NeurIPS 2024 Competition Proposal: URGENT Challenge

Wangyou Zhang Robin Scheibler Kohei Saijo Samuele Cornell
Chenda Li Zhaoheng Ni Anurag Kumar Marvin Sach Wei Wang
Shinji Watanabe Tim Fingscheidt Yanmin Qian

wyz-97@sjtu.edu.cn

Abstract

Speech enhancement (SE) is the task of improving the quality of the desired speech while suppressing other interference signals. Tremendous progress has been achieved in the past decade in deep learning-based SE approaches. However, existing SE studies are often limited in one or multiple aspects of the following: coverage of SE sub-tasks, diversity and amount of data (especially real-world evaluation data), and diversity of evaluation metrics. As the first step to fill this gap, we establish a novel SE challenge, called **URGENT**, to promote research towards universal SE. It concentrates on the universality, robustness, and generalizability of SE approaches. In the challenge, we extend the conventionally narrow SE definition to cover different sub-tasks, thus allowing the exploration of the limits of current SE models. We start with four SE sub-tasks, including denoising, dereverberation, bandwidth extension, and declipping. Note that handling the above sub-tasks within a single SE model has been challenging and underexplored in the SE literature due to the distinct data formats in different tasks. As a result, most existing SE approaches are only designed for a specific subtask. To address this issue, we propose a technically novel framework to unify all these sub-tasks in a single model, which is compatible to most existing SE approaches. Several state-of-the-art baselines with different popular architectures have been provided for this challenge, including TF-GridNet, BSRNN, and Conv-TasNet. We also take care of the data diversity and amount by collecting abundant public speech and noise data from different domains. This allows for the construction of diverse training and evaluation data. Additional real recordings are further used for evaluating robustness and generalizability. Different from existing SE challenges, we adopt a wide range of evaluation metrics to provide comprehensive insights into the true capability of both generative and discriminative SE approaches. We expect this challenge would not only provide valuable insights into the current status of SE research, but also attract more research towards building universal SE models with strong robustness and good generalizability.

Keywords

Speech enhancement, universality, robustness, generalizability

1 Competition description

1.1 Background and impact

We propose the URGENT challenge to promote research on the **U**niversality, **R**obustness, and **G**eneralizability of speech **E**nhanceme**NT** (SE) models. This challenge is most related to the area of speech signal processing and speech enhancement [1], which defines the task of improving a speech signal that has been subject to distortions such as additive noise, acoustic interference, reverberation,

Preprint. Under review.

or bandwidth limitation. Despite their impressive performance on popular benchmarks [2, 3, 4, 5], most existing SE research and challenges have limitations on the coverage of SE sub-tasks, diversity and amount of data, and diversity of evaluation metrics. In particular, universally robust and generalizable SE approaches have been underexplored. We believe this is an important problem to address in order to cope with diverse speech applications in the real world. Advancement in this direction will enable speech processing in more sophisticated real-world scenarios and facilitate numerous downstream speech applications [6], such as automatic speech recognition (ASR), speech translation, and speaker recognition. Therefore, in the proposed challenge, we aim to fill the gap in this direction and try to answer the following research question:

Can we build a single universal SE model to successfully handle various distinct SE sub-tasks in diverse input conditions?

Here, "diverse input conditions" refer to different speech data formats, such as different sampling frequencies (SF). The former relates to the diverse recording devices in real applications, which require the SE model to support different SFs. The latter corresponds to the variations in speech duration, which can lead to significantly different computational costs depending on the SE model architecture. We measure the "successfulness" in the above question from three aspects:

- **universality**: the average performance over all conditions is good.
- **robustness**: the worst performance over all conditions is good.
- **generalizability**: the performance does not degrade significantly for unseen/mismatched devices or environments.

We believe that the proposed challenge would be of interest not only to the speech community, but also to the broader audience in the audio or signal processing community, as they all share highly-related methodologies and can thus draw inspiration and insights from our findings. We conducted a preliminary survey among the speech enhancement community and have received replies from **12 research groups** confirming their willingness to participate in the proposed challenge. Meanwhile, we are also reaching out to 25 potential participants who are active in speech enhancement and have significant records on the challenge activities.

The proposed challenge is highly related to speech applications in our daily life. Since diverse speech devices are used in a wide variety of scenarios, it is likely that the received speech signal in the device (e.g., cell phone) will contain one or multiple distortions caused by background noise, reverberation, clipping (due to limited device capability or heterogenous pre-processing at the device level), and so on. Moreover, different devices may operate at distinct sampling rates, resulting in speech signals of various SFs. The outcome of our proposed challenge is expected to be a single universal SE model that can handle all conditions mentioned above. This is generally favored against conventional task-specific or SF-specific SE models, because it is much easier and more convenient to deploy and maintain a single model than a bunch of specialized models.

1.2 Novelty

The proposed URGENT challenge is a brand new challenge in the SE area, featuring the following three major innovations:

1) Better coverage of SE sub-tasks: As detailed in Section 1.4, we start with four SE sub-tasks in this challenge to explore the universality of SE approaches. This is motivated by the fact that speech signals in real-world applications are likely to contain different SFs and several distortions caused by the environment and recording devices. This necessitates the construction of a universal SE model capable of handling various distortions and input formats. Compared to training separate SE models for each distortion and each input format, constructing a single universal SE model is more resource-efficient and more straightforward to deploy. Importantly, it also alleviates the error propagation problem that usually occurs in cascaded specialized models. We are also interested in investigating whether sharing the knowledge among different sub-tasks in a single SE model can improve the overall performance during the challenge. A technically novel framework is proposed to facilitate this exploration, which is carefully designed to be compatible with existing SE approaches.
2) Larger scale and more diverse data with training data mandated and limited: It is commonly observed that the evaluation of most existing SE approaches is often conducted on fixed, small (e.g.,

Table 1: Detailed information of the corpora used in the challenge. Shaded cells are only used for the test data, while corpora in other cells are shared for training, development, and non-blind test sets.

