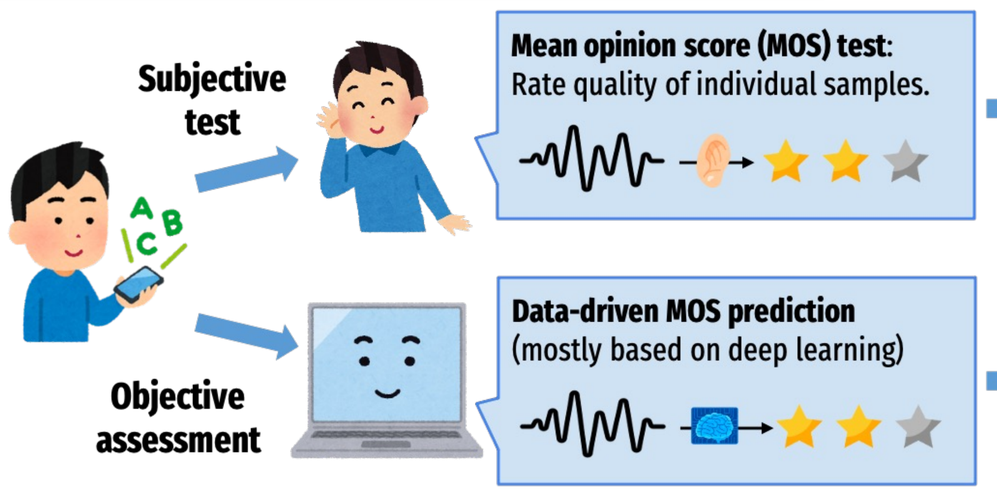
The VoiceMOS Challenge 2024: Beyond Speech Quality Prediction

P4-26-SS05 (#396)

Wen-Chin Huang[1], Szu-Wei Fu[2], Erica Cooper[3], Ryandhimas E. Zezario[4], Tomoki Toda[1], Hsin-Min Wang[4], Junichi Yamagishi[5], Yu Tsao[4]

[1]Nagoya University, JP [2]NVIDIA Taiwan [3]NICT, JP [4]Acadamia Sinica, Taiwan [5]NII, JP

The VoiceMOS Challenge (VMC) series



Generalizable?

2022

- Focus: synthetic speech, supervised setting
- Datasets: BVCC dataset & Blizzard Challenge (BC) '19
  - Large-scale re-evaluation of TTS & VC samples since '08
  - Best system: .979/.975 system-level SRCC

→ Performs well in the supervised setting

2023

- Focus: zero-shot setting
- Tracks: Blizzard challenge (French TTS), Singing Voice Conversion Challenge, clean/noisy/enhanced speech
- Result 1: gap between supervised & zero-shot setting
- Result 2: no consistent performance across all tracks

Track 1: MOS prediction for "zoomed-in" systems

- Motivation: evaluate synthetic systems of high-quality
- New listening tests using the top 50%, 25%, 12% systems in BVCC
- 50% -> validation set; 25% & 12% -> test set

Track 2: MOS prediction for singing voice

- A newly collected dataset named SingMOS: natural singing voices, vocoder analysis-synthesis, singing voice synthesis/conversion samples
- Mandarin & Japanese, 16kHz, 35 systems, 2000/544/645 samples

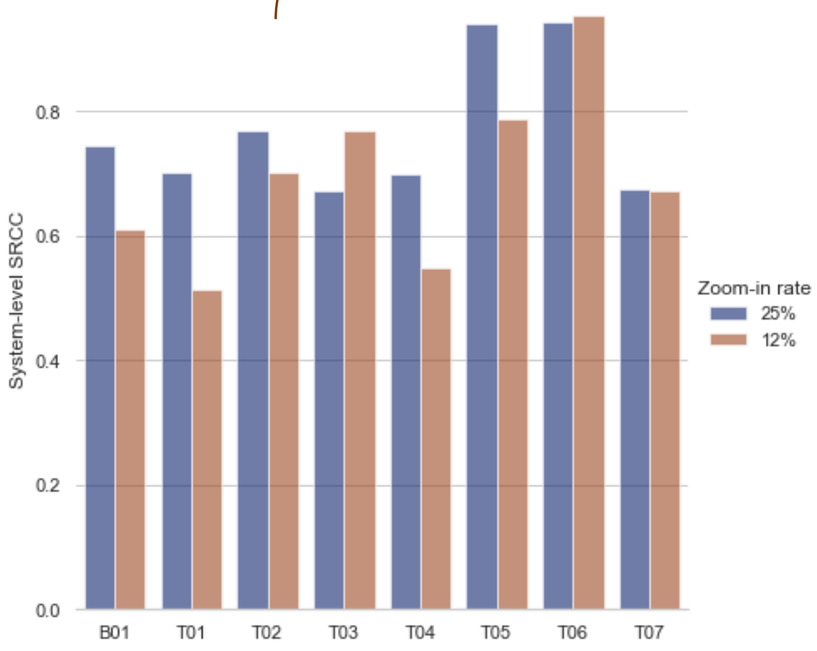
Track 3: semi-supervised MOS prediction for clean/noisy/enhanced speech

