The VoiceMOS Challenge 2022

Wen-Chin Huang[1], Erica Cooper[2], Yu Tsao[3], Hsin-Min Wang[3], Tomoki Toda[1], Junichi Yamagishi[2]
[1]Nagoya University, Japan
[2]National Institute of Informatics, Japan
[3]Academia Sinica, Taiwan

Outline

I. Introduction
II. Challenge description
    A. Tracks and datasets
    B. Rules and timeline
    C. Evaluation metrics
    D. Baseline systems
III. Challenge results
    A. Participants demographics
    B. Results, analysis and discussion
        1. Comparison of baseline systems
        2. Analysis of top systems
        3. Sources of difficulty
        4. Analysis of metrics
IV. Conclusions

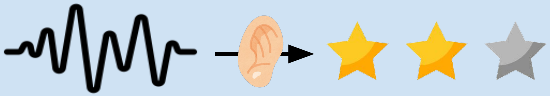
Introduction

Speech quality assessment

Important to evaluate speech synthesis systems, ex. text-to-speech (TTS), voice conversion (VC).

Subjective test

Mean opinion score (MOS) test:
Rate quality of individual samples.

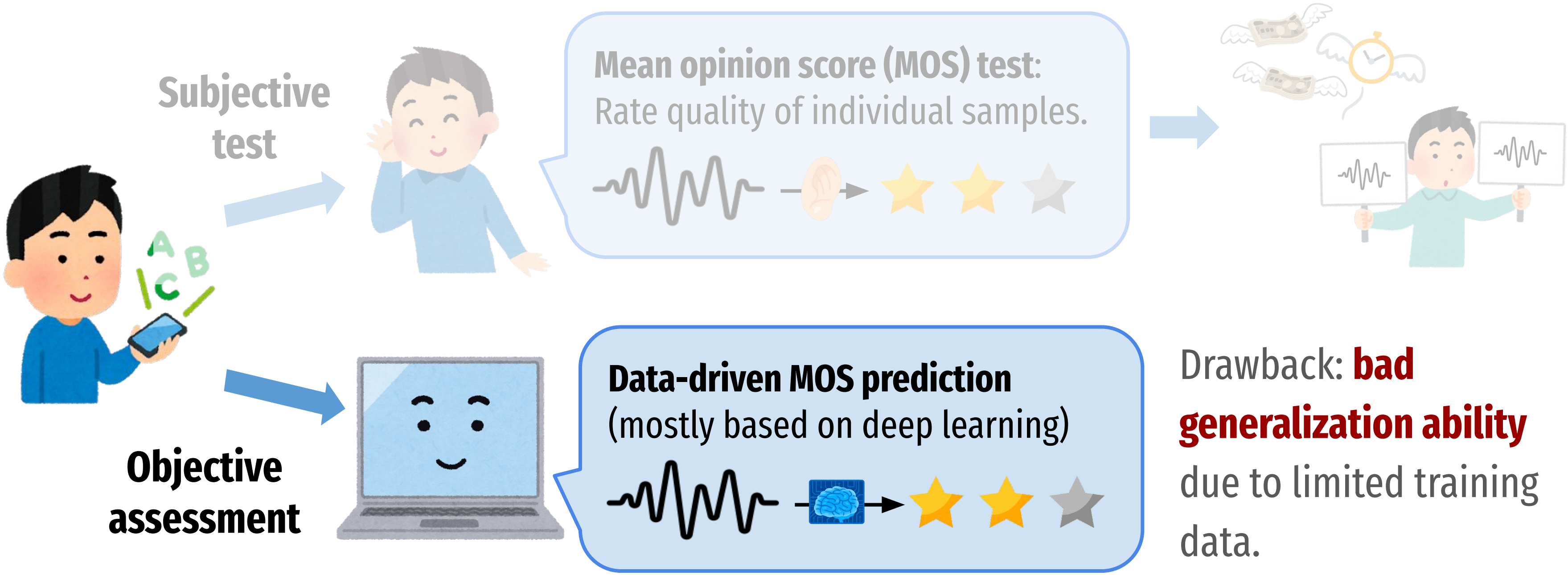


Drawbacks:

1. **Expensive**: Costs too much time and money.
2. **Context-dependent**: numbers cannot be meaningfully compared across different listening tests.

Speech quality assessment

Important to evaluate speech synthesis systems, ex. text-to-speech (TTS), voice conversion (VC).

