.606 | 3.261 | 0.901 | 3.589 | 0.662 | 4.599 | 1.297 | 3.568 | 0.539 | 4.202 | 0.140 | 4.097 | 1.105 | 4.340 | 0.488 | 0.044 |
| Genius-team | 0.589 | 3.178 | 0.818 | 3.535 | 0.608 | 4.511 | 1.208 | 3.550 | 0.521 | 4.140 | 0.079 | 4.060 | 1.068 | 4.322 | 0.471 | 0.044 |
| Hitiot | 0.531 | 2.965 | 0.604 | 3.280 | 0.353 | 4.592 | 1.289 | 3.248 | 0.218 | 4.005 | -0.057 | 3.916 | 0.924 | 4.447 | 0.625 | 0.045 |
| Noisy | 0.411 | 2.360 | 0.000 | 2.927 | 0.000 | 3.302 | 0.000 | 3.029 | 0.000 | 4.061 | 0.000 | 2.992 | 0.000 | 3.852 | 0.000 | 0.051 |
| Njuaa-lab | 0.480 | 2.398 | 0.038 | 2.863 | −0.065 | 3.794 | 0.491 | 2.945 | −0.084 | 3.835 | −0.227 | 3.277 | 0.0285 | 4.222 | 0.371 | 0.049 |
| N&B | 0.385 | 2.346 | −0.014 | 2.737 | −0.190 | 4.221 | 0.918 | 2.836 | −0.194 | 3.657 | −0.045 | 3.119 | 0.127 | 4.132 | 0.280 | 0.050 |
| Kuaishou | 0.381 | 2.363 | 0.002 | 2.684 | −0.244 | 3.685 | 0.383 | 3.109 | 0.080 | 3.206 | −0.855 | 3.444 | 0.452 | 4.374 | 0.523 | 0.048 |

Multi-Dimensional Subjective Test - Extension of ITU-T P.863.2 Annex A/Listening Phase of ITU-T P.804.

**TABLE 6. Real-Time Track Challenge Results**

| Model | Final | Overall | | Signal | | Background | | Average |
|---|---|---|---|---|---|---|---|---|
| | Score | MOS | DMOS | MOS | DMOS | MOS | DMOS | 95%CI |
| Legends-tencent | 0.616 | 3.350 | 0.527 | 3.581 | 0.434 | 4.208 | 0.755 | 0.04 |
| Ctv-tencent | 0.596 | 3.268 | 0.444 | 3.497 | 0.350 | 4.094 | 0.641 | 0.04 |
| Genius-team | 0.583 | 3.190 | 0.366 | 3.471 | 0.324 | 4.073 | 0.620 | 0.04 |
| Hitiot | 0.550 | 3.089 | 0.266 | 3.312 | 0.164 | 4.074 | 0.622 | 0.04 |
| Noisy | 0.496 | 2.824 | 0.000 | 3.147 | 0.000 | 3.453 | 0.000 | 0.04 |
| Njuaa-lab | 0.480 | 2.790 | −0.034 | 3.047 | −0.100 | 3.712 | 0.260 | 0.04 |
| N&B | 0.451 | 2.699 | −0.125 | 2.911 | −0.236 | 3.781 | 0.328 | 0.05 |
| Kuaishou | 0.462 | 2.747 | −0.077 | 2.952 | −0.195 | 3.690 | 0.238 | 0.04 |

Subjective Test Based on ITU-T P.835.

**TABLE 7. Non-Real-Time Track Challenge Results**

| Model | Final | Overall | | Signal | | Background | | Coloration | | Discontinuity | | Loudness | | Reverberation | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Score | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | MOS | DMOS | 95%CI |
| Legends-tencent* | 0.610 | 3.271 | 0.911 | 3.612 | 0.684 | 4.636 | 1.333 | 3.598 | 0.569 | 4.201 | 0.140 | 4.109 | 1.117 | 4.316 | 0.465 | 0.043 |
| Ctv-tencent* | 0.606 | 3.261 | 0.901 | 3.589 | 0.662 | 4.599 | 1.297 | 3.568 | 0.539 | 4.202 | 0.140 | 4.097 | 1.105 | 4.340 | 0.488 | 0.044 |
| Genius-team* | 0.589 | 3.178 | 0.818 | 3.535 | 0.608 | 4.511 | 1.208 | 3.550 | 0.521 | 4.140 | 0.079 | 4.060 | 1.068 | 4.322 | 0.471 | 0.044 |
| N&B | 0.446 | 2.608 | 0.247 | 2.964 | 0.036 | 4.058 | 0.756 | 3.131 | 0.102 | 3.673 | −0.389 | 3.601 | 0.609 | 4.335 | 0.484 | 0.048 |
| Hamburg | 0.445 | 2.570 | 0.210 | 2.988 | 0.061 | 3.765 | 0.463 | 3.330 | 0.300 | 3.764 | −0.387 | 3.435 | 0.443 | 4.241 | 0.390 | 0.048 |
| Noisy | 0.411 | 2.360 | 0.000 | 2.927 | 0.000 | 3.302 | 0.000 | 3.029 | 0.000 | 4.061 | 0.000 | 2.992 | 0.000 | 3.852 | 0.000 | 0.051 |

Multi-Dimensional Subjective Test - Extension of ITU-T P.863.2 Annex A/Listening Phase of ITU-T P.804. Teams With a * Had Identical Submissions to the Real-Time Track.

