# Lessons Learned from the URGENT 2024 Speech Enhancement Challenge

**Wangyou Zhang**[1], Kohei Saijo[2], Samuele Cornell[3], Robin Scheibler[4], Chenda Li[1], Zhaoheng Ni[5], Anurag Kumar[5], Marvin Sach[6], Wei Wang[1], Yihui Fu[6], Shinji Watanabe[3], Tim Fingscheidt[6], Yanmin Qian[1]

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

# CONTENTS

# Observation

1. Most existing speech enhancement (SE) research focuses on a single or limited range of conditions. (Narrow task defintion)

   noisy | anechoic | reverberant | certain sample rate | certain distortion

2. SE models are usually trained on small-sized data or single-domain data. (Lack of data diversity)

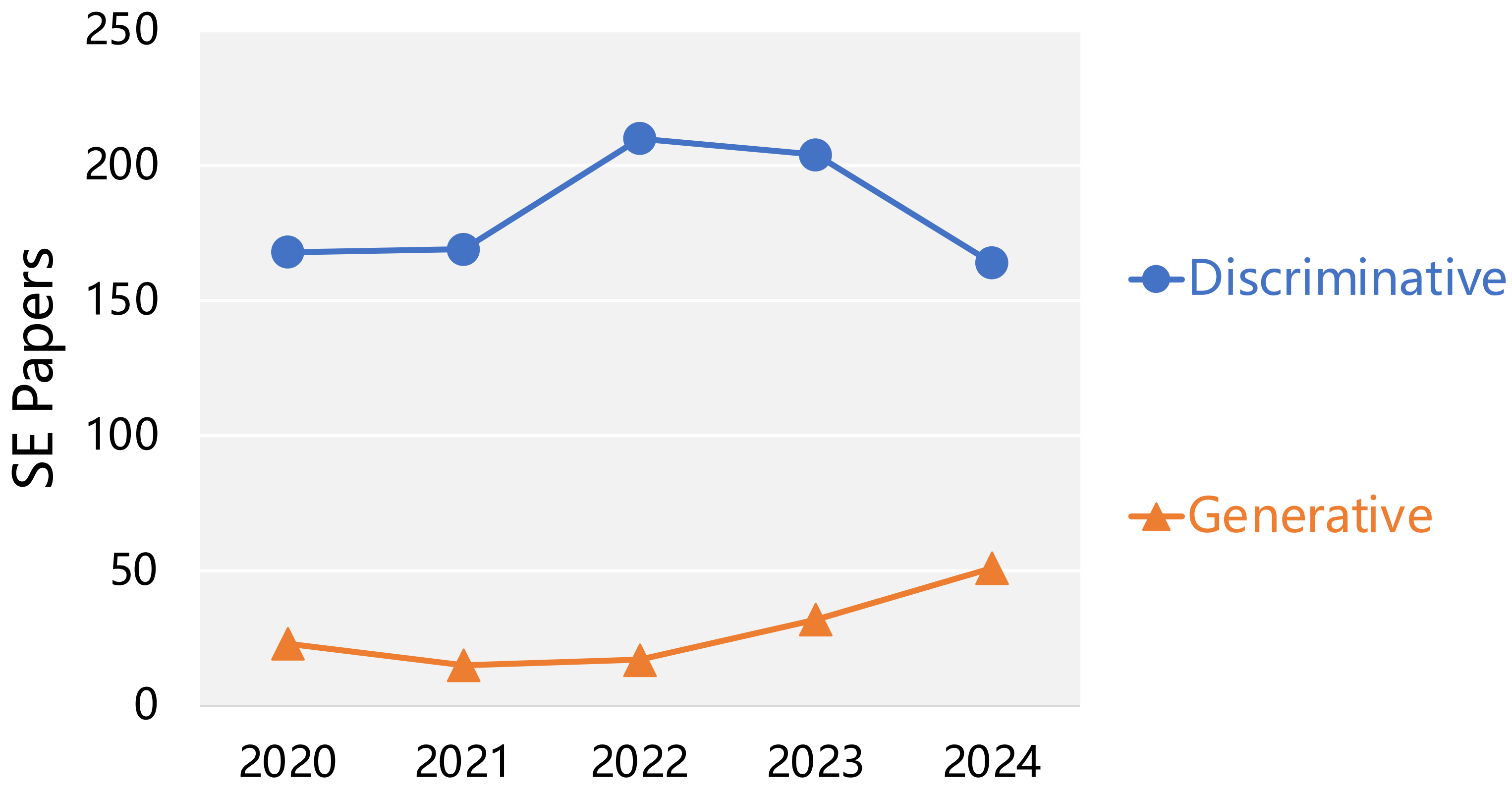
   VCTK+DEMAND | DNS Challenge | CHiME-4 | REVERB | WHAMR!

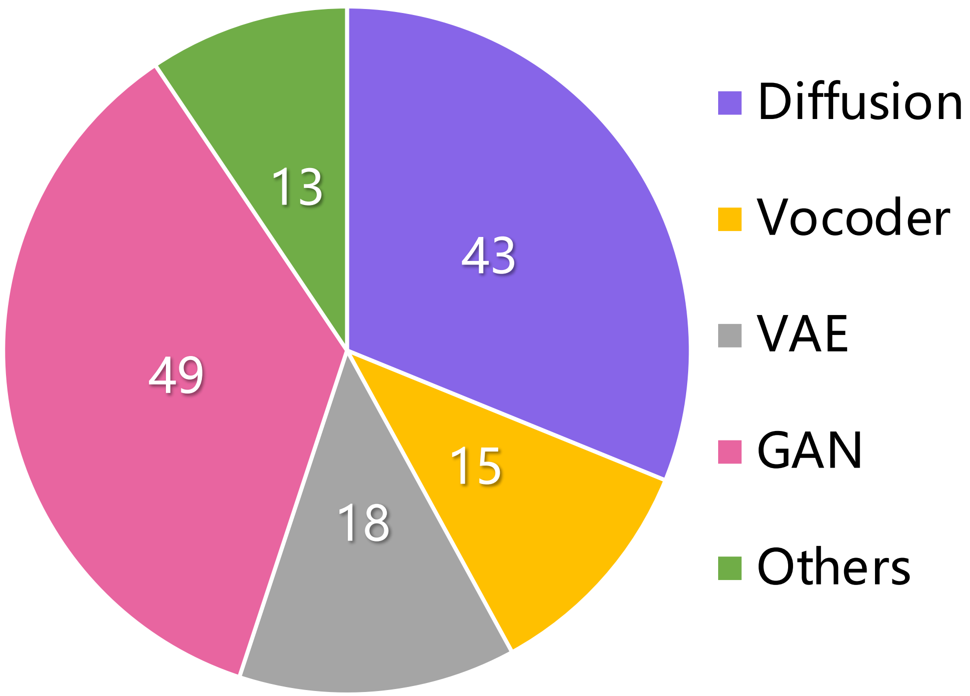
3. The evaluation of SE models is often done only on matched conditions, with just a few metrics. (Limited evaluation)

4. Performance has largely saturated on existing benchmarks, which only reflect limited scenarios in real world. (Outdated benchmarks)