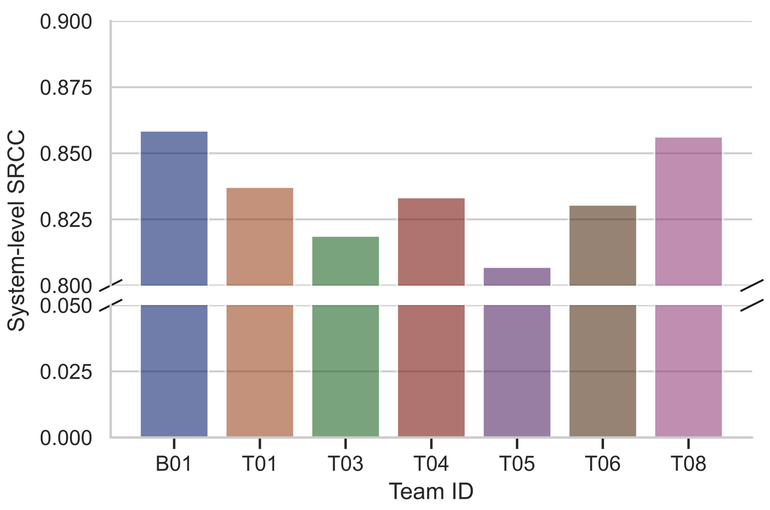
- Setting: very limited amount of training data & zero-shot setting
- Train/valid set: UDASE task of 7th CHiME, 60/40 samples (real noisy samples)
  Test set: VoiceBank-DEMAND, 4 noise types, 5 enhancement systems, 280 samples (artificial samples)
- Beyond quality: speech signal quality (SIG), background intrusiveness (BAK), overall quality (OVRL)

Participants, baselines

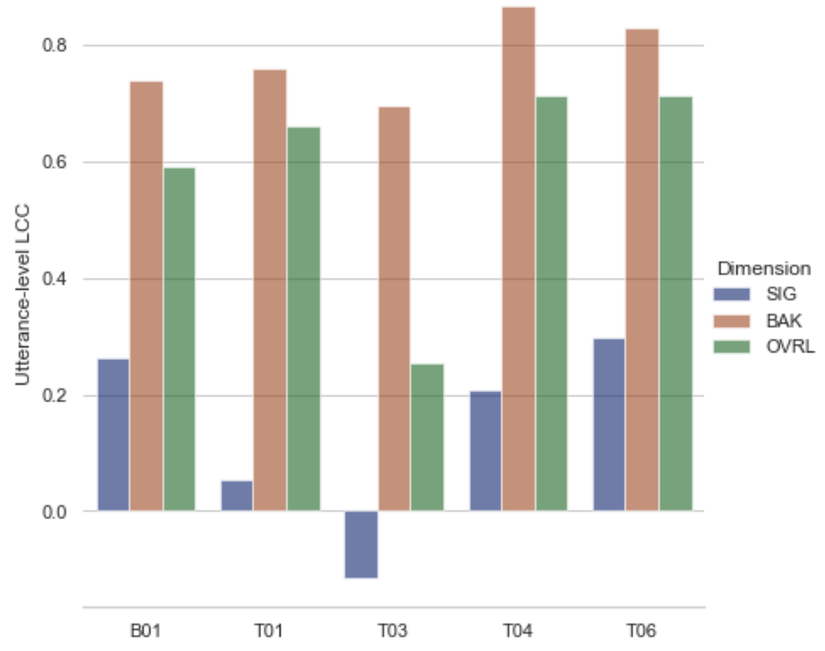
- 8 teams (5 academia, 3 industry)
- Baselines for tracks 1 & 2: SSL-MOS
- Baseline for track 3: VQScore

Main metric:
system-level Spearman correlation coeff. (SRCC)





Main metric:
utterance-level linear correlation coeff. (LCC)



Track 1:

- 12% is harder than 25%
- Baseline (B01) ranked 4th/6th in 25%/12%
  → participants have advanced
- Top systems: T05 & T06

Track 2:

- No team outperformed the baseline (B01)
- Differences were small
- T06 ranked 1st in all utterance-level metrics

Track 3:

- Baseline (B01) was outperformed
- SIG is the most difficult to predict
- No team exceled all aspects
- T06: 1st in SIG & OVRL; T04: 1st in BAK

Top system: T06

- Performed remarkably well in all three tracks.
- Improved version of RAMP: equipping a parametric model (e.g., SSL-MOS) with a non-parametric head based on kNNs.
- Was shown to generalize well to unseen data.

Top system: T05 (P4-28-SS05 (#407))

- Top system in track 1.
- SSL feature + mel spectrogram (EfficientNetV2 encoder).
- Conducted own listening test.

Top system: T08 (P4-27-SS05 (#406))

- Top system in track 2.
- SSL feature + pitch histogram.

Top system: T04

- Top system in track 3.
- Trained separate models for BAK and SIG prediction. OVRL = (BAK+SIG)/2.
- BAK predictor: pre-trained to predict SNR of simulated noisy speech samples.
- SIG predictor: pre-trained to predict spoofed and natural samples from ASVSpoof 2019.
- Both are fine-tuned on the provided training data.

Future directions

- Modern-day speech synthesis systems
- More diverse speech types
- Beyond speech: music, environmental sounds

Challenge HP