The P.804 and P.835 subjective results for both tracks are given in Tables 5–8. The ANOVAs for each track are given in Tables 9 and 10. P.835 results are given for reference only but agree with the P.804 results. Objective results are given in Table 11.

## VIII. ANALYSIS

### A. COMPARISON OF METHODS

A high-level comparison of the top-5 entries is given in Tables 12 and 13. Some observations are given below:

- The top entries improved SIG by DMOS $> 0.6$, unlike previous DNS challenges which had no SIG improvement [2], [3].
- The correlation between the training set hours (the total duration of data used) and the overall score is PCC $= 0.91$. The models with larger training sets tended to do better.
- The correlation between the runtime factor and the overall score is PCC $= -0.60$. We expected the non-real-time track entries to exceed the performance of the real-time track, but that was not the case. We observed a similar fact in the INTERSPEECH 2021 Deep Noise Suppression Challenge [3], where the non-real-time track also performed significantly worse than the real-time track. In both cases, we received more entries in the real-time track than non-real-time track, and there may be more researchers working on real-time speech enhancement than non-real-time speech enhancement.

**TABLE 8. Non-Real-Time Track Challenge Results**

| Model | Final | Overall | | Signal | | Background | | Average |
|---|---|---|---|---|---|---|---|---|
| | Score | MOS | DMOS | MOS | DMOS | MOS | DMOS | 95%CI |
| Legends-tencent* | 0.616 | 3.350 | 0.527 | 3.581 | 0.434 | 4.208 | 0.755 | 0.04 |
| Ctv-tencent* | 0.596 | 3.268 | 0.444 | 3.497 | 0.350 | 4.094 | 0.641 | 0.04 |
| Genius-team* | 0.583 | 3.190 | 0.366 | 3.471 | 0.324 | 4.073 | 0.620 | 0.04 |
| N&B | 0.517 | 2.967 | 0.143 | 3.165 | 0.018 | 3.870 | 0.418 | 0.04 |
| Hamburg | 0.495 | 2.842 | 0.018 | 3.119 | –0.028 | 3.684 | 0.232 | 0.04 |
| Noisy | 0.496 | 2.824 | 0.000 | 3.147 | 0.000 | 3.453 | 0.000 | 0.04 |

Subjective Test Based on ITU-T P.835. Teams With a * Had Identical Submissions to the Real-Time Track.

**TABLE 9. Real-Time Track ANOVA**

| Team | Legends-tencent | Ctv-tencent | Genius-team | Hitiot | Noisy | Njuaa-lab | N&B |
|---|---|---|---|---|---|---|---|
| Ctv-tencent | 0.609 | | | | | | |
| Genius-team | 0.001 | 0.004 | | | | | |
| Hitiot | 0.000 | 0.000 | 0.000 | | | | |
| Noisy | 0.000 | 0.000 | 0.000 | 0.000 | | | |
| Njuaa-lab | 0.000 | 0.000 | 0.000 | 0.000 | 0.681 | | |
| N&B | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | |
| Kuaishou | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.094 |

The Pairwise P-Values are Shown for the Lower-Triangular Matrix.

**TABLE 10. Non-Real-Time Track ANOVA**

| Team | Legends-tencent | Ctv-tencent | Genius-team | N&B | Hamburg |
|---|---|---|---|---|---|
| Ctv-tencent | 0.609 | | | | |
| Genius-team | 0.001 | 0.004 | | | |
| N&B | 0.000 | 0.000 | 0.000 | | |
| Hamburg | 0.000 | 0.000 | 0.000 | 0.285 | |
| Noisy | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

The Pairwise P-Values are Shown for the Lower-Triangular Matrix.

One approach to get better non-real-time models is to take the winner of the real-time track and increase the model size and complexity by 100x, very likely increasing the performance while making it no longer real-time.

- The correlation between the model size and the overall score is PCC = -0.58. Smaller models tended to perform better.
- The correlation between the number of stages and the overall score is PCC = 0.61. More stages tended to perform better.
- The top model by team Legends-tencent [35] significantly improved all measured speech quality dimensions, and did the best in all dimensions except reverberation. Their performance is illustrated in Fig. 3.
- A successful strategy used by teams Legends-tencent [35], Genius-team [36], and HITIoT [38] is a restoration module followed by a speech enhancement module. The generative models for restoration by teams Legends-tencent [35] and Genius-team [36] perform particularly well.
- There is still significant room for improvement in this test set for OVRL and SIG.
- None of the teams used the ICASSP 2022 Acoustic Echo Cancellation Challenge [27] dataset for training, even though it has thousands of clips of real-world speech signal impairments. This is likely because there is no clean speech available for this dataset, and using it would require semi-supervised or unsupervised training. Rather, all teams used the ICASSP 2022 Deep Noise Suppression Challenge [2] for a training set, and the winning team Legends-tencent [35] augmented that with a private training set.