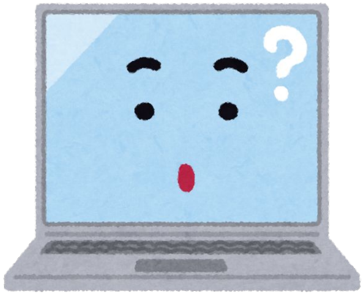


Drawback: **bad generalization ability** due to limited training data.

Goals of the VoiceMOS challenge

Encourage research in automatic data-driven MOS prediction

Compare different approaches using **shared datasets and evaluation**



Focus on the challenging case of **generalizing to a separate listening test**

Promote discussion about the future of this research field

*Accepted as an **Interspeech 2022 special session**!

Challenge description

I. Tracks and datasets
II. Rules and timeline
III. Evaluation metrics
IV. Baseline systems

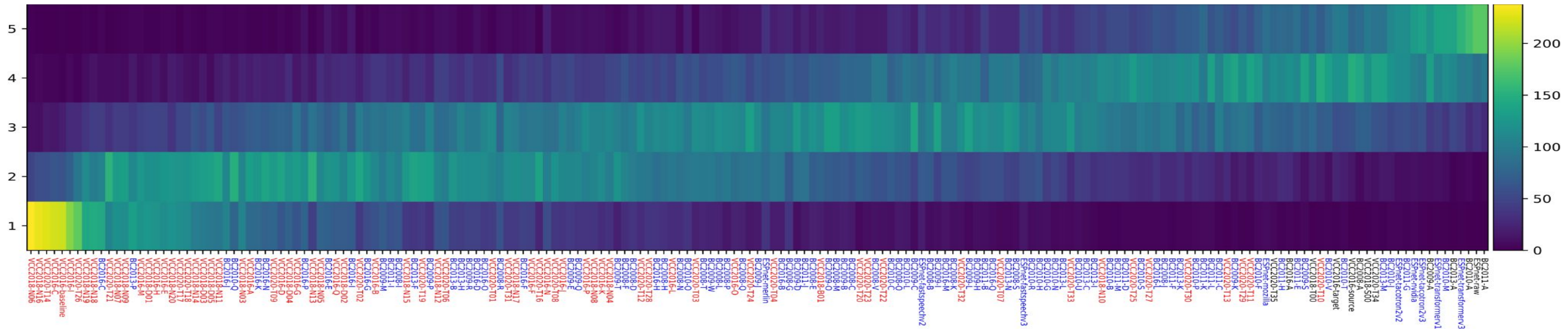
Challenge platform: CodaLab

Open-source web-based platform for reproducible machine learning research.

Tracks and dataset: Main track

The BVCC Dataset

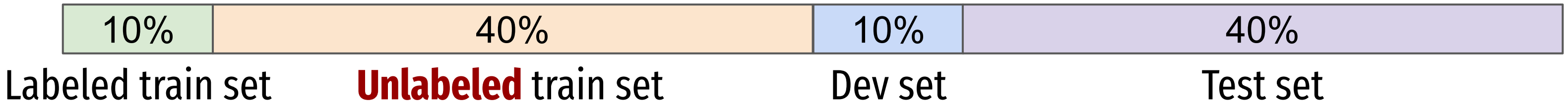
- Samples from 187 different systems all rated together in one listening test
  - Past Blizzard Challenges (text-to-speech synthesis) since 2008
  - Past Voice Conversion Challenges (voice conversion) since 2016
  - ESPnet-TTS (implementations of modern TTS systems), 2020
- Test set contains some **unseen systems, unseen listeners, and unseen speakers** and is balanced to match the distribution of scores in the training set



Tracks and dataset: OOD track

Listening test data from the Blizzard Challenge 2019

- “Out-of-domain” (OOD): **Data from a completely separate listening test**
- Chinese-language synthesis from systems submitted to the 2019 Blizzard Challenge
- Test set has some **unseen systems** and **unseen listeners**



- Designed to reflect a real-world setting where a small amount of labeled data is available
- Study generalization ability to a different listening test context
- Encourage unsupervised and semi-supervised approaches using unlabeled data

Dataset summary

Table 1: *Summary of the main track and out-of-domain (OOD) track datasets.*

Track	Lang	# Samples			# ratings per sample
		Train	Dev	Test	
Main	Eng	4,974	1,066	1,066	8
OOD	Chi	Label: 136 Unlabel: 540	136	540	10-17

Rules and timeline

2021/12/5 → 2022/2/21 → 2022/2/28 → 2022/3/7 →

Training phase

✔Main track train/dev audio+rating
✔OOD track train/dev audio+rating
Leaderboard available
Need to submit at least once

Evaluation phase

✔Main track test audio
✔OOD track test audio
Leaderboard available

Break phase

No submission available.
Conduct result analysis.

