



## Interspeech 2025 URGENT Speech Enhancement Challenge

Kohei Saijo<sup>1</sup>, Wangyou Zhang<sup>2</sup>, Samuele Cornell<sup>3</sup>, Robin Scheibler<sup>4</sup>, Chenda Li<sup>2</sup>, Zhaoheng Ni<sup>5</sup>, Anurag Kumar<sup>5</sup>, Marvin Sach<sup>6</sup>, Yihui Fu<sup>6</sup>, Wei Wang<sup>2</sup>, Tim Fingscheidt<sup>6</sup>, Shinji Watanabe<sup>3</sup>

<sup>1</sup>Waseda University, Japan <sup>2</sup>Shanghai Jiao Tong University, China <sup>3</sup>Carnegie Mellon University, USA  
<sup>4</sup>Google DeepMind, Japan <sup>5</sup>Meta, USA <sup>6</sup>Technische Universität Braunschweig, Germany

saijo@pcl.cs.waseda.ac.jp

### Abstract

There has been a growing effort to develop universal speech enhancement (SE) to handle inputs with various speech distortions and recording conditions. The URGENT Challenge series aims to foster such universal SE by embracing a broad range of distortion types, increasing data diversity, and incorporating extensive evaluation metrics. This work introduces the Interspeech 2025 URGENT Challenge, the second edition of the series, to explore several aspects that have received limited attention so far: language dependency, universality for more distortion types, data scalability, and the effectiveness of using noisy training data. We received 32 submissions, where the best system uses a discriminative model, while most other competitive ones are hybrid methods. Analysis reveals some key findings: (i) some generative or hybrid approaches are preferred in subjective evaluations over the top discriminative model, and (ii) purely generative SE models can exhibit language dependency.

**Index Terms:** URGENT challenge, speech enhancement, multilingual, scalability

### 1. Introduction

Speech enhancement (SE) aims to improve speech quality degraded by noise, reverberation, or other distortions. It has seen significant progress thanks to neural networks. However, most prior work has focused on matched training–inference conditions and limited tasks such as noise suppression and dereverberation.

In recent years, the advent of generative models has broadened the scope of SE to include more complex tasks such as bandwidth extension, packet loss concealment [1], and wind noise reduction [2]. Following this trend, there has been growing interest in developing universal SE models capable of handling multiple tasks and diverse input formats within a single framework. For example, score-based diffusion models have been explored for universal SE [3, 4], enabling a single model to address multiple SE tasks. To support diverse input formats, models that process multiple sampling rates and varying numbers of microphones have also been proposed [5]. However, despite these advances, a well-established benchmark for evaluating generative and universal SE models remains lacking, making fair comparisons challenging.

To address this gap, the URGENT Challenge (Universality, Robustness, and Generalizability of speech Enhancem<sup>EN</sup>T) was launched. The first edition, the URGENT 2024 Challenge [6], aimed to lay the foundation for developing universal SE models by setting two key objectives: (i) handling four types of distortion (additive noise, reverberation, bandwidth limitation, and clipping) and (ii) accommodating input signals with varying sampling rates. The challenge utilized a collection of publicly available datasets, significantly larger than those used in most

prior studies, and employed a comprehensive evaluation framework with as many as 13 evaluation metrics.

To further advance the development of universal SE models, we propose the second edition of the series, Interspeech 2025 URGENT Challenge. This iteration introduces the following modifications based on our preliminary investigations [7]:

- **More data diversity:** There is room for improvement in robustness on, e.g., samples with unseen or rarely seen noise. To address this, the proposed challenge by incorporating more diverse speech and noise data.
- **Leveraging noisy data:** While noisy data is crucial for scaling up the amount of data due to the scarcity of clean data, we found that simple filtering methods (e.g., using DNSMOS scores [8]) did not remove noisy samples very effectively. We intentionally include more noisy data to encourage participants to explore effective ways of utilizing it.
- **Two tracks with different training data scales:** To examine the impact of data scale, the proposed challenge introduces two tracks with different training data sizes: one with  $\sim 2.5k$  hours and the other with  $\sim 60k$  hours of speech..
- **More distortions:** We found that SE models struggle to generalize to unseen distortions. To enhance generalizability, we consider three additional distortions, commonly observed in real recordings due to the recording device or environment.
- **Multilingual data:** While the first challenge used only English data, the proposed challenge incorporates multilingual data, allowing us to examine the accessibility of SE models and evaluation metrics across the languages.

The challenge has received 22 submissions for one track and 10 for the other. This paper details the challenge design and presents a preliminary analysis of the submissions. The details of the challenge are also available on our website<sup>1</sup>.