### B. DISTRIBUTION OF DIMENSIONS

Fig. 4 shows the distribution of the subjective dimensions compared to overall quality at the model level. All of the dimensions except discontinuity and reverberation have a

**TABLE 11.** **The Objective Results on the Blind Set Obtained With DNSMOS model [18] (MOS _ SIG, MOS _ BAK, MOS _ OVR), and NISQA [20] (NISQA _ MOS Etc.)**

| Team | MOS SIG | MOS BAK | MOS OVR | NISQA MOS | NISQA COLOR | NISQA LOUDNESS | NISQA NOISE | NISQA DISCONTINUITY |
|---|---|---|---|---|---|---|---|---|
| Legends-tencent | 3.958 | 4.376 | 3.710 | 4.037 | 3.801 | 4.132 | 4.360 | 4.338 |
| Ctv-tencent | 3.954 | 4.358 | 3.695 | 3.993 | 3.783 | 4.105 | 4.325 | 4.313 |
| Genius-team | 3.894 | 4.305 | 3.623 | 3.925 | 3.752 | 4.081 | 4.288 | 4.266 |
| Hitiot | 3.708 | 4.344 | 3.487 | 3.392 | 3.294 | 3.758 | 4.146 | 3.803 |
| Kuaishou | 3.661 | 4.057 | 3.387 | 3.452 | 3.284 | 3.805 | 3.621 | 3.760 |
| Hamburg | 3.661 | 3.847 | 3.278 | 3.598 | 3.630 | 3.943 | 3.760 | 3.931 |
| N&B - Track 2 | 3.439 | 4.107 | 3.140 | 2.779 | 2.877 | 3.372 | 3.719 | 3.252 |
| Njuaa-lab | 3.281 | 3.895 | 2.914 | 2.178 | 2.567 | 2.898 | 3.441 | 2.813 |
| N&B - Track 1 | 3.183 | 4.101 | 2.902 | 2.474 | 2.689 | 3.012 | 3.650 | 3.227 |
| Noisy | 3.150 | 3.379 | 2.664 | 2.359 | 2.728 | 2.814 | 3.096 | 3.404 |

**TABLE 12.** **Comparison of the Top Five Teams for Multiple Dimensions**

| Place | Track | Team | Params | Real-time factor | Training set | Training set hours | Stages | Domain | $M$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Real-time | Legends-tencent [35] | 12.1 M | 0.37 | DNS [2], private | 1500 | 3 | time, STFT | 0.610 |
| 2 | Real-time | Genius-team [36] | 5.2 M | 0.36 | DNS [2] | 1500 | 2 | time, STFT | 0.589 |
| 3 | Real-time | HITIoT [37] | 9.2 M | 0.36 | DNS [2] | 1500 | 1 | STFT | 0.531 |
| 4 | Non-real-time | N&B [38] | 10 M | 1.48 | DNS [2] | 421 | 2 | STFT | 0.446 |
| 5 | Non-real-time | Hamburg [39] | 55.7 M | 30.1 | VCTK [40] | 28.2 | 1 | STFT | 0.445 |
| PCC to $M$ | | | -0.58 | -0.60 | | 0.91 | 0.61 | | |

We Included the Pearson Correlation Coefficient With Respect to the Final Score $M$.

**TABLE 13.** **Models Used by the Top Five Teams**

| Team | Model |
|---|---|
| Legends-tencent [35] | AGC → GSM-GAN (Restore) → Enhance |
| Genius-team [36] | TRGAN (Restore) → MTFAA-Lite (Enhance) |
| Hitot [37] | Half temporal, half frequency attention U-Net |
| N&B [38] | GateDCCRN (Repairing) → GateDCCRN, S-DCCRN (Denoising) |
| Hamburg [39] | Generative diffusion model (modified NCSN++) |

significant linear correlation to the overall quality (see Fig. 5). The high correlation between signal and overall quality (0.98 at the model level and 0.93 at the clip level) can be attributed to the preponderance of signal impairments in this dataset, as opposed to other datasets such as DNS Challenges where background noise was the focus of the challenge. A majority (82%) of the clips in this dataset were found to have lower signal quality than background noise (SIG < BAK), whereas this number was below 30% in the last DNS challenges. Given that the minimum of signal and background quality is a strong determinant of perceived overall quality [1], the observed high correlation between signal and overall quality in this dataset was expected.

**TABLE 14.** **Correlations Between Subjective Scores Obtained From P.804 and P.835 Subjective Tests on Shared Dimensions in Model Level for All Entries**

| Dimension | PCC | SRCC | Kendall Tau-b | Tau-b95 |
|---|---|---|---|---|
| Background/Noisiness | 0.964 | 0.926 | 0.825 | 0.853 |
| Signal | 0.954 | 0.933 | 0.801 | 0.914 |
| Overall | 0.965 | 0.940 | 0.825 | 0.822 |
| M (challenge metric) | 0.961 | 0.946 | 0.825 | - |

Tau-B95 is Kendall Tau-B Applied to Corrected Ranked-Order by Considering 95% Confidence Interval of Subjective Scores According to [41].

