The VoiceMOS Challenge 2023:

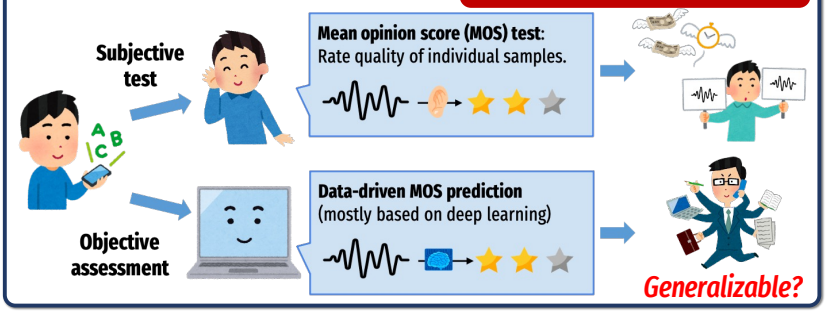
Zero-shot Subjective Speech Quality Prediction for Multiple Domains

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, Junichi Yamagishi
Nagoya University; National Institute of Informatics, Japan; Academia Sinica, Taiwan

2023 ASRU Workshop on Automatic Speech Recognition and Understanding

The VoiceMOS Challenge (VMC)

Focuses on automatic quality assessment for synthetic speech



VoiceMOS Challenge 2022

- Attracted more than 20 participants
- Main track: the BVCC dataset
  - Large-scale re-evaluation of TTS & VC samples from 2008
  - Best system: .979 system-level SRCC

  → Current technology performs well in an in-domain setting
- Out-of-domain track: Blizzard Challenge 2019
  - Chinese TTS; Small amount of labeled data (136 samples)
  - Best system: .975 system-level SRCC

  → Current technology performs well in a fine-tuning setting

Still far from a truly generalizable, zero-shot predictor!

Track 1: French text-to-speech

- In collaboration with Blizzard Challenge (TTS challenge founded in 2008); Theme this year: French
- Two sub tracks: (1a) Speaker-dependent; (1b) Speaker-adaptive

Track 2: Singing voice conversion

- In collaboration with the Voice Conversion Challenge (a VC challenge founded in 2016)
- Theme this year: singing voice conversion

Track 3: noisy/enhanced speech

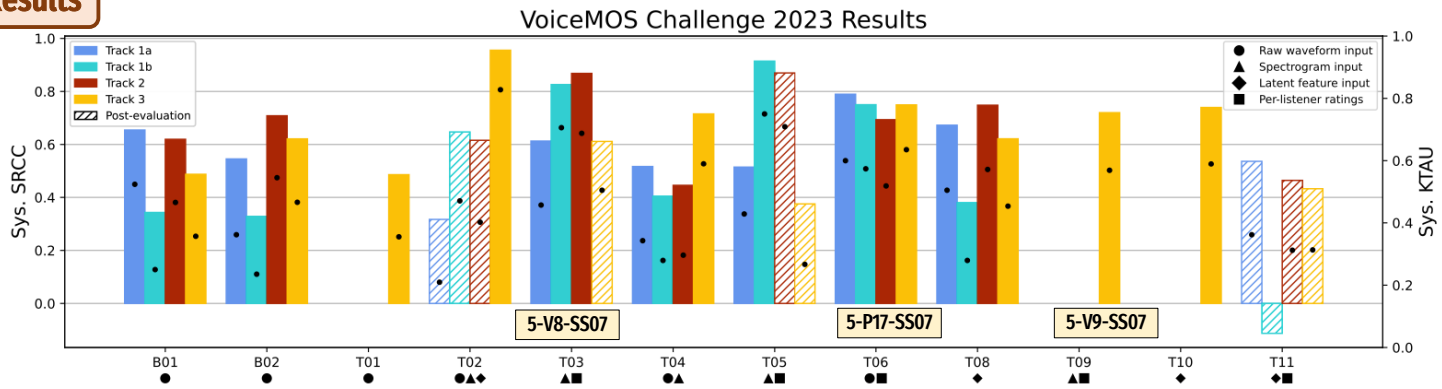
- Dataset: TMHINT-QI (Mandarin)
- Training set
  - 4 noise types (babble, street, pink, and white) at 4 SNR levels (-2, 0, 2, and 5)
  - 5 enhancement (SE) models: KLT, MMSE, FCN, DDAE, and Transformer
- Testing set
  - Same noise generation process as in training set
  - 5 SE models: MMSE, FCN, Trans, DEMUCS, and CMGAN

Datasets

No official training set for tracks 1 & 2!

Track	Type	Lang	Systems	Samples per system	# ratings per sample
Track 1a Track 1b	TTS	Fre	Hub: 21 Spoke: 17	42 34	15
Track 2	Singing VC	Eng	In-dom: 25 Cross-dom: 24	80	6
Track 3	Noisy & enhanced	Chi	97	20	5.3

Results



Best vs. Average scores

Best score for each track / *Average score for each track*

Track	SRCC	KTAU	Team	Avg SRCC	Avg KTAU
1a	0.79	0.60	T06	0.57	0.42
1b	0.91	0.75	T05	0.50	0.39
2	0.87	0.69	T03	0.67	0.50
3	0.95	0.83	T02	0.63	0.49

Track 3 has highest best score

Track 2 has highest average score

- The goal of having **one model that can predict MOS well for different domains** has **not yet been reached**.
- **Domain mismatch** between singing voice conversion and text-to-speech synthesis was **not as large** as we expected.
- The **difference** in predictability between **speaker-dependent** and **speaker-adaptive** French TTS was surprising.
- Approaches using **listener information** and a **mix of training datasets** tended to be more successful.
- We can **observe a gap** between the case where a **small amount of in-domain training data** is available, and the **completely zero-shot setting**.

Feedback

- Unfamiliarity with French TTS, Singing voice and speech enhancement is a difficulty
- High listener variance in Track 3
- Not sure what is a proper training set for tracks 1 & 2

Future directions

- Keep on improving zero-shot ability
- Moving from MOS to other listening tests (MUSHRA, AB preference tests, ...)
- More fine-grained instructions (expressivity, etc.)

Challenge HP