Type	Corpus	Condition	SF (kHz)	Duration (After pre-processing)	License
Speech	LibriVox data from DNS5 challenge [12]	Audiobook	8~48	~350 h	CC BY 4.0
	LibriTTS reading speech [26]	Audiobook	8~24	~200 h	CC BY 4.0
	CommonVoice 11.0 English portion [27]	Crowd-sourced voices	8~48	~550 h	CC0
	VCTK reading speech [28]	Newspaper, etc.	48	~80 h	ODC-By
	WSJ reading speech [29, 30]	WSJ news	16	~85 h	LDC User Agreement
	Clean speech corpora only for the blind test set	Unseen domains	8~48	$>$ 100 h	CC/Apache/MIT licenses
	Noisy real speech corpora only for the blind test set	Unseen domains	8~48	$>$ 100 h	CC/Apache/MIT licenses
Noise	Audioset+FreeSound noise in DNS5 challenge	Crowd-sourced + Youtube	8~48	~180 h	CC BY 4.0
	WHAM! noise [4]	4 Urban environments	48	~70 h	CC BY-NC 4.0
	Noise corpora only for the blind test set	$>$ 20 unseen types	8~48	$>$ 150 h	CC/Apache/MIT licenses
RIR	Simulated RIRs from DNS5 challenge	SLR28	48	~60k samples	CC BY 4.0
	Other RIRs simulated by participants	-	8~48	-	-
	RIRs only for test data	Real-recorded RIRs	16~48	$>$ 10k samples	CC/Apache/MIT licenses

~10 h) or medium datasets (e.g., ~100 h). This can potentially lead to heavily over-fitted model designs for these datasets. Moreover, such evaluations are usually restricted to matched conditions where the speech quality, linguistic content, noise family, and other characteristics can be highly similar to the training condition. This also impedes the understanding of the generalizability of different SE approaches to unseen "speech-in-the-wild" conditions [7]. In this challenge, we aim to explicitly evaluate this aspect with public data from different domains as detailed in Section 1.3. Meanwhile, we mandate and restrict the training material allowed for system development. This enables a comparable and conclusive analysis of the capabilities of different SE systems.

3) Extensive evaluation metrics: One major drawback of existing challenges is that they often only use one or two objective metrics to evaluate challenge submissions, which may not well reflect the comprehensive performance of SE models. For instance, it is possible to train SE models based on a specific evaluation metric (e.g., scale-invariant signal-to-noise ratio [8]) discriminatively, which thus leads to high scores in that metric but also biases the evaluation. In our challenge, we aim to break this convention and propose to adopt four distinct categories of metrics to evaluate SE models from different perspectives. These include a variety of non-intrusive SE metrics, intrusive SE metrics, downstream task-independent (e.g., phoneme similarity[1]) metrics, and downstream task-dependent metrics. The extensive evaluation metrics distinguish our challenge from existing ones, and fit perfectly to our multi-task design. Moreover, our ranking rule based on these metrics also takes care of both generative and discriminative SE approaches, thus encouraging efforts in both directions to tackle this challenge.

While there have been several SE challenges in the literature, our proposed challenge is distinct from them as mentioned above. In particular, existing SE challenges generally focus on specific scenarios, such as denoising and dereverberation [5, 9, 10, 11, 12], speech restoration [13, 14], packet loss concealment [15], acoustic echo cancellation [16, 17, 18, 19], hearing aids [20, 21], 3D SE [22, 23, 24], far-field multi-channel SE for video conferencing [25], and unsupervised domain adaptation for denoising [7][2]. These challenges have fostered the development of state-of-the-art (SOTA) SE models in the past. Complementarily, the proposed URGENT challenge provides unique insights into the universality, generalizability, and robustness of SE approaches with a wide range of scenario evaluation metrics.

1.3 Data

We have also taken care of the data diversity and amount during the challenge design. For simulating the training, development, and non-blind test data, we collect abundant public speech and noise data from different domains, including LibriVox & CommonVoice & VCTK & WSJ speech corpora and Audioset & FreeSound & WHAM! noise corpora. The detailed information of the corpora is listed in Table 1.

Almost all corpora in Table 1 adopt an open license that allows free use of the data (at least for non-commercial use). The only exception is the WSJ corpus, which is owned by LDC and thus

[1]Unlike downstream tasks such as ASR, the phoneme information is generally language independent. Therefore, we denote it as a downstream task-independent metric.

[2]Our challenge shares some similarities with the CHiME-7 UDASE task [7] which explores unsupervised adaptation of SE models to real scenarios. However, our challenge considers more distortions than just noise and reverberation, and adopts more diverse training/evaluation data as well as much more metrics.

requires a purchased license. Regarding this specific corpus, we have contacted the LDC on their data policy, and they have granted our challenge participants a free temporary license of the WSJ corpus during the challenge period. Therefore, we can ensure that all corpora listed in Table 1 can be freely accessible during the challenge. Moreover, WSJ is a commonly used corpus in ASR and other speech tasks, which is very likely to be available in different research groups in this area. So it is still possible for the community to conduct post-challenge evaluation.