## C. CORRELATION BETWEEN P.804 AND P.835

The correlation between quality scores collected using P.804- and P.835-based subjective tests, for all entries are reported in Table 14. We observed a strong correlation between all the shared scores between the two subjective methodologies. Considering the rankings of participating teams, only the rank of N&B and Kuaishou teams from the real-time track would

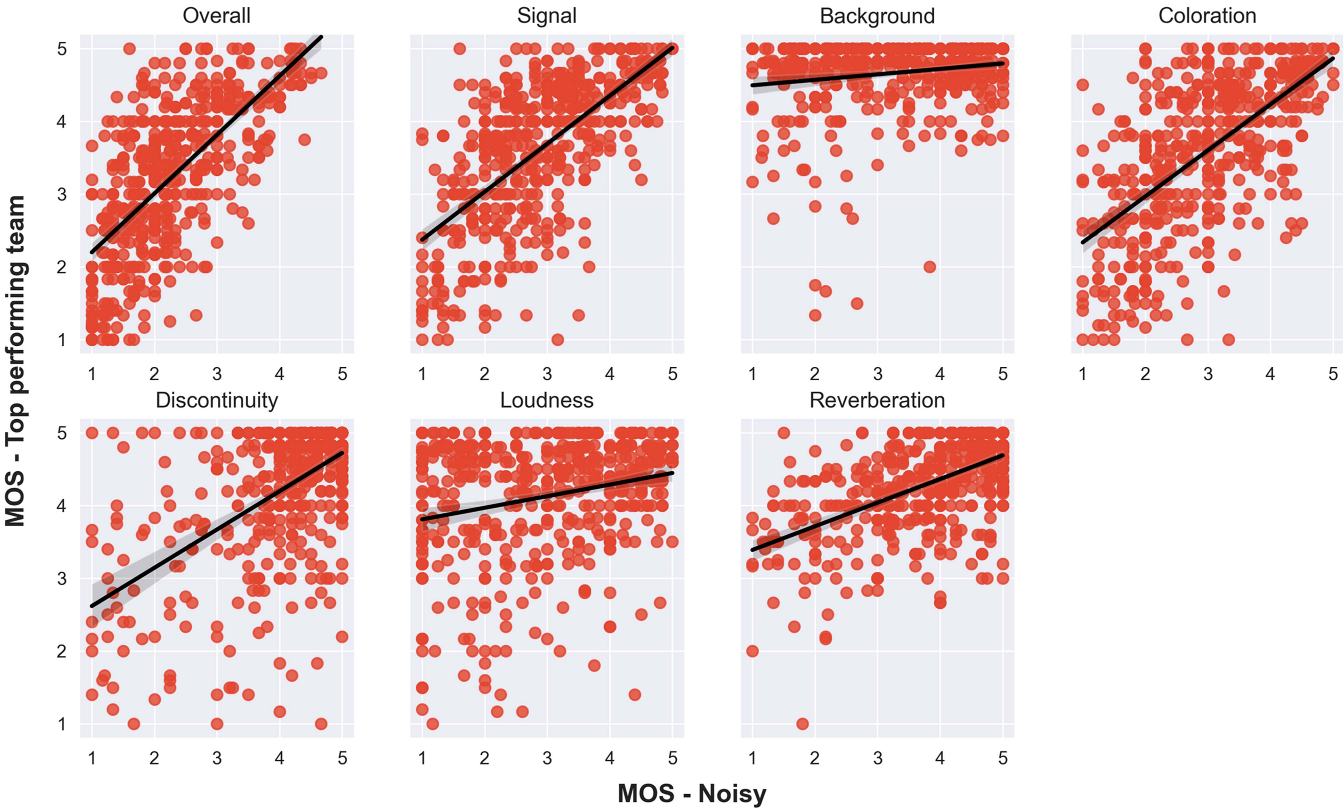


**FIGURE 3.** Distribution of subjective ratings before (X-axis) and after applying the winning model by Legends-tencent [35] (Y-axis). Each dot is a single audio clip, and a best-fit line is shown. No processing would be a diagonal line from (1,1) to (5,5). Background is close to ideal, while loudness degrades excellent loudness (MOS = 5) inputs.

**TABLE 15.** The PCC Between the Subjective P.804 Results and the Objective Metrics Estimated With DNSMOS P.835 [18] and NISQA [20] Models

| Subjective metric | Objective metric | PCC | |
|---|---|---|---|
| | | Clip level | Model level |
| P.804 Overall | DNSMOS P.835 OVRL | 0.695 | 0.884 |
| P.804 Overall | NISQA MOS | 0.681 | 0.766 |
| P.804 Signal | DNSMOS P.835 SIG | 0.656 | 0.799 |
| P.804 Noisiness | DNSMOS P.835 BAK | 0.545 | 0.933 |
| P.804 Noisiness | NISQA NOISE | 0.586 | 0.938 |
| P.804 Coloration | NISQA COLOR | 0.663 | 0.872 |
| P.804 Discontinuity | NISQA DISCONTINUITY | 0.478 | 0.310 |
| P.804 Loudness | NISQA LOUDNESS | 0.700 | 0.784 |

**TABLE 16.** The Loading of Quality Scores on Three-Factor Structure Using Maximum Likelihood Extraction Method With Varimax Rotation. KMO Value = 0.65

| Quality score | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| Signal | 0.824 | 0.481 | |
| Noisiness | | | 0.742 |
| Coloration | 0.787 | | |
| Discontinuity | | 0.936 | |
| Loudness | 0.476 | | |
| Reverberation | | | 0.413 |

KMO Value = 0.65. Factor Loading > 0.3 is Presented.

swap when scores from P.835 test are used (tied rank using P.804 ratings).