Post-evaluation phase

✔Main track test rating
✔OOD track test rating
✔Evaluation results
Submission & leaderboard reopen

General rules

External data is permitted.
All participants need to submit system description.

Evaluation metrics

System-level and **Utterance-level**

- **Mean Squared Error (MSE):** difference between predicted and actual MOS
- **Linear Correlation Coefficient (LCC):** a basic correlation measure
- **Spearman Rank Correlation Coefficient (SRCC):** non-parametric; measures **ranking order**
- **Kendall Tau Rank Correlation (KTAU):** more robust to errors

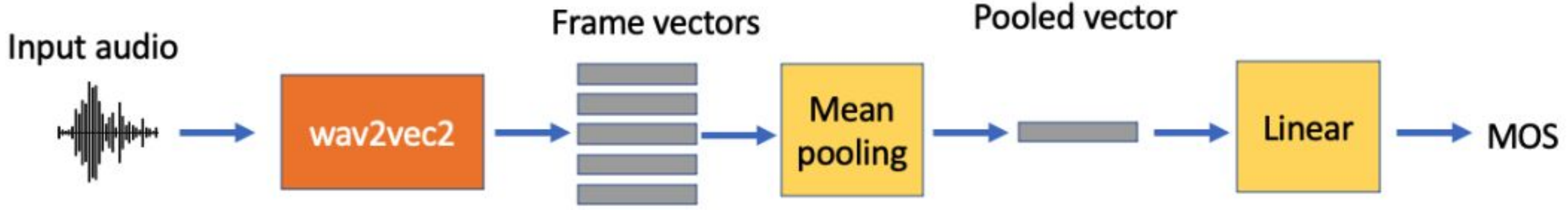
```python
import numpy as np
import scipy.stats

# `true_mean_scores` and `predict_mean_scores` are both 1-d numpy arrays.

MSE = np.mean((true_mean_scores - predict_mean_scores)**2)
LCC = np.corrcoef(true_mean_scores, predict_mean_scores)[0][1]
SRCC = scipy.stats.spearmanr(true_mean_scores, predict_mean_scores)[0]
KTAU = scipy.stats.kendalltau(true_mean_scores, predict_mean_scores)[0]
```

Following prior work, we picked **system-level SRCC** as the main evaluation metric.

Baseline system: SSL-MOS

Fine-tune a self-supervised learning based (SSL) speech model for the MOS prediction task

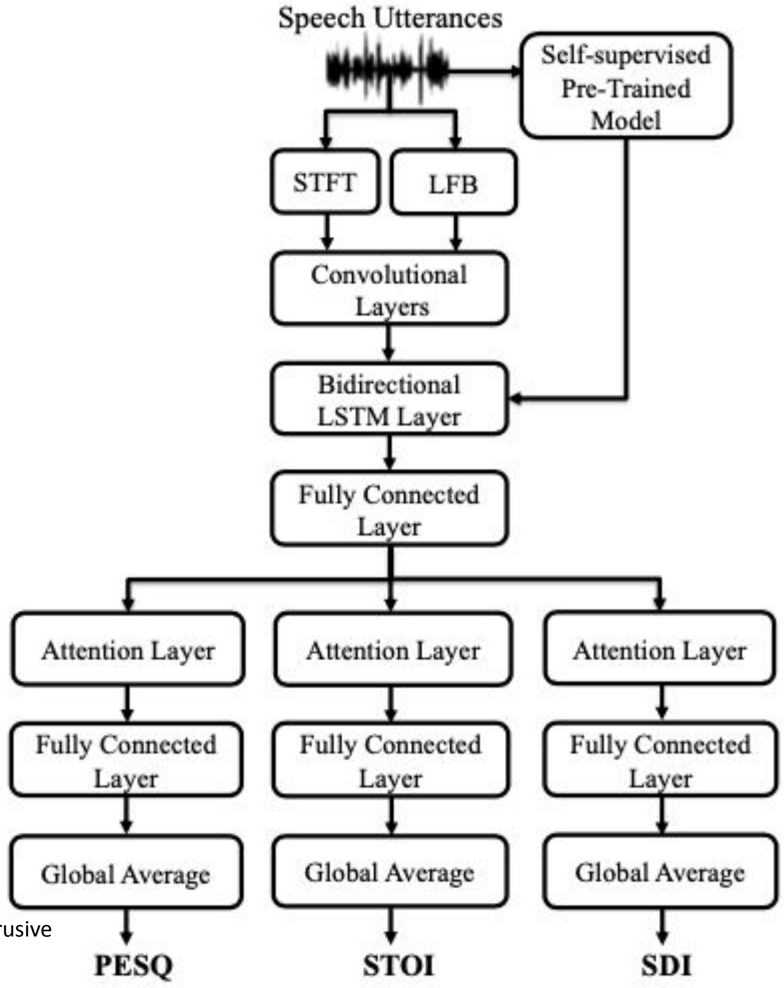
- Pretrained wav2vec2
- Simple mean pooling and a linear fine-tuning layer
- Wav2vec2 model parameters are updated during fine-tuning