The collected raw speech data are further pre-processed through the following steps to detect the true SF and remove low-quality samples:

1) We first estimate the effective bandwidth of each speech and noise sample based on the energy thresholding algorithm proposed in [31]. This is critical for our proposed method to successfully handle data with different SFs. Then, we resample each speech and noise sample accordingly to the best matching SF, which is defined as the lowest SF among {8, 16, 22.05, 24, 32, 44.1, 48} kHz that can fully cover the estimated effective bandwidth.
2) A voice activity detection (VAD) algorithm[3] is further used to detect "bad" speech samples that are actually non-speech or mostly silence, which will be removed from the data.
3) Finally, the non-intrusive DNSMOS scores (OVRL, SIG, BAK) [32] are calculated for each remaining speech sample. This allows us to detect noisy and low-quality speech samples via thresholding each score.

We finally curated a list of speech (~1300 hours) and noise (~250 hours) samples based on the above procedure, and they will be used for simulating the training and test data in the challenge. Note that these do not include the corpora listed in the shaded cells in Table 1, which will be further added to the blind test set for a comprehensive evaluation in diverse conditions.

During the challenge, the participants are allowed to freely simulate degraded training and development data based on the listed corpora. We will provide data preparation scripts as well as baseline implementations to ease this process. However, unlike existing DNS challenges, they are not allowed to use external corpora other than the listed ones. This enables a comparable and conclusive experimental analysis.

The evaluation will be conducted based on the analysis of results on both simulated and real-recorded test data. Our current non-blind test set consists of more than 9000 samples covering 4 different distortions and SFs ranging from 8 kHz to 48 kHz. This is much larger and more diverse than existing SE challenges, which often use only hundreds of test samples for evaluation. In addition, we will also include real-recorded ASR corpora that adopt an open license as the blind test set, which allows for evaluation on more realistic scenarios. This sufficient amount and diversity of test data will allow us to draw conclusive and statistically significant results, from which we can provide solid insights to the community.

1.4 Tasks and application scenarios

Our broader definition of the SE task is illustrated in Figure 1, and the SE model $\mathrm{SE}(\cdot)$ can be formulated in the general form below:

$$\hat{\mathbf{x}} = \mathrm{SE}(\mathcal{F}(\mathbf{x}))\,, \tag{1}$$

where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are respectively the desired clean speech and enhanced speech. The former is degraded by a distortion model $\mathcal{F}(\cdot)$, and the degraded speech $\mathcal{F}(\mathbf{x})$ is fed as input to SE models. While this definition seems similar to the commonly-adopted SE definition in the literature, ours has two key distinctions. First, conventional SE approaches often only support a single SF, while our model allows the model input $\mathcal{F}(\mathbf{x})$ to have various sampling frequencies (SF). This enables us to model the often sub-optimal SFs (less than 44.1 kHz) that are usually encountered in real-world devices. Second, we cover diverse distortions (i.e., additive noise, reverberation, clipping, and bandwidth limitation) using the distortion model $\mathcal{F}$, while conventional SE studies largely focus on noise or reverberation.

As mentioned earlier in Section 1.2, these definitions can better capture the complicated forms of degraded speech in realistic scenarios. For example, we are surrounded by a wide variety of speech devices in the real world, which can have very different SFs ranging from 8 kHz to as high as 48 kHz. They capture speech signals along with background noise, reverberation, and many other

[3] `https://github.com/wiseman/py-webrtcvad`

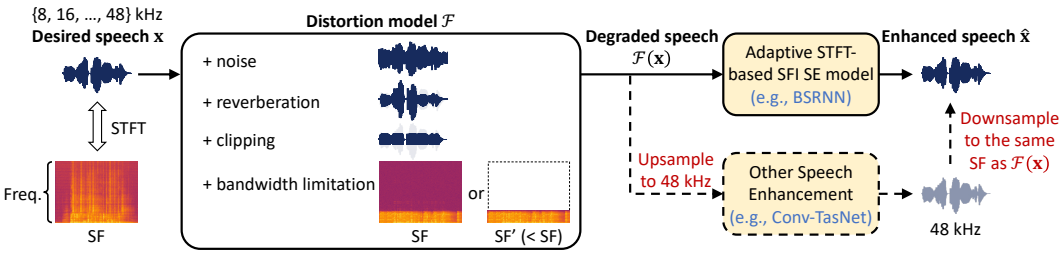

Figure 1: URGENT speech enhancement task definition.

forms of distortions that can be introduced by the devices themselves (e.g., clipping due to low device capability and bandwidth limitation due to low microphone quality). The proposed task will thus cover these variations caused by both the environment and devices, leading to better-suited SE techniques for real-world applications.

Despite the clear definition of this task, it has been a challenging topic in the SE literature, with only very limited explorations [33, 34] that do not cover all of the aforementioned variations. To address this issue, we propose a novel SE framework that has been proven successful through our preliminary experiments in unifying all aforementioned sub-tasks as well as different SF variations.

As illustrated in Figure 1, in this challenge, we unify different sub-tasks via a distortion model $\mathcal{F}(\cdot)$ that provides a consistent data format for different SE sub-tasks. This allows us to unify multiple sub-tasks in a single model by simply combining data with different distortions. In particular, we take care of the bandwidth limitation distortion[4], as it may result in changes in the data format (e.g., data shape, and SF) unlike other distortions. This distortion is defined as removing high-frequency components from the speech data via low-pass filtering. This process often corresponds to signal downsampling, as shown in Figure 1. However, it can also happen with high-SF signals where the upper frequencies are missing due to poor microphone devices. To unify the data format for all different distortions, we propose to always ensure the SFs of the desired, degraded, and enhanced speech are the same as illustrated in Figure 1. In this way, we can easily build a multi-task SE system with a unified data format.