### 2. Related works

Existing SE challenges have greatly advanced SE studies for specific scenarios, such as denoising and dereverberation [9, 10], speech restoration [11, 12], packet loss concealment [1], and so on. The URGENT challenge complements existing ones by focusing on universality, generalizability, and robustness across diverse scenarios and evaluation metrics.

Most previous SE studies use only monolingual data, and thus language dependency of SE models is still under-explored. While the language dependency of discriminative SE has been reported to be small [13], that of generative SE lacks thorough investigation. Additionally, although often overlooked, some

<sup>1</sup><https://urgent-challenge.github.io/urgent2025/>

metrics can be language dependent [14, 15]. Through this challenge, we collect the objective and subjective evaluation scores of various SE systems in multiple languages, which will later be analyzed to assess the language dependency of metrics<sup>2</sup>.

One of the focuses of this challenge is how to leverage possibly noisy data. While there are some possible approaches such as data filtering based on non-intrusive SE metrics [8, 16, 17] or unsupervised learning [18, 19], a comprehensive comparison of them is currently lacking, particularly on large-scale data. The proposed challenge aims to clarify which of these methods are more effective, or if new ones can be devised, by encouraging participants to explore how to leverage noisy data.

This challenge also explores the impact of training data size on final performance. While this has been investigated in other fields [20, 21], to our best knowledge, only an investigation on small-scale data has been done in SE field [22]. The proposed challenge offers two tracks with different data amounts, encouraging the investigation of data scalability.

### 3. Challenge Design

#### 3.1. Task definition

In the URGENT challenge, the SE process is defined as [6]:

$$\hat{\mathbf{x}} = \text{SE}(\mathcal{F}(\mathbf{x})), \quad (1)$$

where  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  are the desired and enhanced speech signals.  $\mathcal{F}(\cdot)$  is the distortion model that degrades the desired signal. Our definition differs from the commonly adopted one in the literature in that the SE model has to handle (i) inputs with *various sampling frequencies (SF)* (8, 16, 22.05, 24, 32, 44.1, and 48kHz), while most existing SE systems often only consider a fixed SF, (ii) inputs with *seven types of distortions* (detailed in Section 3.2), while most existing SE systems consider only noise or reverberation, and (iii) *multilingual speech*, while most existing SE studies consider only monolingual data.

#### 3.2. Distortions

The challenge considers the following seven distortions in the distortion model  $\mathcal{F}$ : additive noise, reverberation, clipping, bandwidth limitation, codec loss (MP3 and OGG), packet loss, and wind noise<sup>3</sup>. Compared to the first challenge, which included only the first four distortions [6], the present challenge introduces three more challenging but realistic conditions. Each noisy speech is degraded by up to five distortions simultaneously.

#### 3.3. Data

##### 3.3.1. Training data

A collection of public corpora listed in Table 1 is used as training data. Participants can freely simulate noisy speech using these data. It is also allowed to freely simulate distortions mentioned in Section 3.2. However, to allow a fair comparison, using datasets other than those listed in Table 1 is prohibited.

As some speech corpora include noisy speech, we preprocessed them following the methodology of the first challenge [6]. Specifically, for DNS5 Challenge data and CommonVoice, we

<sup>2</sup>We were able to collect the subjective listening scores only on English data during the challenge period due to the Interspeech challenge time constraint. We plan to collect data for other languages and analyze the metrics after the challenge concludes.

<sup>3</sup>We used the wind-noise simulator provided in [2]: <https://github.com/sp-uhh/storm>, which is based on <https://github.com/audiolabs/SC-Wind-Noise-Generator>

<sup>4</sup>This research did not train on any music, except to the extent that some music data from the FSD50K and FMA datasets was processed as background noise for speech enhancement purpose.

**Table 1:** Corpora used for training/validation/non-blind testing, where shaded cells are used only for non-blind testing<sup>4</sup>. Data amounts of MLS and CommonVoice are limited in the first track but not in the second.  $\diamond$ : We crawled the high-quality (less compressed) version of MLS (MLS-HQ) from LibriVox.