### D. CORRELATION OF SUBJECTIVE AND OBJECTIVE DATA

In Table 11 we present the objective results on the blind set using DNSMOS [18] (MOS _ SIG, MOS _ BAK, MOS _ OVR), and NISQA model [20] (NISQA _ MOS, etc.). Similar to the subjective results, the Legends-tencent, Ctv-tencent, and Genius-team teams attained the best metrics estimated with DNSMOS and NISQA. Moreover, in Table 15 we compute the PCC between the subjective P.804 metrics and the metrics obtained with DNSMOS [18] and NISQA [20]. The correlations range from PCC 0.478 to 0.700, which demonstrates why we still require a subjective test for accurately evaluating speech quality.

### E. MODEL OF OVERALL AND OTHER DIMENSIONS

We performed Explanatory Factory Analysis (EFA) [42] to investigate the underlying structure between the quality dimensions, namely if there is a shared variance between the sub-dimensions. We used the Maximum Likelihood extraction method with Varimax rotation and extracted three factors as suggested by the Scree plot [43]. The result of Bartlett's

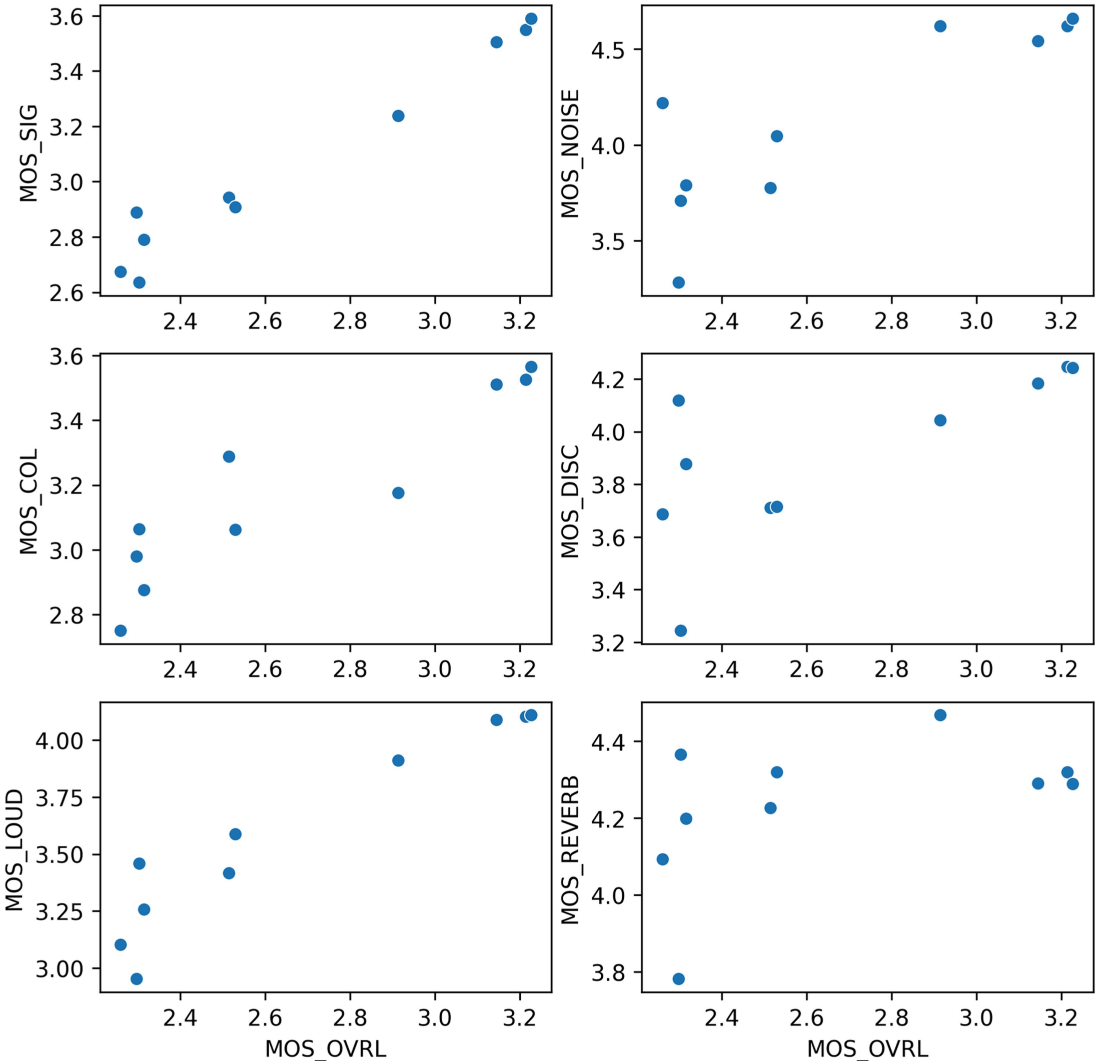


**FIGURE 4.** Distribution of subjective test dimensions for all entries in model level.

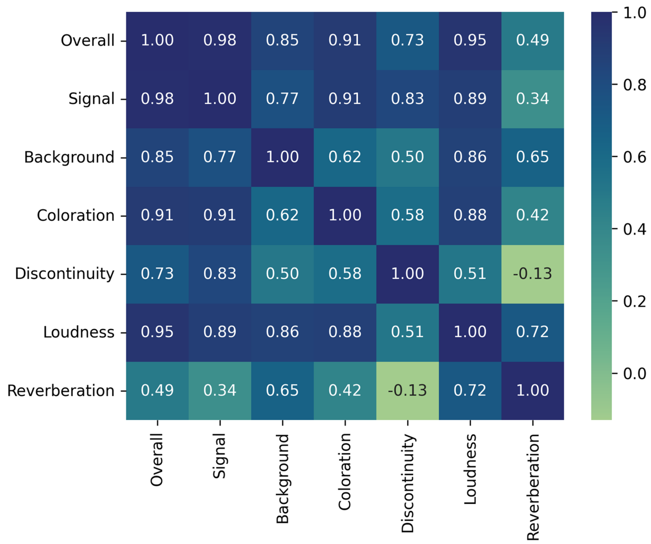


**FIGURE 5.** Pearson correlation between different subjective test dimensions for all entries in model level.

**TABLE 17.** Average Performance of Different Regressors Predicting Overall Quality Given the Six Sub-Dimensions in 5-Fold Cross-Validation

| | Clip level | | | Model level | | |
|---|---|---|---|---|---|---|
| Regressor | PCC | RMSE | $R^2$ | PCC | RMSE | $R^2$ |
| Linear regression | 0.947 | 0.318 | 0.894 | 0.993 | 0.051 | 0.951 |
| Polynomial (n = 4) | 0.959 | 0.276 | 0.920 | 0.996 | 0.047 | 0.969 |
| Random forest | 0.960 | 0.276 | 0.921 | 0.977 | 0.131 | 0.754 |

In the Model, Level 3-Fold Cross-Validation is Used.

test of sphericity was significant and the KMO value was 0.65 indicating that the data is adequate for explanatory factor analysis. The loading of quality scores on each factor is presented in Table 16. In total 62% of the variance in data is explained by the three factors. Factor 1 represents the coloration with high loading from signal, coloration, and loudness. Discontinuity is loaded on factor 2 with some cross-loading from the signal indicating no or limited shared variance between discontinuity ratings and both coloration

**TABLE 18. Average Coefficient and Importance of Features in Linear Regression and Random Forest Models Predicting Overall Quality, Respectively**

| Feature | Clip level Linear regression coefficient | Clip level Random forest features imp. | Model level Linear regression coefficient | Model level Random forest features imp. |