Furthermore, we provide participants, via our baseline code, with two alternative model designs as shown in the right part of Figure 1.

1) For most conventional SE models that are only designed for a single SF, we adopt a simple yet effective strategy via pre-processing and post-processing [35]. That is, we always upsample the model input to the highest SF (i.e., 48 kHz), so that the model only needs to process 48 kHz data during both training and inference. Then, we downsample the model output back to the original input SF for model training and evaluation.
2) For specific time-frequency domain architectures that are invariant to the input data shape (both time and frequency dimensions), we adopt another strategy named sampling-frequency-independent (SFI) processing. Inspired by existing works on SFI SE approaches [36, 37, 38, 39], we adopt the SFI short-time Fourier transform (STFT) [37, 39] to replace the default STFT layer in the SE models, where the STFT window size and hop size are adaptively adjusted according to the input SF to have a fixed duration. For these models, no further pre-processing or post-processing is needed.

With the above carefully designed framework, we can now easily build an SE system to handle different sub-tasks and SFs. Our preliminary exploration (submitted to the Interspeech 2024 conference [40][5]) also verified the effectiveness of this framework with several popular SE architecture, including Conv-TasNet [41], BSRNN [38], and TF-GridNet [42]. The first baseline adopts model design 1), while the other two baselines adopt model design 2) for handling different SFs.

1.5 Metrics

To obtain a comprehensive evaluation of the above baseline models, we adopt the following four categories of evaluation metrics that capture different aspects of the enhanced speech quality:

[4]It corresponds to the bandwidth extension (BWE) sub-task.

[5]Note that the paper in submission [40] only serves as a preliminary investigation, while the actual challenge has not been launched. And we are proposing to launch it as a NeurIPS competition.

Table 2: Preliminary evaluation on non-blind test data. "SBS." denotes the SpeechBERTScore. Results with * are not fully comparable due to different data and training setups. The per-metric ranking is denoted by red numbers in parentheses. The per-category average ranking and the overall ranking are denoted by 🏅 and 🏆, respectively.

Model	Non-intrusive SE metrics		Intrusive SE metrics						Downstream-indep.		Downstream-task-dep.	
	DNSMOS ↑	**NISQA** ↑	**POLQA** ↑	**PESQ** ↑	**ESTOI** ↑	**SDR (dB)** ↑	**MCD** ↓	**LSD** ↓	**SBS.** ↑	**PhnSim** ↑	**SpkSim** ↑	**WAcc** (%) ↑
Noisy input	1.64 (6)	1.76 (6)	2.50 (4)	1.63 (5)	0.704 (4)	6.11 (5)	6.76 (5)	3.99 (5)	**0.87** (1)	0.68 (5)	0.72 (3)	82.18 (3)
OM-LSA [54]	2.19 (5)	2.09 (5)	2.37 (5)	1.81 (4)	0.702 (5)	10.88 (4)	5.26 (4)	3.64 (4)	0.85 (4)	0.71 (4)	0.65 (5)	78.61 (4)
VoiceFixer [33]*	**2.93** (1)	**3.65** (1)	1.97 (6)	1.50 (6)	0.527 (6)	-9.59 (6)	9.16 (6)	7.54 (6)	0.81 (6)	0.59 (6)	0.54 (6)	66.19 (6)
Conv-TasNet	2.31 (4)	2.71 (4)	3.12 (3)	2.42 (3)	0.799 (3)	14.42 (3)	3.23 (3)	2.73 (3)	0.85 (4)	0.73 (3)	0.70 (4)	76.82 (5)
BSRNN	2.41 (3)	3.05 (3)	3.49 (2)	2.66 (2)	0.833 (2)	14.89 (2)	2.75 (2)	2.66 (2)	**0.87** (1)	0.80 (2)	0.77 (2)	82.53 (2)
TF-GridNet	2.43 (2)	3.06 (2)	**3.54** (1)	**2.76** (1)	**0.841** (1)	**15.42** (1)	**2.70** (1)	**2.39** (1)	**0.87** (1)	**0.81** (1)	**0.78** (1)	**82.87** (1)

Model	🏆	🏅Non-intrusive SE metrics	🏅Intrusive SE metrics	🏅Downstream-task-independent	🏅Downstream-task-dependent
Noisy input	4.175	6.0	4.7	3.0	3.0
OM-LSA [54]	4.450	5.0	4.3	4.0	4.5
VoiceFixer [33]*	4.750	**1.0**	6.0	6.0	6.0
Conv-TasNet	3.750	4.0	3.0	3.5	4.5
BSRNN	2.125	3.0	2.0	1.5	2.0
TF-GridNet	**1.250**	2.0	**1.0**	**1.0**	**1.0**

- intrusive SE metrics: POLQA [43], PESQ [44], extended short-time objective intelligibility (ESTOI) [45], signal-to-distortion ratio (SDR) [46], mel cepstral distortion (MCD) [47], log-spectral distance (LSD) [48];
- non-intrusive SE metrics: DNSMOS [32], NISQA [49];
- downstream-task-independent metrics: phoneme similarity (PhnSim, equal to "1-LPD" in [50]), SpeechBERTScore [51];
- downstream-task-dependent metrics: speaker similarity (SpkSim), word accuracy (WAcc)[6].