# Observation (Cont'd)

5. Recent advances in generative methods for speech enhancement



# Goal

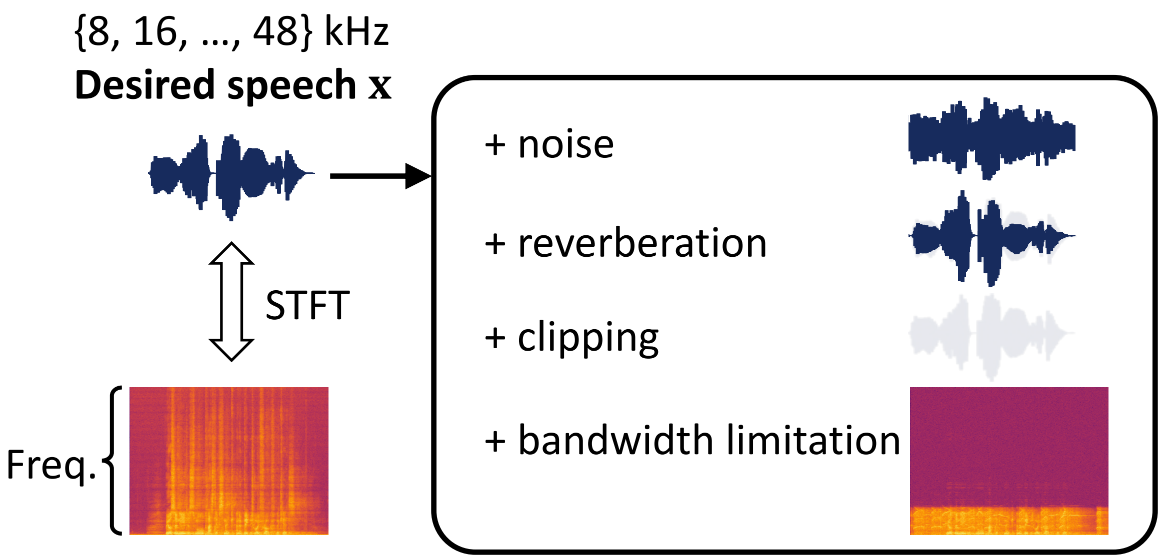
Conventional Speech Enhancement ⇒ Universally Robust Speech Enhancement w/ Generalizability

| Conventional Speech Enhancement | | Universally Robust Speech Enhancement w/ Generalizability |
| --- | --- | --- |
| • Only designed for a limited number of subtasks<br>• Only support one sampling frequency | **Universality** → | • Explicitly designed for various subtasks<br>• Support different input formats |
| • Only evaluated in limited data/conditions<br>• Limited evaluation metrics | **Robustness & Generalizability** → | • Evaluated in a wide range of conditions<br>• Diverse evaluation metrics |
| • Dominated by discriminative methods<br>• Mostly trained on single-domain / limited data | **Diversity** → | • Generative methods are encouraged<br>• Large-scale multi-domain data |

# URGENT Challenge – Task definition

- 4 sub-tasks
- A comprehensive range of sampling frequencies

Universally Robust
Speech Enhancement
w/ Generalizability




**Universality**
- Explicitly designed for various subtasks
- Support different input formats

**Robustness & Generalizability**
- Evaluated in a wide range of conditions
- Diverse evaluation metrics

**Diversity**
- Generative methods are encouraged
- Large-scale multi-domain data

# URGENT Challenge – Evaluation metrics

- 5 categories of multifaceted metrics
- A ranking-based overall evaluation protocol

Universally Robust Speech Enhancement w/ Generalizability

**Non-intrusive**: DNSMOS, NISQA

**Intrusive**: POLQA, PESQ, ESTOI, SDR, MCD, LSD

**Downstream-task-independent**: SpeechBERTScore, LPS

**Downstream-task-dependent**: SpkSim, WAcc

**Subjective**: MOS

**Universality**
- Explicitly designed for various subtasks
- Support different input formats

**Robustness & Generalizability**
- Evaluated in a wide range of conditions
- Diverse evaluation metrics

**Diversity**
- Generative methods are encouraged
- Large-scale multi-domain data

# URGENT Challenge – Data

| Type | Corpus | Condition |
|---|---|---|
| ~1300 hours<br>Speech | LibriVox data from DNS5 challenge | Audiobook |
| | LibriTTS reading speech | Audiobook |
| | CommonVoice 11.0 English portion | Crowd-sourced voices |
| | VCTK reading speech | Newspaper, etc. |
| | WSJ reading speech | WSJ news |
| ~250 hours<br>Noise | Audioset+FreeSound noise in DNS5 challenge | Crowd-sourced + Youtube |
| | WHAM! noise | 4 Urban environments |
| ~60k RIRs<br>RIR | Simulated RIRs from DNS5 challenge | SLR28 |

Universally Robust Speech Enhancement w/ Generalizability

**Universality**
- Explicitly designed for various subtasks
- Support different input formats

**Robustness & Generalizability**
- Evaluated in a wide range of conditions
- Diverse evaluation metrics

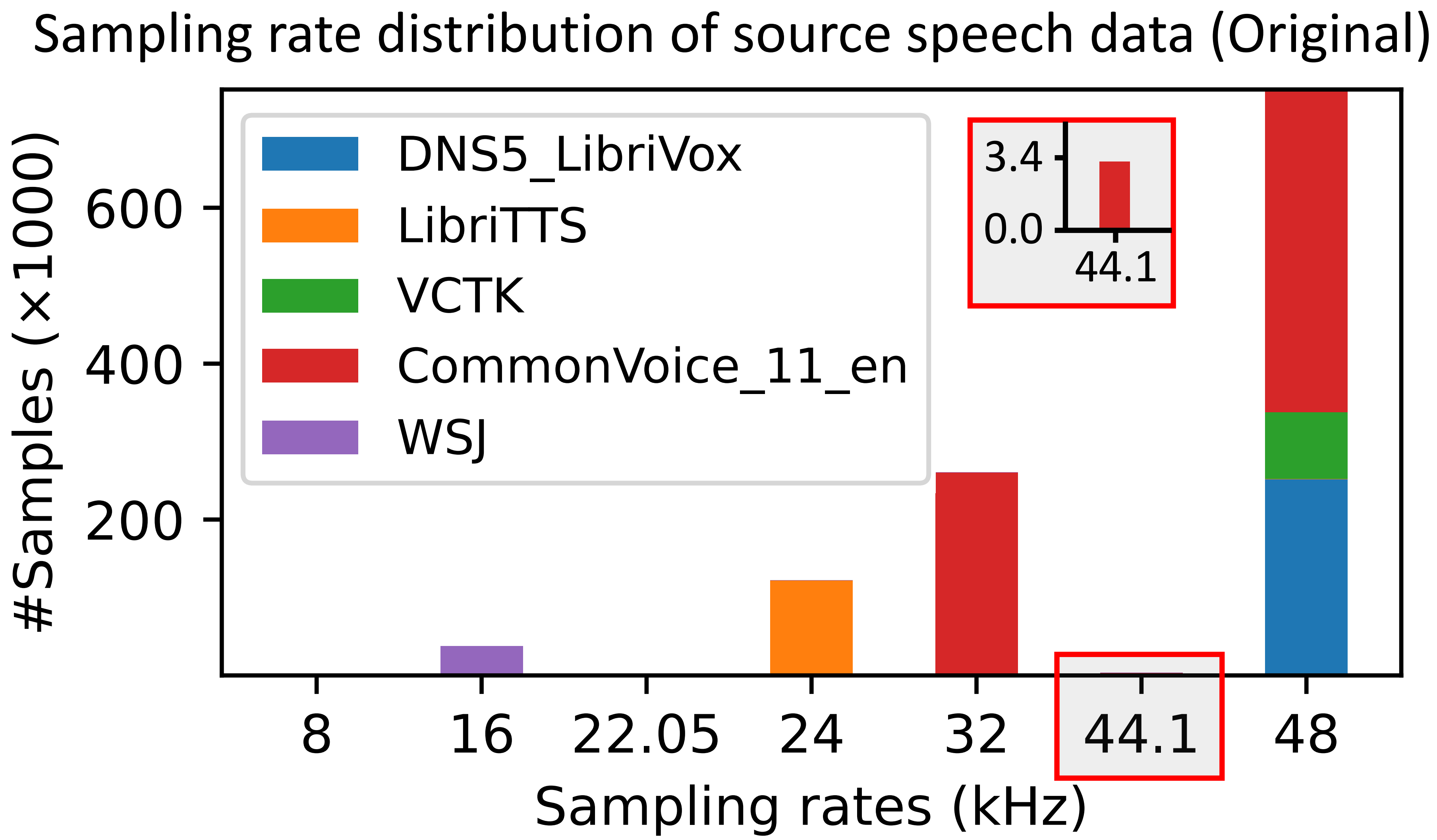
**Diversity**
- Generative methods are encouraged
- Large-scale multi-domain data