Type	Corpus	Duration
Speech	LibriVox data from DNS5 Challenge (en) [10]	~350 h
	LibriTTS reading speech (en) [23]	~200 h
	VCTK reading speech (en) [24]	~80 h
	WSJ reading speech (en) [25, 26]	~85 h
	EARS expressive speech (en) [27]	~107 h
	MLS-HQ <sup>5</sup> (en, de, fr, es) [28]	~450 (48600) h
	CommonVoice 19.0 (en, de, fr, es, zn) [29]	~1300 (9500) h
Noise	Noise from DNS5 Challenge [10]	~180 h
	WHAM! noise [30]	~70 h
	FSD50K [31]	~100 h
	Free Music Archive (FMA, Medium partition) [32]	~200 h
	Wind noise simulated by participants	-
	DEMAND [33]	-
RIR	TUT Urban Acoustic Scenes 2018 [34]	-
	Simulated RIRs from DNS5 Challenge	~60k samples
	Other RIRs simulated by participants	-
	SLR28 real RIRs	-
	BRUDEX [35]	-
	MYRIAD V2 econ [36]	-

(i) resampled the speech to the lowest SF that can fully cover the effective frequency range, (ii) filtered out non-speech or samples dominated by silence using voice activity detection (VAD)<sup>5</sup>, and (iii) filtered out noisy samples based on the DNSMOS score [8]. We also applied the resampling process to LibriTTS, MLS-HQ, DNS noise, FSD50K, and FMA.

Note that the VAD- and DNSMOS-based data filtering mentioned above is imperfect. Indeed, we found many samples with audible background noise still remaining after the data filtering. Although such noisy samples may adversely affect SE model training, we intentionally keep such samples to encourage participants to explore how to leverage noisy samples effectively.

To investigate the language dependency of SE models, we use two multilingual corpora, MLS-HQ<sup>6</sup> and CommonVoice 19.0 which cover five languages in total (en, de, fr, es, and zn).

##### 3.3.2. Validation/non-blind test data

In the challenge, it is allowed to freely simulate participants' own validation sets using the validation portion of the corpora in Table 1 to select the system to submit. However, we provide the official validation set containing 1000 noisy speech to allow a fair comparison. When making the official validation set, since many CommonVoice samples are noisy, we manually selected only clean ones. Samples from the other datasets are randomly selected after applying the data filtering mentioned in Section 3.3.1. The non-blind test set is made in a similar manner as the official validation set using the test portion of the corpora in Table 1. However, we add some noise and RIRs shown in the shaded cells of Table 1 to gauge robustness against unseen data.

##### 3.3.3. Blind test data

The blind test set, used to determine the final ranking, is primarily from domains unseen during training to assess the robustness and generalizability. Specifically, the blind test set

- Consists of 50% simulated data and 50% real recordings, both sourced from publicly available datasets. All synthetic data and most real recordings come from datasets not listed in

<sup>5</sup><https://github.com/wiseman/py-webrtcvad>

<sup>6</sup>We prepared the high-quality version of MLS (MLS-HQ) by crawling less-compressed speech from LibriVox (<https://librivox.org>) and segmenting them using the same timestamps as MLS.

Table 1. Since real recordings with certain distortions (e.g., packet loss) are scarce, we artificially apply clipping, packet loss, bandwidth limitations, and codec loss.

- Includes Japanese data as an unseen language.
- Has 150 samples per language (totaling 900 samples).
- Considers also unseen distortions due to neural audio codecs, specifically Encodec [37] and Descript Audio Codec [38], as codec lossy distortion in addition to MP3 and OGG.

### 3.4. Two tracks with different data scales

We provide two tracks with different training data scales but the same test set:

- **Track 1:** We limit the duration of MLS and CommonVoice, resulting in  $\sim 2.5$ k hours of speech.
- **Track 2:** We do not limit the duration of MLS and CommonVoice datasets, resulting in  $\sim 60$ k hours of speech.

This design allows us to explore (i) the impact of data scale and (ii) data-hungry methods such as large-scale self-supervised-learning-based methods.

### 3.5. Evaluation metrics and ranking strategy

We use multiple evaluation metrics to perform a more comprehensive evaluation. For instance, the hallucination of a generative model could be hard to detect by non-intrusive metrics but is easily penalized by other metrics<sup>7</sup>. We adopt the following 14 metrics categorized into five categories:

- 1) **Non-intrusive SE metrics:** DNSMOS [8], NISQA [16], and UTMOS [17].
- 2) **Intrusive SE metrics:** perceptual objective listening quality assessment (POLQA) [39], perceptual evaluation of speech quality (PESQ) [40], extended short-time objective intelligibility (ESTOI) [41], signal-to-distortion ratio (SDR) [42], mel cepstral distortion (MCD) [43], and log-spectral distance (LSD) [44].
- 3) **Downstream-task-independent metrics:** Levenshtein phone similarity (LPS) [45] and SpeechBERTScore (SBS) [46] with mHuBERT-147 [47].
- 4) **Downstream-task-dependent metrics:** speaker similarity (SpkSim) with RawNet3 [48] and character accuracy (CAcc, 1 - character error rate) with OWSM v3.1 [49].
- 5) **Subjective metric:** absolute category rating mean opinion score (MOS) via ITU-T P.808 test [50, 51], implemented on the Amazon Mechanical Turk (on only English data<sup>2</sup>).