The intrusive SE metrics reflect the objective quality of the enhanced speech from the signal perspective, which require well-aligned reference speech as an additional input. The non-intrusive SE metrics emphasize the speech naturalness and overall quality, which are predicted by pre-trained neural networks, thus not requiring any reference speech. The downstream-task-independent metrics measure how the high-level task-agnostic representations (e.g., phoneme predictions and discrete tokens) of the enhanced speech match those of the reference speech. For example, the PhnSim metric compares the sequence-level phoneme similarity between the enhanced and reference speech, which has proven effective for evaluating generative SE approaches in the correctness of their generated contents [50]. The SpeechBERTScore metric compares the similarity between semantic embeddings of the enhanced and reference speech. It is worth noting that while they need the reference speech for metric calculation, no strict alignment between the enhanced and reference speech signals is required. The downstream-task-dependent metrics measure either a task-specific characteristic of the enhanced speech (e.g., speaker similarity) or the compatibility with a downstream task (e.g. ASR performance in terms of WAcc). These metrics allow us to easily exploit real-recorded data for extensive evaluation. We use the RawNet3 [52] model pre-trained on VoxCeleb datasets for cosine-based speaker similarity calculation and the OWSM v3.1 [53] model for WAcc calculation. Among all the metrics above, a lower value in MCD and LSD indicates a better speech quality, while for other metrics, a higher value represents better SE performance.

To show the effectiveness of these proposed metrics, we provide the evaluation results from our preliminary experiments based on the simulated data as introduced in Section 1.3. As shown in Table 2, we trained three different models, Conv-TasNet, BSRNN, and TF-GridNet and compared their performance with other baseline approaches. Among them, OM-LSA [54] a typical denoising method based on classical signal processing, which serves as a weak baseline that can only cope with the denoising sub-task. VoiceFixer is a generative SE approach with vocoder-based resynthesis[7], which was trained to handle the same set of distortions as introduced in Section 1.4. Since VoiceFixer was trained on a different dataset in [33], we cannot make a fair comparison with it. However, we can still take it as a reference to verify the efficacy of other baseline models.

The results in Table 2 show that different metrics tend to capture different aspects of the enhanced speech. For example, non-intrusive SE metrics (DNSMOS and NISQA) favors the generative

[6]WAcc is equal to 1− word error rate (WER).

[7]Available at `https://github.com/haoheliu/voicefixer`. We adopted "mode 0" as it performs best.

SE approach (VoiceFixer) as it can generate more natural speech compared to discriminative SE approaches (Conv-TasNet, BSRNN, and TF-GridNet). Intrusive SE metrics measure the signal-level quality which are more strict about the sample-wise alignment between enhanced and reference speech. Thus, discriminative approaches outperform VoiceFixer due to their capability of preserving the sample-wise alignment. Downstream-task-independent metrics (SpeechBERTScore and PhnSim) measure the overall and fine-grained quality in terms of contents, while downstream-task-dependent metrics (SpkSim and WAcc) measure the task-specific performance. In particular, we find PhnSim and SpkSim to be good consistency indicators of the enhanced speech. It is well known that generative approaches tend to yield more natural speech but with potentially modified contents or styles. Such an inconsistency can be well captured through these metrics. For example, VoiceFixer achieves low scores in both metrics, indicating that its output contain different phoneme-level contents and speaker characteristics from the original speech, while other discriminative approaches can achieve better consistency. In addition, the SOTA discriminative model (TF-GridNet) achieves the best performance among all models, which is also reflected by the final ranking. This further verifies the effectiveness of the adopted metrics and the proposed ranking strategy in assessing the efficacy of different approaches.

Facilitated by the above preliminary investigation, we propose the following ranking strategy inspired by the Friedman test [55, 56]:

1. First, we calculate the ranking for each metric independently.
2. Then, we average the ranking for each of the four categories (i.e., intrusive SE, non-intrusive SE, downstream-task-independent, and downstream-task-dependent metrics). This results in four category-dependent rankings.
3. Finally, we average the category-dependent rankings to obtain the final ranking.

1.6 Baselines, code, and material provided

As mentioned earlier in Section 1.4, we will provide three baselines (i.e., Conv-TasNet, BSRNN, and TF-GridNet) along with the corresponding open-source implementation in the ESPnet toolkit [57] and training recipe. We are also considering adding an additional baseline based on the diffusion-based generative SE approach [58].

The recipe includes scripts for data downloading, preprocessing, simulation, model training, and metric evaluation[8]. They will be made available to participants and the community after the challenge begins. Note that the provided scripts are only for reference, and the participants can freely use their own codebase for system development.

1.7 Website, tutorial and documentation

The challenge website will be released at `https://urgent-challenge.github.io/urgent2024/`, which will provide self-contained information about the URGENT challenge, including basic description, data introduction, detailed baseline documentation, timeline, leaderboard, FAQ, and news. We will also provide a dedicated email address `urgent.challenge@gmail.com` for communication with participants.

2 Organizational aspects

2.1 Protocol

To participate in the challenge, participants can freely submit their results to our leaderboard anytime during the challenge period. Note that we only require the submission of enhanced audios without the corresponding code, thus attracting more potential submissions. The evaluation will be conducted automatically on our server based on the evaluation metrics and ranking strategies proposed in Section 1.5.