# CONTENTS

# Analysis: Data

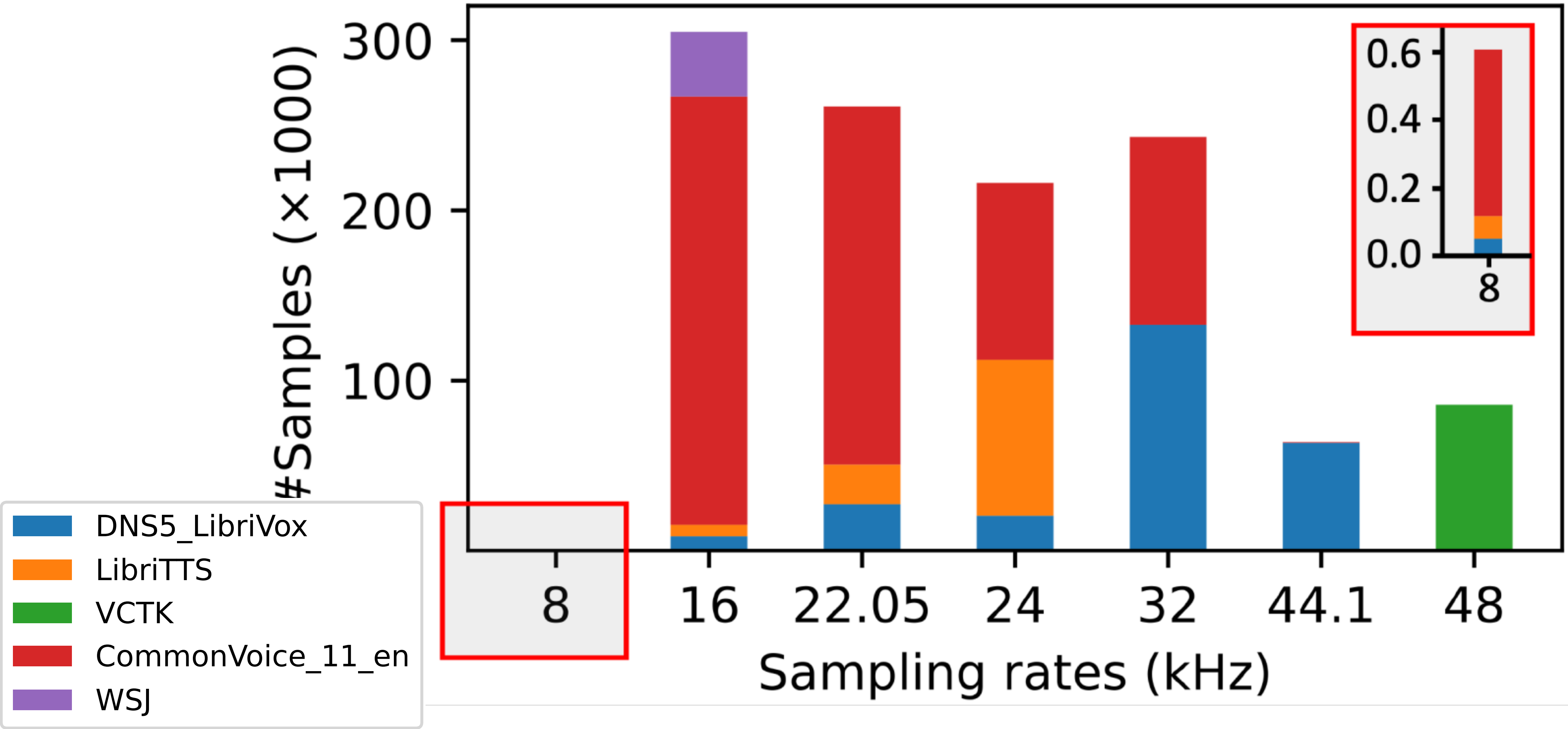
1. Sampling rate
   - The supposed 48 kHz speech can actually only contain much fewer frequency components.
   - It is important to re-estimate the effective bandwidth of collected audio data, even for some widely-used corpora.
2. Label noisiness
   - The noise floor commonly exists in non-studio-quality speech datasets, which may be supposed to be “clean”.
   - The SE model can be then misguided to preserve the noise floor (usually at a low level) in the enhanced speech.

# Analysis: Data (I) – sampling rate



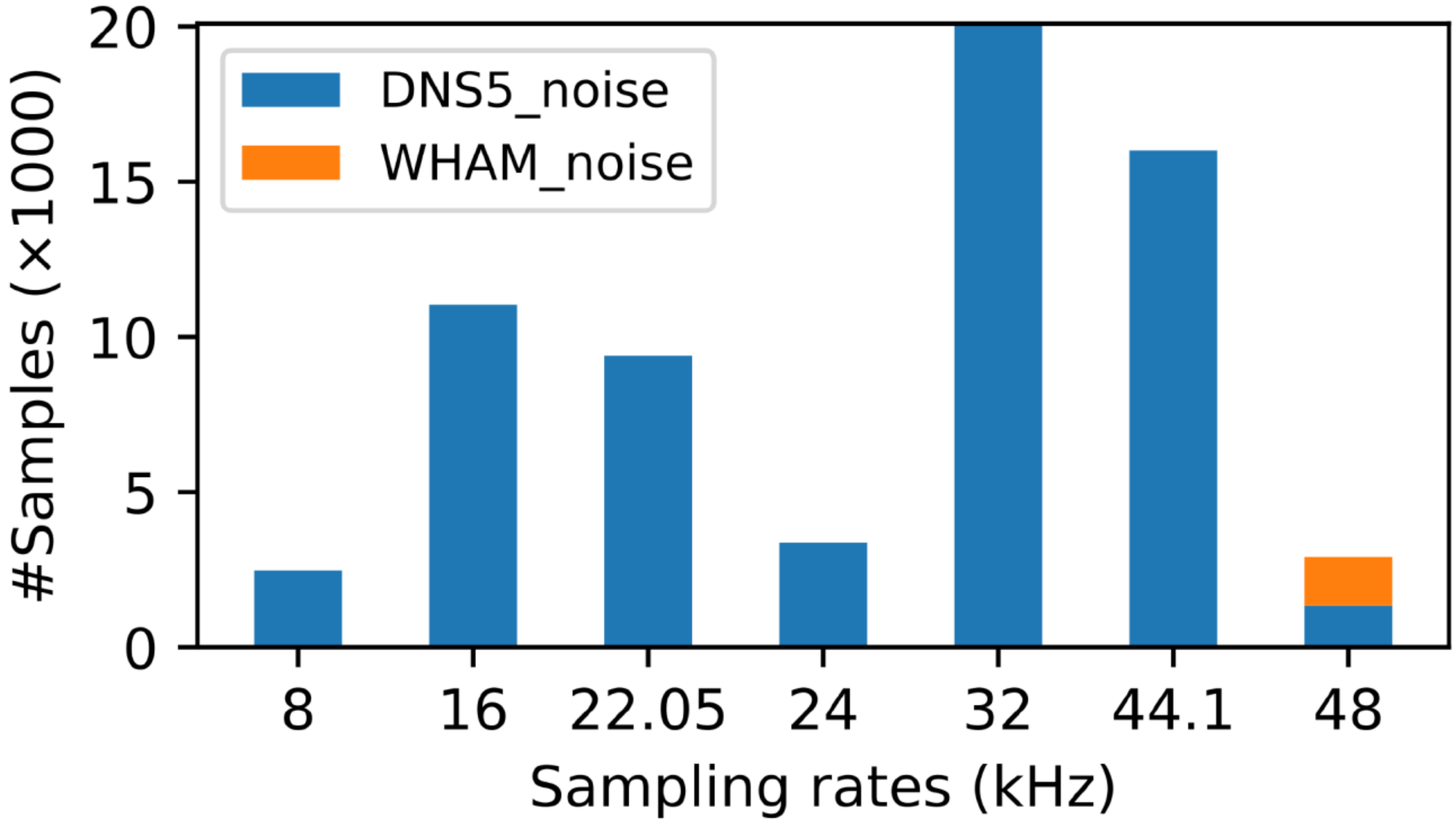
# Analysis: Data (I) – sampling rate

Sampling rate distribution of source speech data (Re-estimated)