Compared to the first challenge [6], we made the following changes to support multilingual data: (i) we use an SSL model trained on multilingual data, mHuBERT-147, in SBS, and (ii) we evaluate CAcc instead of word accuracy. In addition, UTMOS is included, as it showed a high correlation with MOS scores in our preliminary analysis [7]. POLQA and MOS were evaluated only in the blind-testing phase. The final ranking is obtained by aggregating the scores of the 14 metrics mentioned above with the following three steps inspired by the Friedman test [52]:

1. Calculate the ranking for each metric.
2. Average the rankings of the metrics in each category.
3. Average the category-wise rankings obtained in Step 2.

The overall ranking scores obtained in Step 3 are used to determine the final ranking.

<sup>7</sup>Here, "hallucination" refers to discrepancies in spoken content or speaker characteristics between the noisy speech and enhanced speech.

### 3.6. Baseline system

We train the TF-GridNet model [53] with around 8.5M parameters on Track1 training data and provide it as the baseline system. We initialize the model with the pre-trained weights provided in the first challenge to save training time. Training takes around 2.5 days using a single NVIDIA RTX A6000 GPU. The training pipeline is based on ESPnet-SE library [54]. Please refer to the URGENT challenge website<sup>1</sup> for more details.

## 4. Results

In this section, we report the evaluation results of the submitted systems. We also provide some preliminary analysis based on the brief system descriptions we collected from participants.

### 4.1. Overall results

Table 2 shows the evaluation results, where 12 out of 22 systems submitted to Track1 (T\*) and 4 out of 10 systems submitted to Track2 (S\*) are shown due to space limitations. The full results are available on our leaderboard<sup>8</sup>. D, G, and D+G in 'Model type' cell denote discriminative, generative, and hybrid (e.g., discriminative model with adversarial loss or cascade of discriminative and generative models), respectively. T10 and T22 are the baseline system and noisy audios, respectively.

Focusing on the systems that outperformed the baseline (T1-T9), we observed that most were hybrid systems. In particular, the most commonly used approach was to optimize a model with both discriminative losses and an adversarial loss (T2, T4, T5, T6, and T9). These systems can leverage the strong denoising and dereverberation capabilities of well-established discriminative models while also benefitting from generative loss to tackle distortions that are particularly well handled by generative models (e.g., bandwidth limitation, packet loss, etc.). T7 employed a similar framework but used denoising score-matching loss. T3, which showed a strong performance on all the metrics, was a cascade of D, G, and D systems where the second model is a discrete-token-based generative model. However, interestingly, the top system T1 was a purely discriminative model based on a sub-band recurrent neural network (RNN). Although the baseline system T10 has a similar architecture (based on full- and sub-band RNNs), notably, T1 has  $\sim 102$ M parameters, which is much larger than that of T10 ( $\sim 8.5$ M). The larger model may have benefited since the dataset was large.

Although larger-scale training data was available in Track2, no systems in Track2 outperform the best system in Track1, which suggests that scaling up the data does not necessarily lead to better performance in SE when the additional data contain some noisy speech.

### 4.2. Language dependency of SE models

To analyze the language dependency of SE models, we list the DNSMOS and CAcc scores for several systems by language in Table 3. Based on the results, the discriminative models (T1 and T10) appear to be relatively insensitive to language variations. In contrast, the purely generative model based on latent diffusion and vocoding (T13) demonstrates a high degree of language dependency, with its performance markedly declining when applied to an unseen language, namely Japanese. Compared with T13, the hybrid approaches (T2 and T3) exhibit lower language dependency. However, while T3, a cascade system of discriminative and purely generative models, achieves higher CAcc than

<sup>8</sup><https://urgent-challenge.com/competitions/13>

**Table 2: Results of Track1 ( $T^*$ ) and Track2 ( $T^*$ ), where the same numbers in IDs indicate systems submitted by the same team. Only 12 out of 22 submissions to Track1 and 4 out of 10 submissions to Track2 are shown due to space limitation. † in Track2 denotes the same submission as Track1. T10 and T22 are baseline and noisy input, respectively. D, G, and D+G in ‘Model type’ cell denote discriminative, generative, and hybrid of them, respectively. Numbers next to metric scores are ranking in each metric in each track. Note that MOS evaluation is done only for English data.**