|---|---|---|---|---|
| Signal | 0.646 | 0.878 | 0.352 | 0.378 |
| Loudness | 0.146 | 0.044 | 0.251 | 0.162 |
| Coloration | 0.102 | 0.019 | 0.266 | 0.146 |
| Noisiness | 0.100 | 0.027 | 0.134 | 0.248 |
| Discontinuity | 0.065 | 0.016 | 0.190 | 0.051 |
| Reverberation | 0.039 | 0.016 | 0.014 | 0.016 |

**TABLE 19. Average Performance of Different Regressors Predicting Signal Quality Given the Five Sub-Dimensions in 5-Fold Cross-Validation**

| Regressor | Clip level PCC | Clip level RMSE | Clip level $R^2$ | Model level PCC | Model level RMSE | Model level $R^2$ |
|---|---|---|---|---|---|---|
| Linear regression | 0.898 | 0.453 | 0.806 | 0.994 | 0.035 | 0.979 |
| Polynomial (n=4) | 0.907 | 0.434 | 0.821 | 0.992 | 0.043 | 0.964 |
| Random forest | 0.901 | 0.446 | 0.812 | 0.852 | 0.170 | 0.452 |

In the Model, Level 3-Fold Cross-Validation is Used.

**TABLE 20. Average Coefficient and Importance of Features in Linear Regression and Random Forest Models Predicting Signal Quality, Respectively**

| Feature | Clip level Linear regression coefficient | Clip level Random forest features imp. | Model level Linear regression coefficient | Model level Random forest features imp. |
|---|---|---|---|---|
| Loudness | 0.128 | 0.061 | 0.184 | 0.142 |
| Coloration | 0.503 | 0.559 | 0.519 | 0.373 |
| Noisiness | 0.072 | 0.044 | 0.051 | 0.306 |
| Discontinuity | 0.430 | 0.281 | 0.580 | 0.130 |
| Reverberation | 0.099 | 0.054 | 0.158 | 0.049 |

**TABLE 21. Real-Time Track Word Error Rate Challenge Results for the Blind Test Set**

| Team | English | | German | | Dutch | | Spanish | | French | | Average | P.804 Ranking |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #samples | WER | #samples | WER | #samples | WER | #samples | WER | #samples | WER | | |
| Noisy | 395 | 13.82 | 72 | 8.36 | 18 | 10.08 | 10 | 16.46 | 5 | 41.86 | 13.23 | #5 |
| Legends-tencent | 395 | 14.30 | 72 | 8.28 | 18 | 17.62 | 10 | 21.52 | 5 | 39.54 | 13.95 | #1 |
| Njuaa-lab | 395 | 14.26 | 72 | 11.72 | 18 | 11.93 | 10 | 15.19 | 5 | 37.21 | 14.06 | #6 |
| Ctv-tencent | 395 | 14.16 | 72 | 15.67 | 18 | 17.01 | 10 | 21.52 | 5 | 48.84 | 14.98 | #2 |
| Genius-team | 395 | 14.24 | 72 | 17.47 | 18 | 21.42 | 10 | 21.52 | 5 | 46.51 | 15.43 | #3 |
| Hitiot | 395 | 14.49 | 72 | 22.26 | 18 | 25.98 | 10 | 31.65 | 5 | 55.81 | 16.78 | #4 |
| Kuaishou | 395 | 16.44 | 72 | 26.87 | 18 | 30.13 | 10 | 30.38 | 5 | 79.07 | 19.34 | #8 |
| N&B | 395 | 17.57 | 72 | 43.20 | 18 | 54.18 | 10 | 50.63 | 5 | 60.47 | 23.67 | #7 |