# Analysis: Data (I) – sampling rate

Sampling rate distribution of source noise data (Re-estimated)



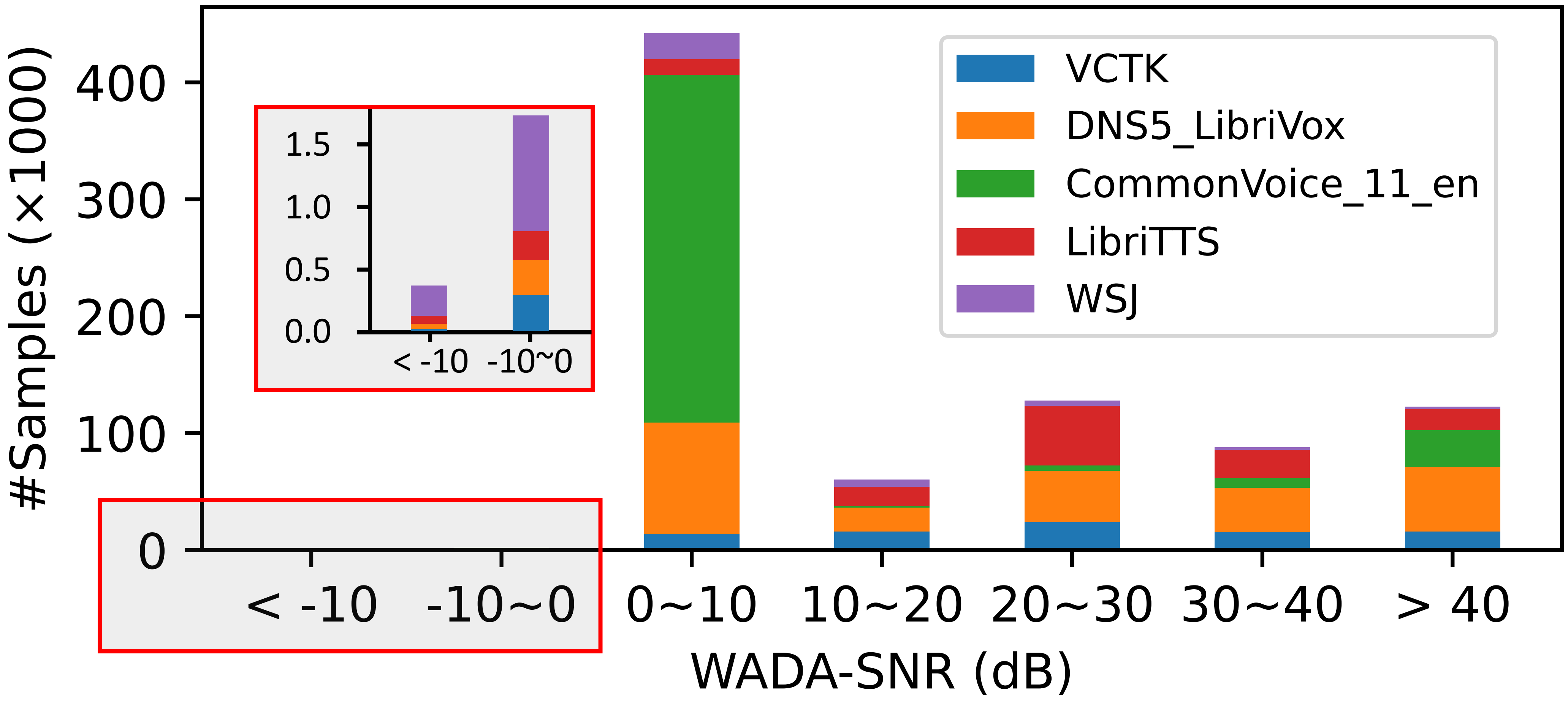
# Analysis: Data

1. Sampling rate
   - The seemingly 48 kHz speech can actually only contain much fewer frequency components.
   - It is important to re-estimate the effective bandwidth of collected audio data, even for some widely-used corpora.

2. Label noisiness
   - The noise floor commonly exists in non-studio-quality speech datasets, which may be supposed to be "clean".
   - The SE model can be then misguided to preserve the noise floor (usually at a low level) in the enhanced speech.