Rank	ID	Model type	Non-intrusive SE metrics			Intrusive SE metrics				Down.-task-indep.		Down.-task-dep.		Subject. MOS <sup>†</sup>	Ranking score <sup>†</sup>		
			DNSMOS <sup>†</sup>	NISQA <sup>†</sup>	UTMOS <sup>†</sup>	POLQA <sup>†</sup>	PESQ <sup>†</sup>	ESTOI <sup>†</sup>	SDR <sup>†</sup>	MCD <sup>†</sup>	LSD <sup>†</sup>	SBS <sup>†</sup>	LPS <sup>†</sup>			SpkSim <sup>†</sup>	CACC (%) <sup>†</sup>
1	T1	D	2.88 (8)	3.22 (6)	2.09 (5)	<b>3.40 (1)</b>	<b>2.64 (1)</b>	<b>0.82 (1)</b>	<b>12.66 (1)</b>	3.67 (2)	2.93 (3)	<b>0.87 (1)</b>	<b>0.74 (1)</b>	<b>0.76 (1)</b>	<b>79.80 (1)</b>	3.24 (5)	2.97 (1)
2	T2	D+G	2.92 (5)	3.24 (5)	2.16 (3)	3.17 (4)	2.47 (4)	0.79 (5)	11.10 (5)	3.96 (7)	2.99 (8)	0.84 (4)	0.70 (5)	0.74 (3)	76.06 (6)	3.32 (3)	4.37 (2)
3	T3	D+G	2.94 (4)	3.25 (4)	2.19 (2)	3.16 (5)	2.45 (6)	0.79 (4)	11.25 (4)	4.79 (11)	3.66 (12)	0.83 (6)	0.71 (3)	0.71 (6)	77.09 (5)	3.44 (2)	4.47 (3)
4	T4	D+G	2.80 (15)	3.01 (10)	2.04 (6)	3.22 (2)	2.47 (3)	0.80 (3)	11.47 (3)	3.90 (5)	2.94 (4)	0.85 (2)	0.71 (4)	0.74 (2)	78.06 (3)	3.28 (4)	4.63 (4)
5	T5	D+G	2.83 (13)	2.92 (12)	2.03 (7)	3.18 (3)	2.48 (2)	0.81 (2)	11.69 (2)	<b>3.64 (1)</b>	2.98 (7)	0.85 (3)	0.72 (2)	0.74 (4)	78.35 (2)	3.04 (10)	5.80 (5)
6	T6	D+G	2.91 (7)	3.28 (3)	1.98 (9)	2.99 (8)	2.29 (8)	0.78 (6)	10.58 (7)	4.22 (8)	3.78 (13)	0.83 (7)	0.70 (6)	0.70 (7)	77.15 (4)	3.21 (6)	6.53 (6)
7	T7	D+G	2.85 (11)	3.08 (8)	1.98 (8)	3.13 (6)	2.45 (5)	0.78 (7)	10.74 (6)	3.95 (6)	2.81 (2)	0.84 (5)	0.68 (7)	0.70 (9)	75.28 (9)	3.19 (7)	7.27 (7)
8	T8	D	2.92 (6)	3.16 (7)	1.97 (10)	2.99 (7)	2.25 (9)	0.76 (9)	10.39 (8)	3.89 (4)	2.95 (5)	0.82 (8)	0.66 (9)	0.69 (10)	75.30 (8)	3.17 (9)	8.23 (8)
9	T9	D+G	3.08 (2)	3.69 (2)	2.11 (4)	2.67 (14)	1.99 (14)	0.74 (12)	7.43 (16)	4.42 (9)	3.14 (9)	0.80 (11)	0.65 (10)	0.71 (5)	72.47 (15)	3.17 (8)	8.70 (9)
10	T10	D	2.85 (10)	2.77 (14)	1.92 (16)	2.99 (9)	2.24 (10)	0.76 (8)	10.24 (9)	3.80 (3)	<b>2.72 (1)</b>	0.82 (9)	0.67 (8)	0.70 (8)	75.60 (7)	2.96 (12)	9.63 (10)
13	T13	G	<b>3.10 (1)</b>	<b>3.74 (1)</b>	<b>2.53 (1)</b>	1.99 (21)	1.34 (21)	0.54 (22)	-12.28 (22)	10.31 (22)	7.14 (22)	0.78 (17)	0.59 (18)	0.47 (22)	67.87 (21)	<b>3.69 (1)</b>	12.53 (13)
22	T22	-	1.90 (22)	1.58 (22)	1.55 (22)	1.83 (22)	1.31 (22)	0.58 (21)	3.24 (20)	9.34 (21)	5.84 (20)	0.70 (22)	0.51 (22)	0.55 (19)	73.41 (12)	2.13 (22)	20.50 (22)
1	T2 <sup>†</sup>	D+G	2.92 (3)	3.24 (3)	2.16 (2)	3.17 (3)	2.47 (2)	0.79 (3)	11.10 (4)	3.96 (3)	2.99 (2)	0.84 (2)	0.70 (3)	0.74 (1)	76.06 (4)	3.32 (2)	2.50 (1)
2	T3 <sup>†</sup>	D+G	2.90 (5)	3.11 (4)	2.13 (3)	3.18 (1)	2.44 (3)	0.79 (2)	11.61 (2)	4.44 (5)	3.44 (5)	0.83 (3)	0.71 (2)	0.72 (3)	77.51 (2)	3.27 (3)	3.00 (2)
3	T5 <sup>†</sup>	D+G	2.83 (8)	2.90 (7)	2.02 (4)	3.18 (2)	2.49 (1)	0.81 (1)	11.76 (1)	3.63 (1)	3.00 (3)	0.85 (1)	0.72 (1)	0.74 (2)	78.40 (1)	2.99 (5)	3.07 (3)
4	T6 <sup>†</sup>	D+G	2.91 (4)	3.31 (2)	1.96 (5)	3.00 (4)	2.31 (4)	0.78 (4)	11.21 (3)	4.14 (4)	3.57 (6)	0.83 (4)	0.70 (4)	0.71 (4)	76.77 (3)	3.25 (4)	3.87 (4)