and loudness. As expected, noisiness built a separate factor orthogonal to others with moderate loading from reverberation. All in all, the results of EFA show that coloration, discontinuity, and noisiness are loaded on different orthogonal factors that align with the literature [44]. Signal scores share variance with coloration, discontinuity, and loudness, whereas reverberation shares variance with noisiness. Note that this factor structure represents the construct of the current training set and its generalizability should be validated in a separate study.

In addition, we used different regressors to predict the overall quality given the subjective scores of the six sub-dimensions per clip. The results of k-fold cross-validations for clip and model levels are reported in Table 17. Given that only a limited number of models are available in the dataset, random forest performed poorly compared to other regressors at the model level. The coefficients of the linear regression model and the feature importance from the random forest model are reported in Table 18. At the clip level, the importance of features mostly agrees with both models. Given the fact that most of the sub-dimensions have cross-loading with the signal quality in the explanatory factor analyses, we created different regressors to predict that. The performance of those regressors is reported in Table 19 and the coefficients in Table 20. As expected, noisiness and reverberation have the smallest coefficients.

**TABLE 22.** Non-Real-Time Track Word Error Rate Challenge Results for the Blind Test Set

| Team | English | | German | | Dutch | | Spanish | | French | | Average | P.804 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | #samples | WER | #samples | WER | #samples | WER | #samples | WER | #samples | WER | | Ranking |
| Noisy | 395 | 13.82 | 72 | 8.36 | 18 | 10.08 | 10 | 16.46 | 5 | 41.86 | 13.23 | #6 |
| Legends-tencent* | 395 | 14.30 | 72 | 8.28 | 18 | 17.62 | 10 | 21.52 | 5 | 39.54 | 13.95 | #1 |
| Ctv-tencent* | 395 | 14.16 | 72 | 15.67 | 18 | 17.01 | 10 | 21.52 | 5 | 48.84 | 14.98 | #2 |
| Genius-team* | 395 | 14.24 | 72 | 17.47 | 18 | 21.42 | 10 | 21.52 | 5 | 46.51 | 15.43 | #3 |
| N&B | 395 | 16.15 | 72 | 30.00 | 18 | 49.89 | 10 | 30.38 | 5 | 65.12 | 20.13 | #4 |
| Hamburg | 395 | 20.65 | 72 | 36.42 | 18 | 22.51 | 10 | 31.94 | 5 | 76.74 | 23.77 | #5 |

The Teams Marked With * Have Identical Submissions for Both Tracks.

**TABLE 23.** Amount of Improvement Remaining (In MOS) to Get Excellent Quality Rated Speech Based on This Challenge

| Area | Headroom |
|---|---|
| Overall | 1.73 |
| Signal | 1.74 |
| Background | 0.36 |
| Coloration | 1.40 |
| Loudness | 0.80 |
| Discontinuity | 0.89 |
| Reverberation | 0.68 |

### F. WORD ERROR RATE

To have a more comprehensive view of the signal enhancement models, in Tables 21 and 22 we included the word error rate (WER) for both tracks. To eliminate potential bias introduced by automatic speech recognition (ASR) systems, we employed human transcripts when calculating the WER. A state-of-the-art speech recognition API from Azure Cognitive service was used for computing WER. In the second track, the rank is identical to the P.804 ranking (excluding noisy), while in the first track, there are some shifts between teams. The best WER result attained by Legends-tencent team is still slightly behind the WER computed on the noisy files, highlighting that there is a huge potential in this research area.

## IX. CONCLUSION

Unlike our previous deep noise suppression challenges, this challenge showed several models with significant improvement in the speech signal. The top models improved all areas we measured: noisiness, discontinuity, coloration, loudness, and reverberation. While the improvements are impressive, there is still significant room for improvement in this test set (see Table 23).

All of the models used in this challenge are relatively small compared to large language models or large multimodal language models. An interesting new area would be to apply a large audio language model (e.g., [45]) for speech restoration and enhancement. Even if it can not be run in real-time or with low latency, there are still many scenarios it can be applied. In addition, all of the models submitted in this challenge used training sets with clean speech available. A good future direction of research is to utilize real-world training sets such as [27], which will require semi-supervised or unsupervised learning.

For future speech signal improvement challenges, we plan to provide an objective metric similar to NISQA. We plan to also add word accuracy rate as an additional metric to optimize. We plan to provide a synthetic data generator and a baseline model to give a better starting point for all participants. As noted above, we hypothesize that large multimodal models could have significant improvements in this area, so keeping a non-real-time track seems important to encourage this exploration.

## REFERENCES


[1] ITU-T Recommendation P.835: Subjective Test Methodology for Evaluating Speech Communication Systems That Include Noise Suppression Algorithm, 2003.

[2] H. Dubey et al., “ICASSP 2022 deep noise suppression challenge,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022.