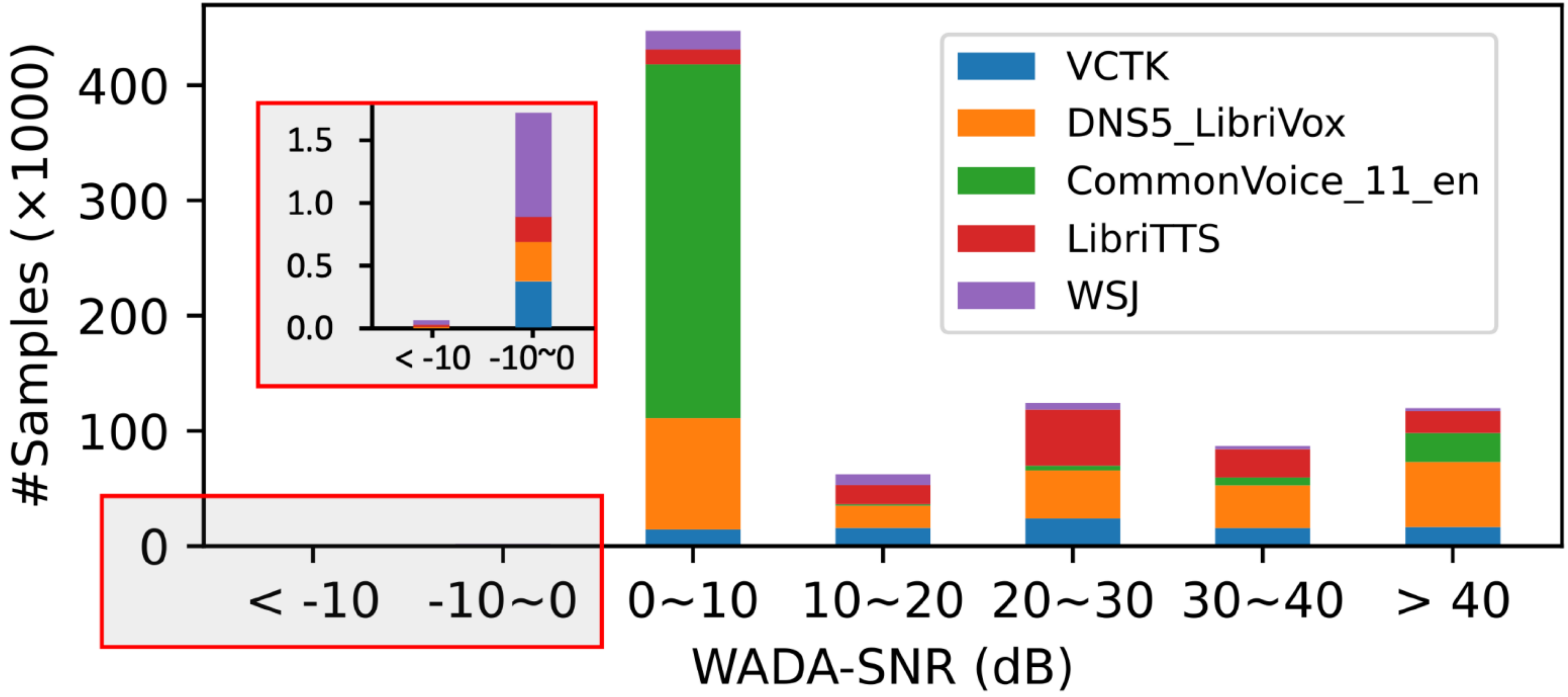
# Analysis: Data (II) – label noisiness

Estimated SNRs of the **original speech labels** in training&validation sets



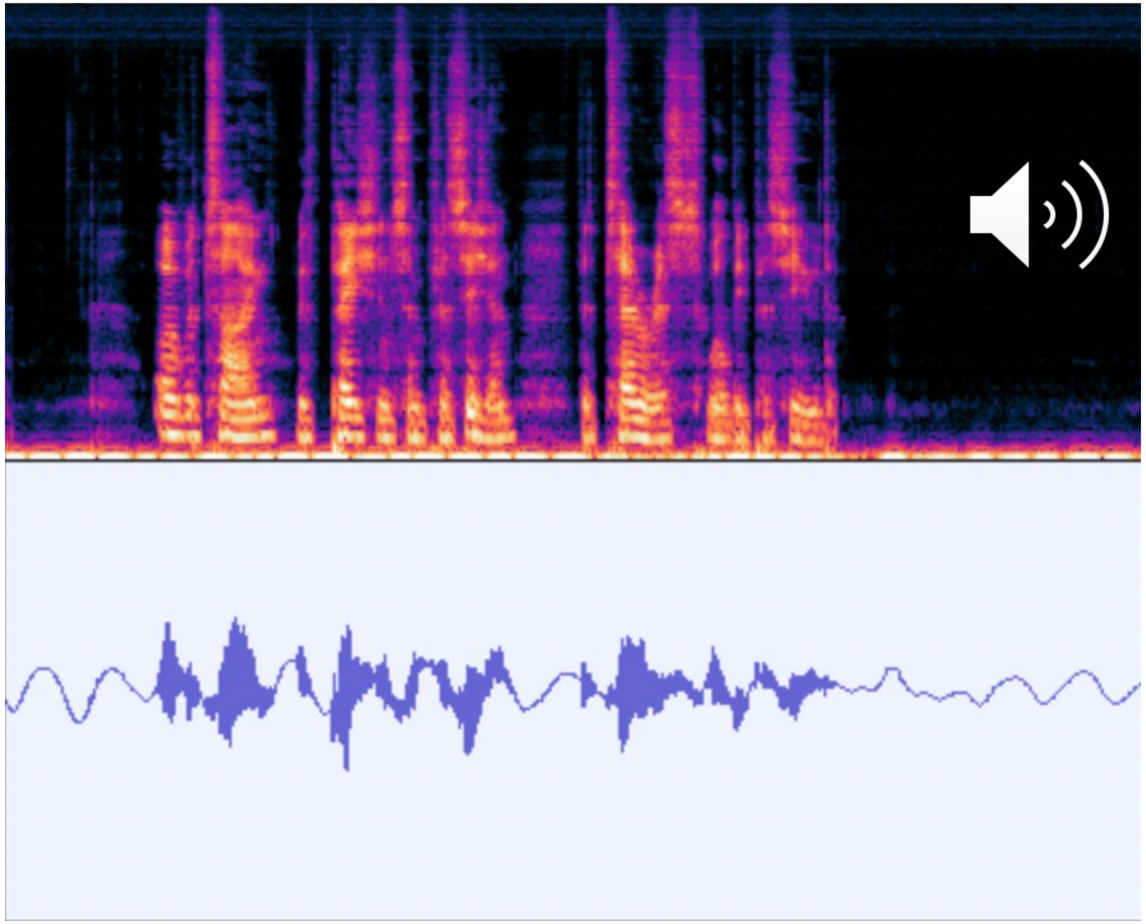
# Analysis: Data (II) – label noisiness

Estimated SNRs of the **enhanced version of speech labels** in training&validation sets

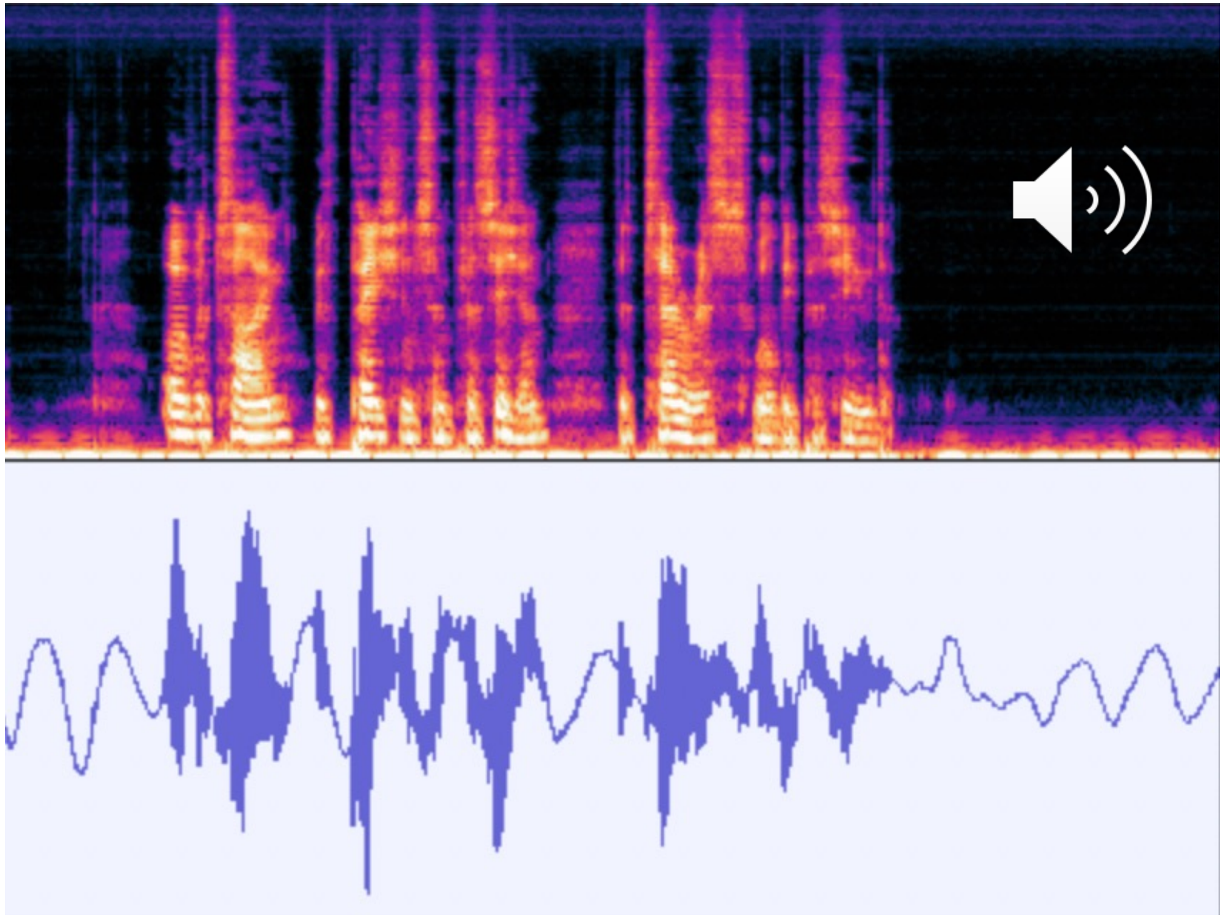


# Analysis: Data (II) – label noisiness

“Clean” speech label from VCTK (Original)