**Table 3: Language-wise DNSMOS and CACC scores of selected models.**

ID	Model type	German		English		French		Spanish		Chinese		Japanese (unseen)	
		DNSMOS <sup>†</sup>	CACC (%) <sup>†</sup>	DNSMOS <sup>†</sup>	CACC (%) <sup>†</sup>	DNSMOS <sup>†</sup>	CACC (%) <sup>†</sup>	DNSMOS <sup>†</sup>	CACC (%) <sup>†</sup>	DNSMOS <sup>†</sup>	CACC (%) <sup>†</sup>	DNSMOS <sup>†</sup>	CACC (%) <sup>†</sup>
T1	D	2.90	82.0	2.80	79.2	2.86	75.8	2.82	86.4	2.98	75.3	2.93	75.1
T2	D+G	2.93	79.4	2.80	73.6	2.92	73.7	2.89	83.4	3.01	64.3	2.96	71.2
T3	D+G	2.96	79.9	2.84	74.4	2.95	74.5	2.89	85.9	3.02	71.3	2.99	67.9
T10	D	2.88	79.0	2.75	73.8	2.81	70.2	2.85	82.3	2.92	70.7	2.89	73.0
T13	G	3.10	74.0	3.17	68.0	3.14	68.6	2.97	80.9	3.16	20.1	3.07	36.8
T22	-	2.02	76.8	1.71	70.3	1.76	67.3	1.97	80.8	1.97	69.3	1.94	74.5

T10 in all languages except Japanese, CACC of T3 for Japanese data is lower than that of T10, indicating that there is still room for improvement in robustness on unseen languages.

Since T13 exhibited a markedly different trend from the others, we conducted a more detailed analysis of its outputs. Audio examples are available on our demo page<sup>9</sup>. We found that the model occasionally hallucinated spoken content, particularly under low-SNR conditions, consistent with the findings in [45]. Notably, for Japanese inputs—which were unseen during training—the output under low-SNR conditions sometimes resembled English or other European languages that dominated the training data. A similar phenomenon was observed in in-painted frames affected by packet loss. These findings suggest that generative SE models may exhibit a non-negligible degree of language dependency.

### 4.3. Preliminary analysis on evaluation metrics

As expected, Table 3 demonstrates that DNSMOS yields high scores even when hallucinations occur. Although only DNSMOS is reported here due to space constraints, the other two non-intrusive metrics exhibited similar trends. These results imply that, particularly when evaluating generative models, it is essential to combine non-intrusive metrics with other metrics that can penalize hallucination, such as CACC.

In the subjective evaluation, some generative approaches were preferred over the top system based on the discriminative model. Note that we conducted subjective evaluations only on English data due to the time constraint of the Interspeech challenge. Generative models are expected not to hallucinate the content heavily on English data as the majority of the training data was English. Since P.808 is a reference-free absolute category rating test where the prompt was ‘‘How do you rate the overall quality of the following speech sample?’’, it would

be difficult to penalize the correctness of the spoken content and speaker consistency as long as the speech sounds natural. Thus, the listeners may have given higher scores to generative models, which typically output speech with less audible distortions than discriminative models. The results suggest that a new subjective listening protocol, which takes into account spoken content and/or speaker consistency, would be needed to evaluate generative models comprehensively.