The current challenge will consist of two major phases. During the first phase, we only release training and validation data to the participants for system development. During the second phase, we release the test data for participants to submit their final results. To prevent cheating and overfitting, we will anonymize the meta information of the test samples, and include real-recorded data as an important part of evaluation, which is unlikely to be seen by the participants.

[8]Will be available at `https://github.com/urgent-challenge/urgent2024_challenge`.

2.2 Rules and Engagement

We only have one track for this challenge, and do not apply any constraint on computational cost or latency. So the participants can freely design their SE approach, either generative or discriminative, to maximize the overall performance. The major requirement is that only the mandated training materials mentioned in Section 1.3 are allowed to be used for system development. This is to ensure a fair comparison among different submissions. No registration is required to participate in the challenge, however, participants will need to contact us if they need a temporary LDC license to access the WSJ data as mentioned in Section 1.3.

The participants can communicate with the organizers via email and GitHub issues/discussions. We are also considering creating an official channel on Slack. To make sure our challenge updates can be delivered to participants from different sources, we will sync up the information via multiple channels, including the "news" section in our official website, and email updates to invited participants.

2.3 Schedule and readiness

The tentative timeline for challenge preparation is:

- **[done]** Finishing training and development data preparation scripts.
- **[done]** Finishing baseline implementation.
- **[done]** Verifying the reproducibility of the scripts.
- **[done]** Cleaning up the script for releasing.
- **[April 30, 2024]** Finishing website preparation and setting it online.
- **[May 20, 2024]** Finishing leaderboard implementation.
- **[May 27, 2024]** Fixing the bind test data.
- **[May 27, 2024]** Announcement of the challenge.

The tentative timeline for running the challenge is:

- **[June 10, 2024]** Challenge begins. Release of all scripts, evaluation plan, and training+development data.
- **[August 19, 2024]** Release of non-blind test data.
- **[September 18, 2024]** Release of blind test data.
- **[September 20, 2024]** Challenge ends.
- **[October 21, 2024]** Notification of final results.

2.4 Competition promotion and incentives

To promote participation in the challenge, we will curate a mailing list based on the publication and challenge records of active research groups in the speech area. We will distribute the call among organization members when sending invitations to potential participants in the list.

We encourage participants to write description and analysis papers with named authors. Since speech research groups are relatively under-represented at NeurIPS, we hope this challenge could be a good opportunity to promote the speech research.

3 Resources

3.1 Resources provided by organizers

We have computational resources on the PSC Bridges2 system via ACCESS allocation CIS210014, supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. These could be used for conducting organizer-side exploratory experiments, evaluation, and analysis. But the participants will have to use their own computational resources for system development. We are also supported by the LDC, which generously provides a temporary license of the WSJ corpus during the challenge period.

3.2 Support requested

We will need supportive materials for visa application to participate in NeurIPS in person. In addition, a poster session is needed for us and challenge winners to present the challenge results.

Organizing team

1. **Wangyou Zhang** (coordinator, data provider, baseline method provider)
   - **Affiliation:** Shanghai Jiao Tong University, China
   - **Email:** wyz-97@sjtu.edu.cn
   - PhD student working on speech enhancement/separation and robust speech recognition in the cocktail party problem. Co-creator and core developer of the ESPnet-SE toolkit.
2. **Robin Scheibler** (beta tester, evaluator)
   - **Affiliation:** LY Corp., Japan
   - **Email:** robin.scheibler@linecorp.com
   - Senior researcher at LY Corp., working on speech and audio processing. Founder and core developer of the Pyroomacoustics toolkit.
3. **Kohei Saijo** (baseline method provider, beta tester)
   - **Affiliation:** Waseda University, Japan
   - **Email:** saijo@pcl.cs.waseda.ac.jp
   - PhD student working on sound source separation.
4. **Samuele Cornell** (beta tester, evaluator)
   - **Affiliation:** Carnegie Mellon University, USA
   - **Email:** cornellsamuele@gmail.com
   - Post-doc researcher working on audio processing and machine learning. Co-creator of Asteroid source separation toolkit and author of SpeechBrain toolkit.
5. **Chenda Li** (baseline method provider, beta tester)
   - **Affiliation:** Shanghai Jiao Tong University, China
   - **Email:** lichenda1996@sjtu.edu.cn
   - PhD student working on speech separation, multi-talker processing. Co-creator and core developer of the ESPnet-SE toolkit.
6. **Zhaoheng Ni** (evaluator)
   - **Affiliation:** Meta, USA
   - **Email:** zni@meta.com
   - Research scientist at Meta Reality Labs, working on generative modeling for audio, single-channel/multi-channel speech enhancement, and so on.
7. **Anurag Kumar** (evaluator)
   - **Affiliation:** Meta, USA
   - **Email:** anuragkr@ieee.org
   - Research lead and scientist at Meta Research, primarily working on deep learning, audio/speech processing and multimodal Learning.
8. **Marvin Sach** (beta tester, evaluator)
   - **Affiliation:** Technische Universität Braunschweig, Germany
   - **Email:** marvin.sach@tu-braunschweig.de
   - PhD student working on speech enhancement.
9. **Wei Wang** (beta tester, evaluator)
   - **Affiliation:** Shanghai Jiao Tong University, China
   - **Email:** wangwei.sjtu@sjtu.edu.cn
   - PhD student working on robust speech processing and self-supervised learning.
10. **Shinji Watanabe** (coordinator, platform administrator)
    - **Affiliation:** Carnegie Mellon University, USA
    - **Email:** shinjiw@ieee.org

- IEEE Fellow, Associate Professor at CMU, primarily working on automatic speech recognition, speech enhancement, spoken language understanding, and machine learning for speech and language processing. Founder of the ESPnet toolkit. Core organizer of CHiME challenge series.