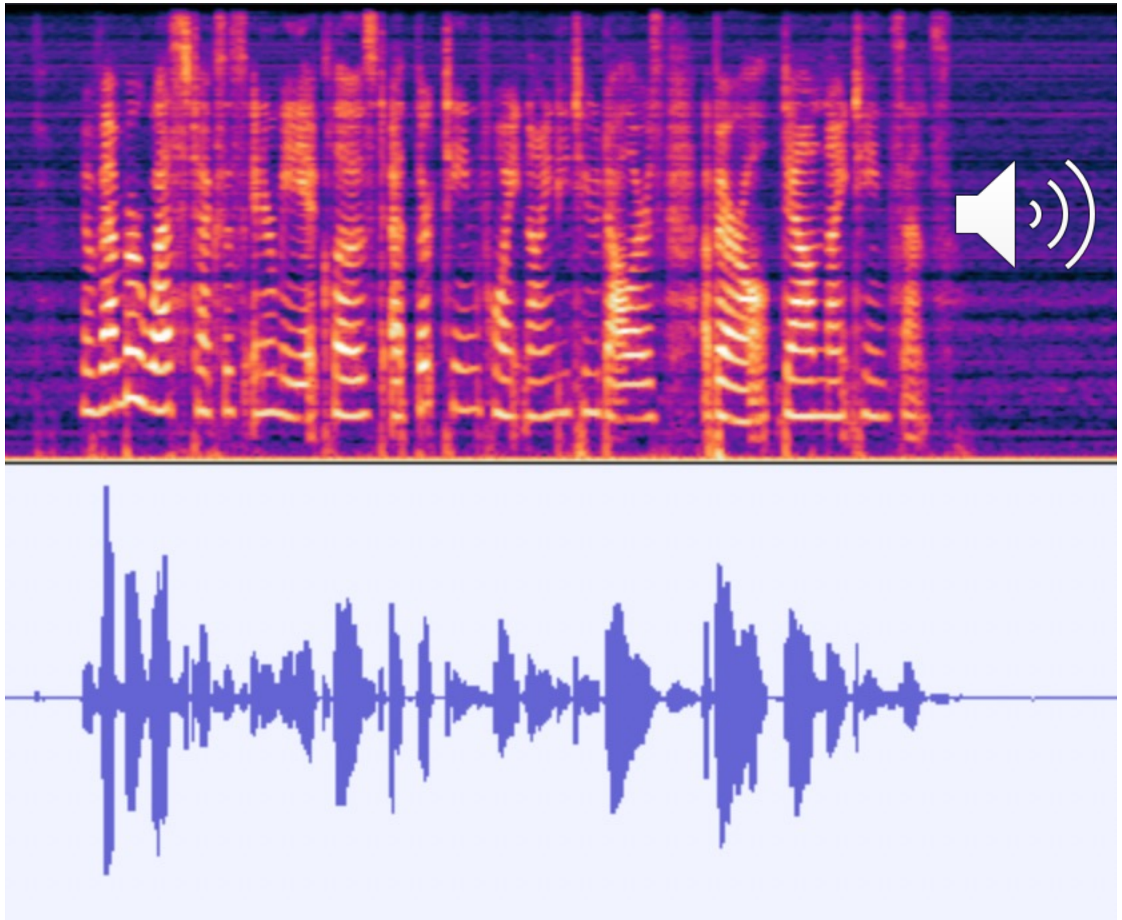


“Clean” speech label from VCTK (Enhanced version)

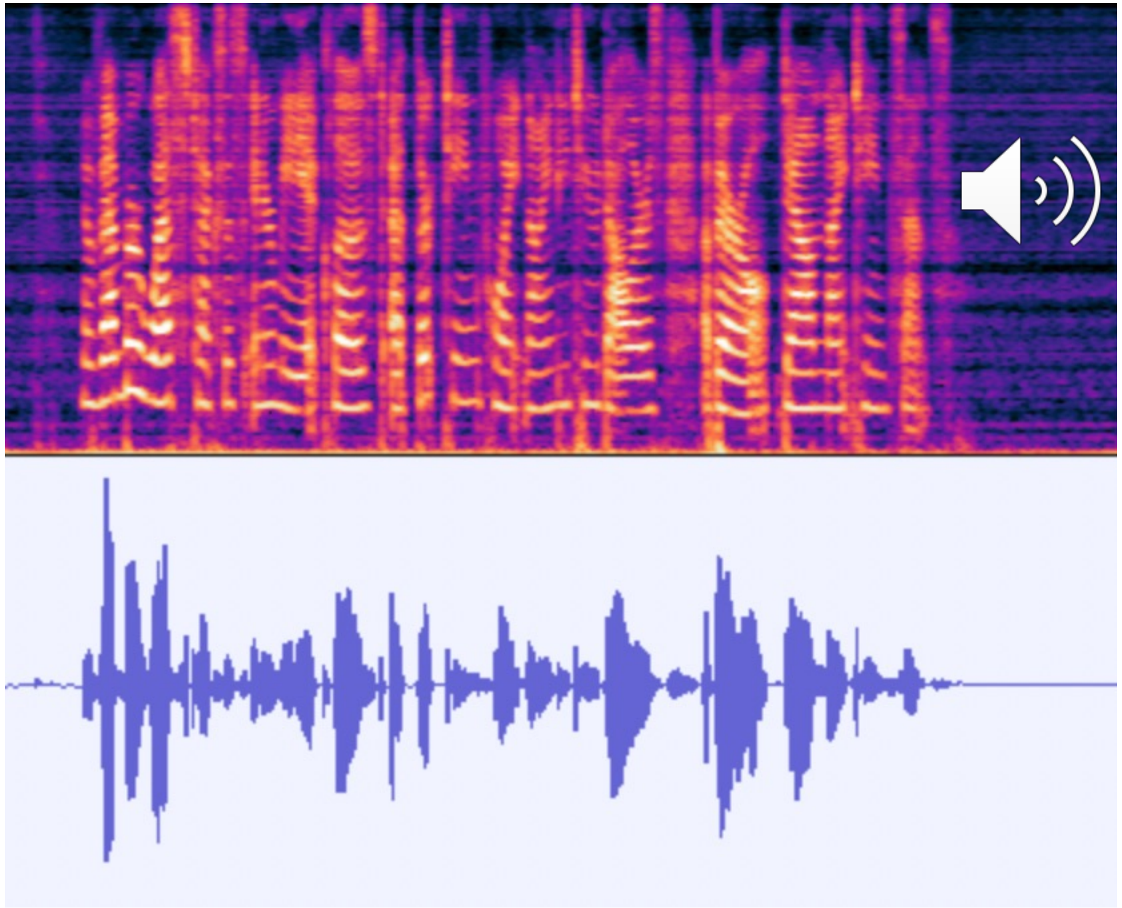


# Analysis: Data (II) – label noisiness

“Clean” speech label  from WSJ
(Original)



“Clean” speech label from WSJ
(Enhanced version)