## 5. Conclusion

We introduced the Interspeech 2025 URGENT Challenge to investigate language dependency, universality on various distortion types, the effectiveness of using noisy data, and the data scalability of SE models. The most popular approach was hybrid models but purely discriminative models gave the best score. The analysis of the submissions implied that (i) generative approaches appeared to be more language-dependent than discriminative ones, (ii) more but possibly noisy data does not necessarily lead to better performance, and (iii) generative models tended to be preferred to discriminative ones in the subjective evaluation (P.808 test). In future work, we will perform a subjective listening test on languages other than English to further analyze the language dependency of the SE systems and metrics.

## 6. Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number JP24KJ2096. The leaderboard evaluation has been supported by the PSC Bridges2 system via ACCESS allocation CIS210014, supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The subjective listening test was funded and executed by Technische Universität Braunschweig.

<sup>9</sup><https://kohei0209.github.io/urgent25-demo/>

## 7. References

- [1] L. Diener *et al.*, “INTER\_SPEECH 2022 audio deep packet loss concealment challenge,” in *Proc. Interspeech*, 2022, pp. 580–584.
- [2] J.-M. Lemerrier *et al.*, “Wind noise reduction with a diffusion-based stochastic regeneration model,” in *Speech Communication; 15th ITG Conference*. VDE, 2023, pp. 116–120.
- [3] J. Serrà *et al.*, “Universal speech enhancement with score-based diffusion,” *arXiv preprint arXiv:2206.03065*, 2022.
- [4] R. Scheibler *et al.*, “Universal score-based speech enhancement with high content preservation,” in *Proc. Interspeech*, 2024, pp. 1165–1169.
- [5] W. Zhang *et al.*, “Toward universal speech enhancement for diverse input conditions,” in *Proc. IEEE ASRU*, 2023.
- [6] W. Zhang *et al.*, “URGENT challenge: Universality, robustness, and generalizability for speech enhancement,” in *Proc. Interspeech*, 2024, pp. 4868–4872.
- [7] W. Zhang *et al.*, “Lessons learned from the URGENT 2024 speech enhancement challenge,” in *Accepted by Interspeech*, 2025.
- [8] C. K. Reddy *et al.*, “DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *Proc. IEEE ICASSP*, 2022, pp. 886–890.
- [9] C. K. Reddy *et al.*, “The INTER\_SPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results,” in *Proc. Interspeech*, 2020, pp. 2492–2496.
- [10] H. Dubey *et al.*, “ICASSP 2023 deep noise suppression challenge,” *IEEE Open Journal of Signal Processing*, pp. 1–13, 2024.
- [11] R. Cutler *et al.*, “ICASSP 2023 speech signal improvement challenge,” *IEEE Open Journal of Signal Processing*, pp. 1–12, 2024.
- [12] N.-C. Ristea *et al.*, “ICASSP 2024 speech signal improvement challenge,” *IEEE Open Journal of Signal Processing*, vol. 6, pp. 238–246, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10830509>
- [13] F.-L. Wang *et al.*, “Disentangling the impacts of language and channel variability on speech separation networks,” in *Proc. Interspeech*, 2022, pp. 5343–5347.
- [14] F. L. Chong *et al.*, “A methodology for improving PESQ accuracy for Chinese speech,” in *TENCON 2005-2005 IEEE Region 10 Conference*. IEEE, 2005, pp. 1–6.
- [15] D. Konane *et al.*, “Impact of languages and accent on perceived speech quality predicted by perceptual evaluation of speech quality (PESQ) and perceptual objective listening quality assessment (POLQA): Case of Moore, Dioula, French and English,” *Open Journal of Applied Sciences*, vol. 11, no. 12, pp. 1324–1332, 2021.
- [16] G. Mittag *et al.*, “NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets,” in *Proc. Interspeech*, 2021, pp. 2127–2131.
- [17] T. Saeki *et al.*, “UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022,” in *Proc. Interspeech*, 2022, pp. 4521–4525.
- [18] T. Fujimura *et al.*, “Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech,” in *Proc. EUSIPCO*. IEEE, 2021, pp. 436–440.
- [19] S. Wisdom *et al.*, “Unsupervised sound separation using mixture invariant training,” *Proc. NIPS*, vol. 33, pp. 3846–3857, 2020.
- [20] J. Kaplan *et al.*, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
- [21] X. Zhai *et al.*, “Scaling vision transformers,” in *Proc. CVPR*, 2022, pp. 12 104–12 113.