11. **Tim Fingscheidt** (coordinator, platform administrator)
    - **Affiliation:** Technische Universität Braunschweig, Germany
    - **Email:** `t.fingscheidt@tu-bs.de`
    - ITG Fellow, Professor at Technische Universität Braunschweig, primarily working on speech enhancement, speech coding and quality metrics, and machine learning for speech and computer vision.
12. **Yanmin Qian** (coordinator, platform administrator)
    - **Affiliation:** Shanghai Jiao Tong University, China
    - **Email:** `yanminqian@sjtu.edu.cn`
    - Full Professor at Shanghai Jiao Tong University, primarily working on speech recognition and translation, speaker and language recognition, speech separation and enhancement, natural language understanding and multi-media signal processing. Founding member of the Kaldi toolkit.

References

[1] Philipos C Loizou. Speech enhancement: theory and practice. CRC press, 2013.

[2] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. Computer Speech & Language, 46:535–557, 2017.

[3] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In Proc. Interspeech, pages 352–356, 2016.

[4] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending speech separation to noisy environments. In Proc. Interspeech, pages 1368–1372, 2019.

[5] Chandan K.A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In Proc. Interspeech, pages 2492–2496, 2020.

[6] Yen-Ju Lu, Xuankai Chang, Chenda Li, Wangyou Zhang, Samuele Cornell, Zhaoheng Ni, Yoshiki Masuyama, Brian Yan, Robin Scheibler, Zhong-Qiu Wang, Yu Tsao, Yanmin Qian, and Shinji Watanabe. ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding. In Proc. Interspeech, pages 5458–5462, 2022.

[7] Simon Leglaive, Léonie Borne, Efthymios Tzinis, Mostafa Sadeghi, Matthieu Fraticelli, Scott Wisdom, Manuel Pariente, Daniel Pressnitzer, and John R Hershey. The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement. In Proc. CHiME, 2023.

[8] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR–half-baked or well done? In Proc. IEEE ICASSP, pages 626–630, 2019.

[9] Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. ICASSP 2021 deep noise suppression challenge. In Proc. IEEE ICASSP, pages 6623–6627, 2021.

[10] Chandan K.A. Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. INTERSPEECH 2021 deep noise suppression challenge. In Proc. Interspeech, pages 2796–2800, 2021.

[11] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner. ICASSP 2022 deep noise suppression challenge. In Proc. IEEE ICASSP, pages 9271–9275, 2022.

[12] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Hannes Gamper, Mehrsa Golestaneh, and Robert Aichner. ICASSP 2023 deep noise suppression challenge. arXiv preprint arXiv:2303.11510, 2023.

[13] Ross Cutler, Ando Saabas, Babak Naderi, Nicolae-Cătălin Ristea, Sebastian Braun, and Solomiya Branets. ICASSP 2023 speech signal improvement challenge. arXiv preprint arXiv:2303.06566, 2023.

[14] Nicolae Catalin Ristea, Ando Saabas, Ross Cutler, Babak Naderi, Sebastian Braun, and Solomiya Branets. ICASSP 2024 speech signal improvement challenge. arXiv preprint arXiv:2401.14444, 2024.

[15] Lorenz Diener, Sten Sootla, Solomiya Branets, Ando Saabas, Robert Aichner, and Ross Cutler. INTERSPEECH 2022 audio deep packet loss concealment challenge. In Proc. Interspeech, pages 580–584, 2022.

[16] Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Sten Sootla, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sorensen, Robert Aichner, and Sriram Srinivasan. INTERSPEECH 2021 acoustic echo cancellation challenge. In Proc. Interspeech, pages 4748–4752, 2021.

[17] Kusha Sridhar, Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Hannes Gamper, Sebastian Braun, Robert Aichner, and Sriram Srinivasan. ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results. In Proc. IEEE ICASSP, pages 151–155, 2021.

[18] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sørensen, and Robert Aichner. ICASSP 2022 acoustic echo cancellation challenge. In Proc. IEEE ICASSP, pages 9107–9111, 2022.

[19] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Evgenii Indenbom, Nicolae-Catalin Ristea, Jegor Gužvin, Hannes Gamper, Sebastian Braun, and Robert Aichner. ICASSP 2023 acoustic echo cancellation challenge. arXiv preprint arXiv:2309.12553, 2023.

[20] Simone Graetzer, Jon Barker, Trevor J. Cox, Michael Akeroyd, John F. Culling, Graham Naylor, Eszter Porter, and Rhoddy Viveros Muñoz. Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing. In Proc. Interspeech, pages 686–690, 2021.

[21] Michael A Akeroyd, Will Bailey, Jon Barker, Trevor J Cox, John F Culling, Simone Graetzer, Graham Naylor, Zuzanna Podwińska, and Zehai Tu. The 2nd clarity enhancement challenge for hearing aid speech intelligibility enhancement: Overview and outcomes. In Proc. IEEE ICASSP, 2023.

[22] Eric Guizzo, Riccardo F. Gramaccioni, Saeid Jamili, Christian Marinoni, Edoardo Massaro, Claudia Medaglia, Giuseppe Nachira, Leonardo Nucciarelli, Ludovica Paglialunga, Marco Pennese, Sveva Pepe, Enrico Rocchi, Aurelio Uncini, and Danilo Comminiello. L3DAS21 challenge: Machine learning for 3D audio signal processing. In Proc. MLSP, pages 1–6. IEEE, 2021.