# CONTENTS

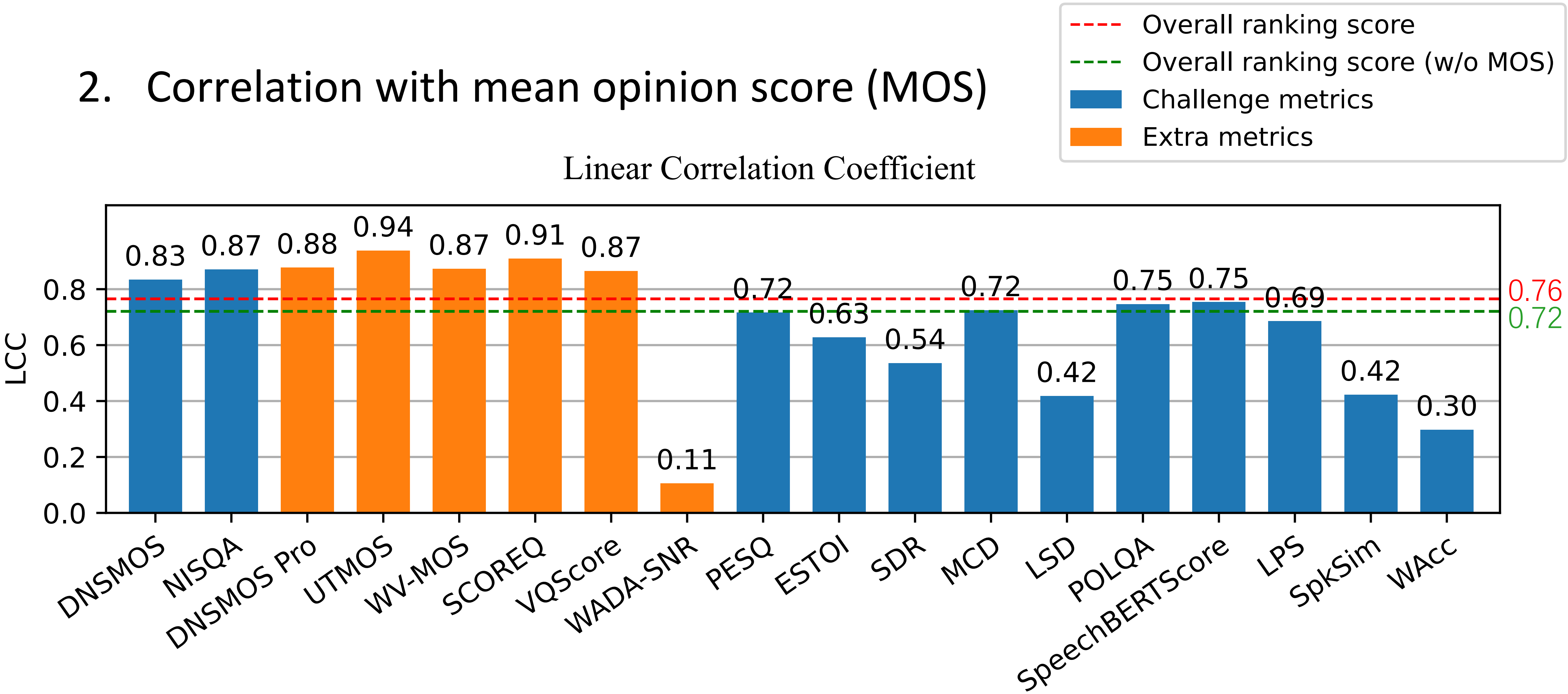
# Analysis: Evaluation Metrics

1. Final leaderboard https://urgent-challenge.com/competitions/5#final_results

| Rank | Team ID | *Non-intrusive SE metrics* | | *Intrusive SE metrics* | | | | | | *Downstream-task-indep.* | | *Downstream-task-dep.* | | *Subjective* | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **DNSMOS↑** | **NISQA↑** | **PESQ↑** | **ESTOI↑** | **SDR↑** | **MCD↓** | **LSD↓** | **POLQA↑** | **SBS.↑** | **LPS↑** | **SpkSim↑** | **WAcc (%)↑** | **MOS↑** | **ranking score↓** |
| 1 | T1 | 3.06 (2) | 3.66 (3) | 2.65 (3) | 0.87 (2) | **14.58 (1)** | **3.04 (1)** | 2.92 (7) | 3.51 (2) | 0.84 (3) | 0.82 (4) | 0.80 (3) | 73.57 (2) | **3.52 (1)** | 2.43 |
| 2 | T2 | 3.00 (6) | 3.59 (6) | **2.80 (1)** | **0.87 (1)** | 14.52 (2) | 3.15 (3) | 2.78 (4) | **3.69 (1)** | **0.85 (1)** | **0.83 (1)** | **0.82 (1)** | 72.91 (4) | 3.46 (3) | 2.90 |
| 3a | T3a | 2.98 (9) | 3.44 (7) | 2.55 (6) | 0.85 (4) | 13.31 (4) | 3.33 (6) | 2.99 (9) | 3.34 (6) | 0.84 (5) | 0.83 (2) | 0.77 (7) | **74.03 (1)** | 3.44 (4) | 5.07 |
| 3b | T3b | 2.95 (11) | 3.35 (11) | 2.66 (2) | 0.86 (3) | 13.54 (3) | 3.14 (2) | **2.70 (1)** | 3.45 (3) | 0.85 (2) | 0.83 (3) | 0.81 (2) | 73.10 (3) | 3.40 (7) | 5.07 |
| 4 | T4 | 2.98 (8) | 3.37 (10) | 2.60 (4) | 0.85 (5) | 13.14 (5) | 3.21 (4) | 2.75 (3) | 3.43 (4) | 0.84 (4) | 0.81 (5) | 0.78 (5) | 71.67 (5) | 3.34 (10) | 6.53 |
| 5 | T5 | 3.02 (4) | 3.60 (5) | 2.32 (9) | 0.82 (8) | 11.38 (10) | 3.34 (7) | 3.45 (14) | 3.16 (8) | 0.82 (9) | 0.78 (9) | 0.76 (8) | 67.96 (8) | 3.47 (2) | 6.57 |
| 6 | T6 | 3.00 (7) | 3.35 (12) | 2.52 (8) | 0.84 (6) | 12.63 (6) | 3.32 (5) | 2.92 (8) | 3.31 (7) | 0.83 (6) | 0.80 (6) | 0.78 (6) | 70.13 (6) | 3.41 (6) | 6.83 |
| 7 | T7 | 2.90 (16) | 3.38 (9) | 2.55 (5) | 0.83 (7) | 12.42 (7) | 3.61 (10) | 2.86 (5) | 3.36 (5) | 0.83 (7) | 0.79 (7) | 0.79 (4) | 69.19 (7) | 3.44 (5) | 7.30 |
| 8 | T8 | 2.96 (10) | 3.15 (15) | 2.55 (7) | 0.80 (11) | 10.72 (11) | 3.83 (11) | 2.73 (2) | 3.15 (9) | 0.81 (11) | 0.75 (11) | 0.74 (11) | 66.15 (13) | 3.36 (9) | 10.60 |
| 9 | T9 | 2.92 (14) | 3.42 (8) | 2.26 (11) | 0.80 (12) | 12.23 (8) | 4.12 (12) | 3.54 (16) | 3.04 (11) | 0.79 (12) | 0.74 (12) | 0.71 (12) | 67.03 (11) | 3.33 (11) | 11.43 |
| 10 | T10 | 2.88 (18) | 3.17 (14) | 2.32 (10) | 0.81 (9) | 11.50 (9) | 3.46 (8) | 3.00 (10) | 3.06 (10) | 0.82 (10) | 0.77 (10) | 0.75 (9) | 67.45 (10) | 3.24 (13) | 11.57 |
| 11 | T11 | 3.06 (3) | **3.94 (1)** | 1.88 (19) | 0.76 (15) | 7.49 (20) | 4.96 (20) | 4.76 (20) | 2.64 (17) | 0.75 (20) | 0.70 (17) | 0.58 (21) | 60.28 (19) | 3.39 (8) | 13.40 |
| 12 | T12 | 2.92 (12) | 2.47 (21) | 2.14 (12) | 0.80 (10) | 9.73 (15) | 3.53 (9) | 3.36 (13) | 2.74 (14) | 0.83 (8) | 0.78 (8) | 0.75 (10) | 67.68 (9) | 2.87 (21) | 13.43 |
| 13 | T13 | 2.89 (17) | 3.23 (13) | 2.03 (16) | 0.77 (14) | 10.43 (13) | 4.63 (16) | 3.83 (19) | 2.69 (15) | 0.77 (14) | 0.72 (14) | 0.67 (16) | 62.68 (15) | 3.32 (12) | 14.40 |
| 14 | T14 | 2.88 (19) | 2.95 (18) | 2.13 (13) | 0.78 (13) | 10.62 (12) | 4.13 (13) | 3.24 (12) | 2.89 (12) | 0.77 (13) | 0.73 (13) | 0.70 (13) | 66.89 (12) | 3.06 (17) | 14.70 |
| 15 | Baseline | 2.83 (21) | 3.07 (17) | 2.07 (14) | 0.76 (16) | 10.13 (14) | 4.22 (15) | 3.09 (11) | 2.81 (13) | 0.77 (16) | 0.70 (16) | 0.70 (14) | 62.97 (14) | 3.12 (16) | 15.77 |
| 16 | T16 | 2.92 (13) | 2.73 (19) | 2.04 (15) | 0.76 (17) | 9.47 (16) | 4.82 (19) | 3.55 (17) | 2.66 (16) | 0.77 (15) | 0.71 (15) | 0.67 (17) | 62.24 (16) | 2.95 (19) | 16.63 |
| 17 | T17 | **3.26 (1)** | 3.83 (2) | 1.36 (22) | 0.60 (21) | 0.41 (22) | 6.27 (21) | 5.43 (21) | 1.74 (22) | 0.68 (21) | 0.56 (21) | 0.48 (23) | 40.73 (21) | 3.05 (18) | 16.80 |
| 18 | T18 | 3.02 (5) | 3.61 (4) | 1.47 (21) | 0.51 (23) | -6.16 (23) | 8.44 (22) | 7.12 (23) | 1.93 (21) | 0.67 (22) | 0.53 (22) | 0.54 (22) | 32.08 (22) | 3.17 (15) | 17.13 |
| 19 | T19 | 2.85 (20) | 3.12 (16) | 1.97 (18) | 0.74 (18) | 9.43 (17) | 4.65 (17) | 3.74 (18) | 2.59 (18) | 0.76 (18) | 0.69 (18) | 0.67 (18) | 60.28 (19) | 3.21 (14) | 17.23 |
| 20 | T20 | 2.91 (15) | 2.55 (20) | 2.00 (17) | 0.73 (19) | 9.03 (19) | 4.18 (14) | 2.89 (6) | 2.57 (19) | 0.77 (17) | 0.68 (20) | 0.68 (15) | 60.64 (18) | 2.91 (20) | 17.63 |
| 21 | T21 | 2.53 (22) | 2.39 (22) | 1.84 (20) | 0.73 (20) | 9.08 (18) | 4.74 (18) | 3.51 (15) | 2.47 (20) | 0.75 (19) | 0.68 (19) | 0.65 (19) | 59.95 (20) | 2.82 (22) | 20.20 |
| 22 | Noisy input | 1.70 (23) | 1.53 (23) | 1.26 (23) | 0.58 (22) | 0.98 (21) | 9.71 (23) | 5.46 (22) | 1.58 (23) | 0.59 (23) | 0.52 (23) | 0.64 (20) | 61.92 (17) | 1.88 (23) | 21.97 |