- [22] W. Zhang *et al.*, “Beyond performance plateaus: A comprehensive study on scalability in speech enhancement,” in *Proc. Interspeech*, 2024, pp. 1740–1744.
- [23] H. Zen *et al.*, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” in *Proc. Interspeech*, 2019, pp. 1526–1530.
- [24] C. Veaux *et al.*, “The Voice Bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *Proc. O-COCOSDA/CASLRE*, 2013, pp. 1–4.
- [25] LDC, *LDC Catalog: CSR-I (WSJ0) Complete*, University of Pennsylvania, 1993.
- [26] Philadelphia: Linguistic Data Consortium, *LDC Catalog: CSR-II (WSJ1) Complete LDC94SI3A*, 1994.
- [27] J. Richter *et al.*, “EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation,” in *Proc. Interspeech*, 2024, pp. 4873–4877.
- [28] V. Pratap *et al.*, “MLS: A large-scale multilingual dataset for speech research,” in *Proc. Interspeech*, 2020, pp. 2757–2761.
- [29] R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” in *Proc. LREC*, 2020, pp. 4218–4222.
- [30] G. Wichern *et al.*, “WHAM!: Extending speech separation to noisy environments,” in *Proc. Interspeech*, 2019, pp. 1368–1372.
- [31] E. Fonseca *et al.*, “FSD50K: an open dataset of human-labeled sound events,” *IEEE/ACM Trans. ASLP*, vol. 30, pp. 829–852, 2021.
- [32] M. Defferrard *et al.*, “FMA: A dataset for music analysis,” *arXiv preprint arXiv:1612.01840*, 2016.
- [33] J. Thiemann *et al.*, “The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings,” in *Proc. Mtgs. Acoust.*, vol. 19, no. 1. AIP Publishing, 2013.
- [34] A. Mesaros *et al.*, “A multi-device dataset for urban acoustic scene classification,” in *DCASE*, 2018, pp. 9–13.
- [35] D. Fejgin *et al.*, “Brudex database: Binaural room impulse responses with uniformly distributed external microphones,” in *Speech Communication; 15th ITG Conf.*, 2023, pp. 126–130.
- [36] T. Dietzen *et al.*, “MYRIAD: A multi-array room acoustic database,” *EURASIP J. Audio Speech Music Process.*, vol. 2023, article no. 17, pp. 1–14, Apr. 2023.
- [37] A. Défossez *et al.*, “High fidelity neural audio compression,” *Transactions on Machine Learning Research*, 2023.
- [38] R. Kumar *et al.*, “High-fidelity audio compression with improved rvqgan,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 980–27 993, 2023.
- [39] J. G. Beerends *et al.*, “Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—temporal alignment,” *Journal of the Audio Eng. Society*, vol. 61, no. 6, pp. 366–384, 2013.
- [40] A. W. Rix *et al.*, “Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs,” in *Proc. IEEE ICASSP*, vol. 2, 2001, pp. 749–752.
- [41] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Trans. ASLP*, vol. 24, no. 11, pp. 2009–2022, 2016.
- [42] E. Vincent *et al.*, “Performance measurement in blind audio source separation,” *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [43] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proc. of IEEE Pacific Rim Conf. on Communications Computers and Signal Process.*, 1993, pp. 125–128.
- [44] A. Gray and J. Markel, “Distance measures for speech processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
- [45] J. Pirklbauer *et al.*, “Evaluation metrics for generative speech enhancement methods: Issues and perspectives,” in *Speech Communication; 15th ITG Conference*, 2023, pp. 265–269.
- [46] T. Saeki *et al.*, “SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics,” in *Proc. Interspeech*, 2024, pp. 4943–4947.
- [47] M. Zanon Boito *et al.*, “mHuBERT-147: A compact multilingual HuBERT model,” in *Proc. Interspeech*, 2024, pp. 3939–3943.
- [48] J.-w. Jung *et al.*, “Pushing the limits of raw waveform speaker recognition,” in *Proc. Interspeech*, 2022, pp. 2228–2232.
- [49] Y. Peng *et al.*, “OWSM v3.1: Better and faster open Whisper-style speech models based on e-branchformer,” in *Proc. Interspeech*, 2024, pp. 352–356.
- [50] “ITU-T recommendation P.808, subjective evaluation of speech quality with a crowdsourcing approach,” ITU-T, Jun. 2018.
- [51] B. Naderi and R. Cutler, “An open source implementation of ITU-T recommendation P.808 with validation,” in *Proc. Interspeech*, 2020, pp. 2862–2866.
- [52] M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.
- [53] Z.-Q. Wang *et al.*, “TF-GridNet: Integrating full-and sub-band modeling for speech separation,” *IEEE/ACM Trans. ASLP*, vol. 31, pp. 3221–3236, 2023.
- [54] C. Li *et al.*, “ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration,” in *Proc. IEEE SLT*, 2021, pp. 785–792.