[3] C. K. Reddy et al., “INTERSPEECH 2021 deep noise suppression challenge,” in *Proc. InterSpeech*. 2021, pp. 2796–2800.

[4] H. Dubey et al., “ICASSP 2023 deep noise suppression challenge,” 2023, *arXiv:2303.11510*.

[5] ITU-T Recommendation P.804: Subjective Diagnostic Test Method for Conversational Speech Quality Analysis, 2017.

[6] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features,” in *Proc. IEEE Workshop Appl. Signal Process. to Audio Acoust.*, 2021, pp. 166–170.

[7] J. Serrá, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, “Universal speech enhancement with score-based diffusion,” 2022. [Online]. Available: https://openreview.net/forum?id=7BfWbjOqgMf

[8] “ITU-T Recommendation P.800: Methods for subjective determination of transmission quality,” 1996.

[9] “ITU-T Recomendation P.808: Subjective evaluation of speech quality with a crowdsourcing approach,” 2018.

[10] B. Naderi and R. Cutler, “An open source implementation of ITU-T recommendation P.808 with validation,” in *Proc. InterSpeech*, 2020, pp. 2862–2866.

[11] B. Naderi and R. Cutler, “Subjective evaluation of noise suppression algorithms in crowdsourcing,” in *Proc. InterSpeech*, 2021.

[12] “ITU-T Recommendation P.863.2: Extension of ITU-T P.863 for multi-dimensional assessment of degradations in telephony speech signals up to full-band,” 2022.

[13] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2001, vol. 2, pp. 749–752.

[14] J. G. Beerends, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, “Perceptual objective listening quality assessment (POLQA), The third generation ITU-T standard for end-to-end speech quality measurement Part I–Temporal alignment,” *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384, 2013.

[15] "ITU-T Recommendation P.563: 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," 2011.

[16] 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.*, 2019, pp. 631–635.

[17] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *InterSpeech*, 2021, pp. 6493–6497.

[18] 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.

[19] G. Yi et al., "ConferencingSpeech 2022 challenge: Non-intrusive objective speech quality assessment (NISQA) challenge for online conferencing applications," in *Proc. InterSpeech*, 2022.

[20] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. InterSpeech*, 2021.

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

[22] 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.

[23] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "INTERSPEECH 2022 audio deep packet loss concealment challenge," in *Proc. InterSpeech*, 2022, p. 5.

[24] K. Kinoshita et al., "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 7, Dec. 2016.

[25] 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.

[26] R. Cutler et al., "INTERSPEECH 2021 acoustic echo cancellation challenge," in *Proc. InterSpeech*, 2021.

[27] R. Cutler et al., "ICASSP 2022 acoustic echo cancellation challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022.

[28] R. Cutler et al., "ICASSP 2023 acoustic echo cancellation challenge," *IEEE Open J. Signal Process.*, early access, Mar. 13, 2024, doi: 10.1109/OJSP.2024.3376289.

[29] H. Li and J. Yamagishi, "DDS: A new device-degraded speech dataset for speech enhancement," in *Proc. InterSpeech*, 2022.

[30] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?–A dataset, insights, and challenges," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1006–1010, Aug. 2015.

[31] B. Naderi, R. Cutler, and N.-C. Ristea, "Multi-dimensional speech quality assessment in crowdsourcing," Sep. 2023, *arXiv:2309.07385[cs, eess]*.

[32] B. Naderi, R. Z. Jiménez, M. Hirth, S. Möller, F. Metzger, and T. Hoßfeld, "Towards speech quality assessment using a crowdsourcing approach: Evaluation of standardized methods," *Qual. User Experience*, vol. 6, no. 1, Nov. 2020, Art. no. 2.

[33] B. Naderi and S. Möller, "Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task," in *Proc. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–6.

[34] B. Naderi, T. Polzehl, I. Wechsung, F. Köster, and S. Möller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," in *Proc. InterSpeech*, 2015.

[35] J. Chen et al., "Gesper: A unified framework for general speech restoration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[36] W. Zhu, Z. Wang, J. Lin, C. Zeng, and T. Yu, "SSI-NET: A multi-stage speech signal improvement system for ICASSP 2023 SSI challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[37] Z. Zhang, S. Xu, X. Zhuang, Y. Qian, L. Zhou, and M. Wang, "Half-temporal and half-frequency attention $U^2$Net for speech signal improvement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[38] M. Liu et al., "Two-stage neural network for ICASSP 2023 speech signal improvement challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[39] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, T. Peer, and T. Gerkmann, "Speech signal improvement using causal generative diffusion models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.

[40] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019, doi: 10.7488/ds/2645.

[41] B. Naderi and S. Möller, "Transformation of mean opinion scores to avoid misleading of ranked based statistical techniques," in *Proc. 12th Int. Conf. Qual. Multimedia Experience*, 2020, pp. 1–4.

[42] M. W. Watkins, "Exploratory factor analysis: A guide to best practice," *J. Black Psychol.*, vol. 44, no. 3, pp. 219–246, Apr. 2018.

[43] R. B. Cattell, "The scree test for the number of factors," *Multivariate Behav. Res.*, vol. 1, no. 2, pp. 245–276, Apr. 1966, doi: 10.1207/s15327906mbr010210.

[44] M. Wältermann, *Dimension-Based Quality Modeling of Transmitted Speech*. New York, NY, USA: Springer, Jan. 2013.

[45] 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.