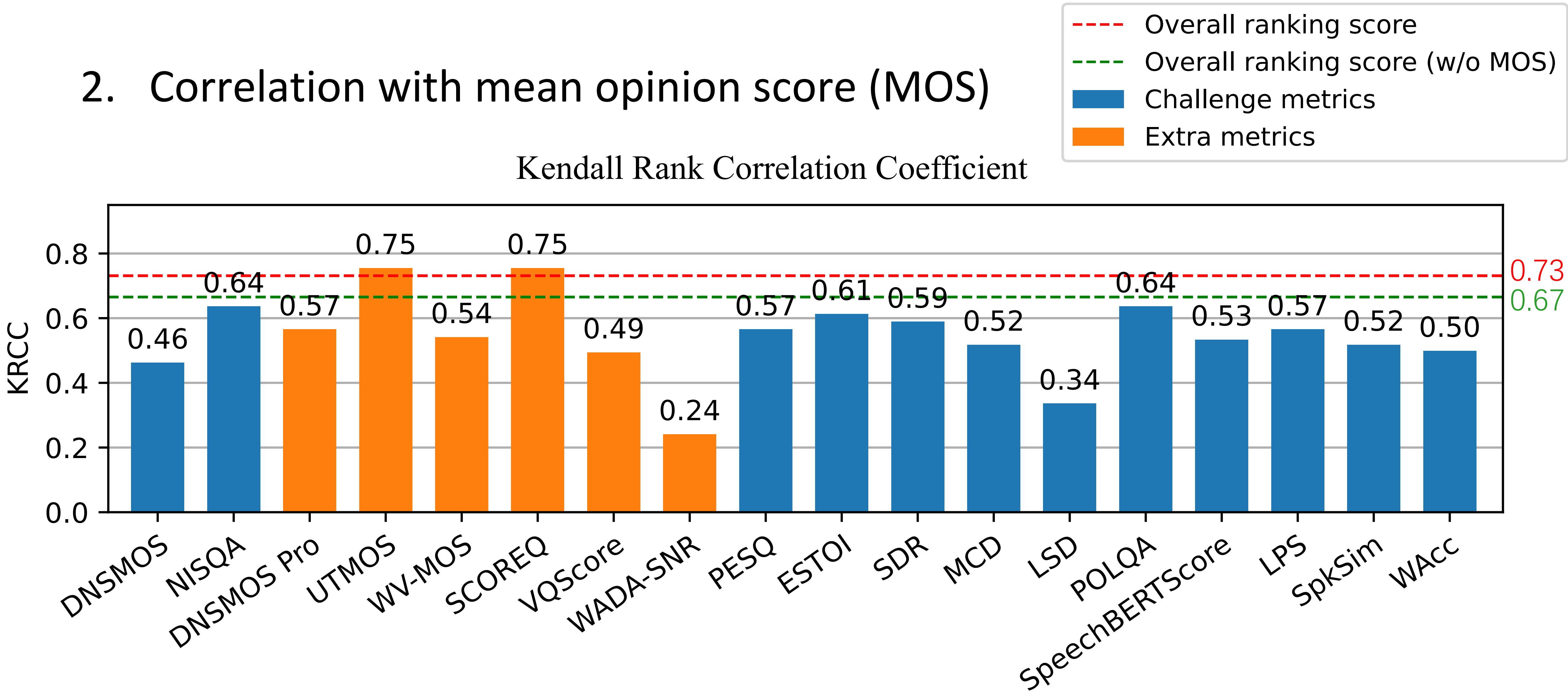
# Analysis: Evaluation Metrics

2. Correlation with mean opinion score (MOS)



# Analysis: Evaluation Metrics

2. Correlation with mean opinion score (MOS)



# Takeaways

- It is feasible to build a single universal SE system to handle various
  - Sampling rates
  - SE subtasks (e.g.., denoising, dereverberation, declipping, bandwidth extension)
- Data quality (effective bandwidth, label noisiness, etc.) might be an obstacle to improving SE performance.
- Another comprehensive summary paper is submitted to NeurIPS 2025, containing details of the top-performing systems and a new SQA dataset.
- What to explore next?
  - ❖ More languages, more distortions, more diverse data, etc.
  - ❖ ⇒ **2nd URGENT Challenge**