[23] Eric Guizzo, Christian Marinoni, Marco Pennese, Xinlei Ren, Xiguang Zheng, Chen Zhang, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello. L3DAS22 challenge: Learning 3D audio sources in a real office environment. In Proc. IEEE ICASSP, pages 9186–9190, 2022.

[24] Christian Marinoni, Riccardo Fosco Gramaccioni, Changan Chen, Aurelio Uncini, and Danilo Comminiello. Overview of the L3DAS23 challenge on audio-visual extended reality. arXiv preprint arXiv:2402.09245, 2024.

[25] Wei Rao, Yihui Fu, Yanxin Hu, Xin Xu, Yvkai Jv, Jiangyu Han, Zhongjie Jiang, Lei Xie, Yannan Wang, Shinji Watanabe, Zheng-Hua Tan, Hui Bu, Tao Yu, and Shidong Shang. ConferencingSpeech challenge: Towards far-field multi-channel speech enhancement for video conferencing. In Proc. IEEE ASRU, pages 679–686, 2021.

[26] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In Proc. Interspeech, pages 1526–1530, 2019.

[27] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4218–4222, 2020.

[28] Christophe Veaux, Junichi Yamagishi, and Simon King. The Voice Bank corpus: Design, collection and data analysis of a large regional accent speech database. In Proc. O-COCOSDA/CASLRE, pages 1–4, 2013.

[29] LDC. LDC Catalog: CSR-I (WSJ0) Complete. University of Pennsylvania, 1993.

[30] Philadelphia: Linguistic Data Consortium. LDC Catalog: CSR-II (WSJ1) Complete LDC94S13A, 1994.

[31] Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. Hi-Fi multi-speaker English TTS dataset. In Proc. Interspeech, pages 2776–2780, 2021.

[32] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In Proc. IEEE ICASSP, pages 886–890, 2022.

[33] Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. VoiceFixer: A unified framework for high-fidelity speech restoration. In Proc. Interspeech, pages 4232–4236, 2022.

[34] Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. arXiv preprint arXiv:2206.03065, 2022.

[35] Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, and Arvindh Krishnaswamy. A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech. In Proc. Interspeech, pages 2482–2486, 2020.

[36] Koichi Saito, Tomohiko Nakamura, Kohei Yatabe, Yuma Koizumi, and Hiroshi Saruwatari. Sampling-frequency-independent audio source separation using convolution layer based on impulse invariant method. In Proc. EUSIPCO, pages 321–325, 2021.

[37] Jouni Paulus and Matteo Torcoli. Sampling frequency independent dialogue separation. In Proc. EUSIPCO, pages 160–164, 2022.

[38] Jianwei Yu and Yi Luo. Efficient monaural speech enhancement with universal sample rate band-split RNN. In Proc. IEEE ICASSP, 2023.

[39] Wangyou Zhang, Kohei Saijo, Zhong-Qiu Wang, Shinji Watanabe, and Yanmin Qian. Toward universal speech enhancement for diverse input conditions. In Proc. IEEE ASRU, 2023.

[40] URGENT challenge organizers. URGENT challenge: Universality, robustness, and generalizability for speech enhancement. In In submission, 2024.

[41] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. ASLP., 27(8):1256–1266, 2019.

[42] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. TF-GridNet: Integrating full-and sub-band modeling for speech separation. IEEE/ACM Trans. ASLP., 31:3221–3236, 2023.

[43] John G Beerends, Christian Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. 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 Engineering Society, 61(6):366–384, 2013.

[44] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In Proc. IEEE ICASSP, volume 2, pages 749–752, 2001.

[45] Jesper Jensen and Cees H Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans. ASLP., 24(11):2009–2022, 2016.

[46] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. IEEE Trans. ASLP., 14(4):1462–1469, 2006.

[47] Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, pages 125–128, 1993.

[48] Augustine Gray and John Markel. Distance measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(5):380–391, 1976.

[49] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In Proc. Interspeech, pages 2127–2131, 2021.

[50] Jan Pirklbauer, Marvin Sach, Kristoff Fluyt, Wouter Tirry, Wafaa Wardah, Sebastian Moeller, and Tim Fingscheidt. Evaluation metrics for generative speech enhancement methods: Issues and perspectives. In Speech Communication; 15th ITG Conference, pages 265–269, 2023.

[51] Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics. arXiv preprint arXiv:2401.16812, 2024.

[52] Jee-weon Jung, Youjin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition. In Proc. Interspeech, pages 2228–2232, 2022.

[53] Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, Jee-weon Jung, and Shinji Watanabe. OWSM v3. 1: Better and faster open Whisper-style speech models based on E-Branchformer. arXiv preprint arXiv:2401.16658, 2024.

[54] Israel Cohen and Baruch Berdugo. Speech enhancement for non-stationary noise environments. Signal Processing, 81(11):2403–2418, 2001.

[55] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32(200):675–701, 1937.

[56] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, 11(1):86–92, 1940.

[57] Chenda Li, Jing Shi, Wangyou Zhang, Aswin Shanmugam Subramanian, Xuankai Chang, Naoyuki Kamo, Moto Hira, Tomoki Hayashi, Christoph Boeddeker, Zhuo Chen, and Shinji Watanabe. ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. In Proc. IEEE SLT, pages 785–792, 2021.

[58] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Trans. ASLP., 31:2351–2364, 2023.