E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in Proc. ICASSP, 2022

Baseline system: MOSANet

- Originally developed for noisy speech assessment
- Cross-domain input features:
    - Spectral information
    - Complex features
    - Raw waveform
    - Features extracted from SSL models

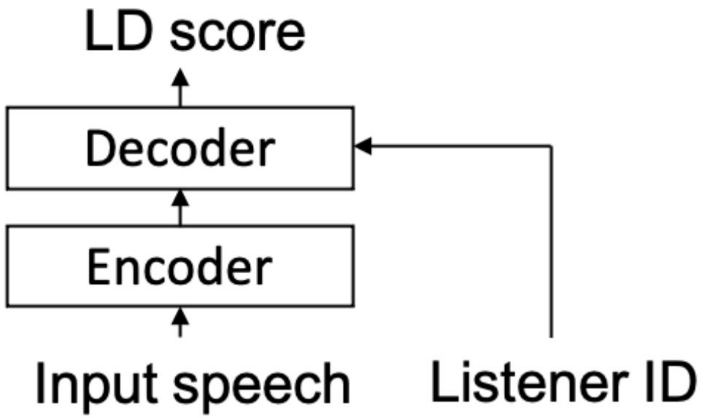


R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, “Deep Learning-based Non-Intrusive Multi-Objective Speech Assessment Model with Cross-Domain Features,” arXiv preprint arXiv:2111.02363, 2021.

Baseline system: LDNet

Listener-dependent modeling

- Specialized model structure and inference method allows **making use of multiple ratings per audio sample.**
- **No external data is used!**



W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, “LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech,” in Proc. ICASSP, 2022

Challenge results

I. Participants demographics
II. Results, analysis and discussion
   A. Comparison of baseline systems
   B. Analysis of top systems
   C. Sources of difficulty
   D. Analysis of metrics

Participants demographics

Number of teams: 22 teams + 3 baselines

14 teams are from academia, 5 teams are from industry, 3 teams are personal

Main track: 21 teams + 3 baselines

OOD track: 15 teams + 3 baselines

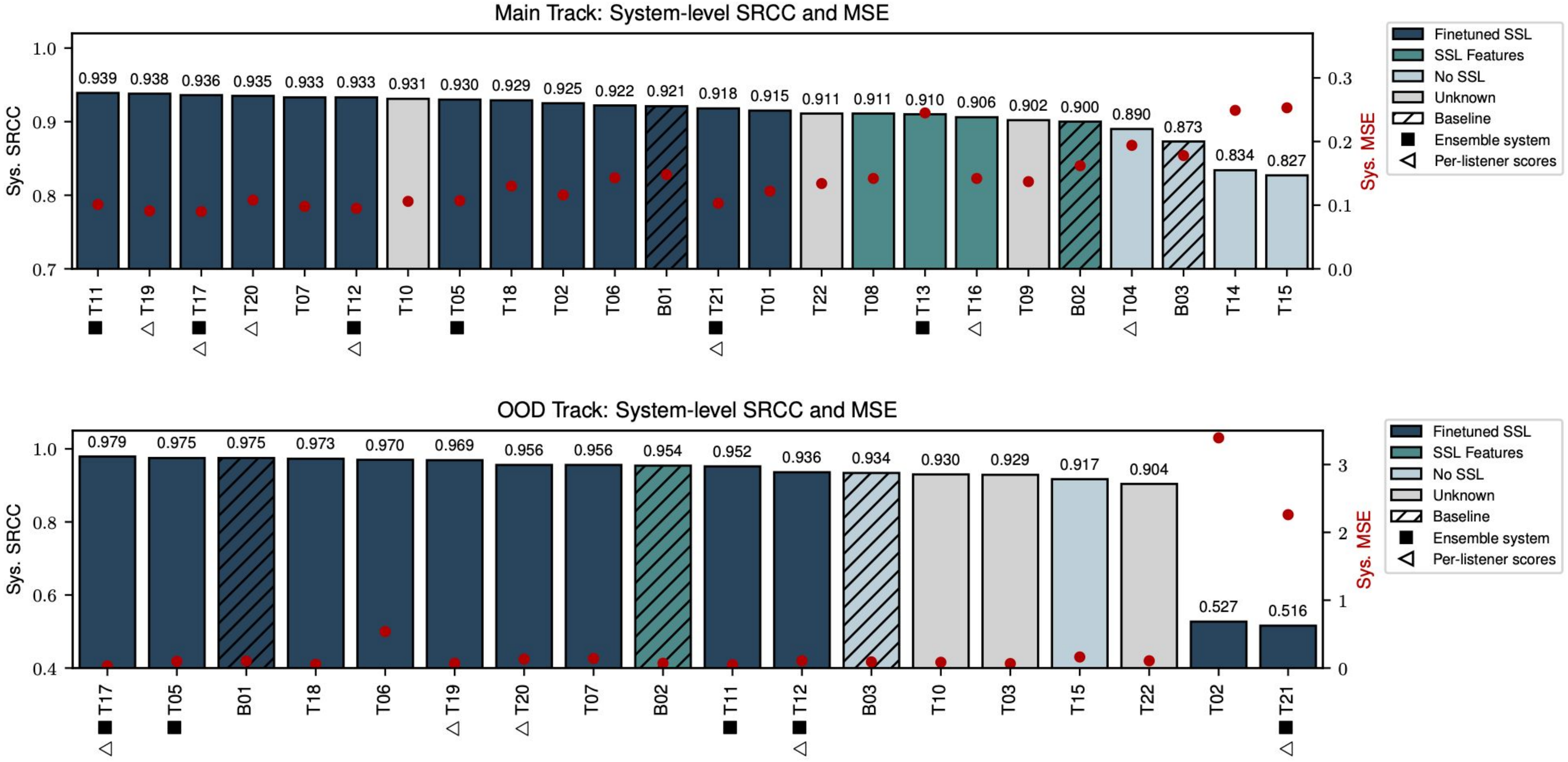
Baseline systems:

- B01: SSL-MOS
- B02: MOSANet
- B03: LDNet

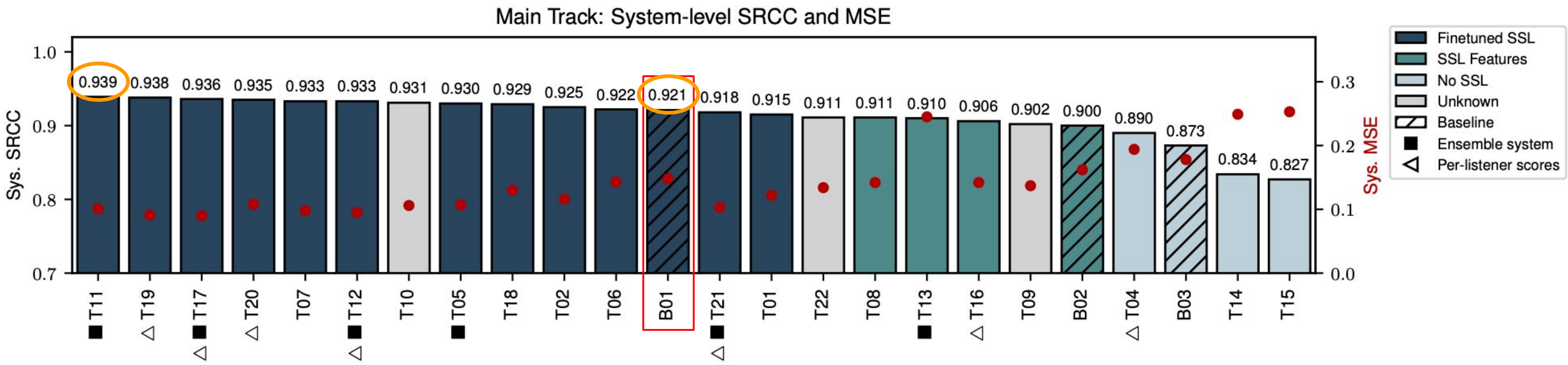
Table 4: *List of participant affiliations in random order.*

Affiliation	Main track	OOD track
Ajmide Media, China	Y	Y
Budapest University of Technology and Economics, Hungary	Y	Y
Bytedance AI-Lab, China	Y	Y
Charles University, Prague, Czech Republic	Y	N
Denso IT Laboratory, Japan	Y	Y
Duke Kunshan University	Y	N
Google; University College Dublin	Y	N
Inner Mongolia University, China	Y	N
Japan Advanced Institute of Science and Technology, Japan	Y	N
National Taiwan University, Taiwan	Y	Y
Netease, China	Y	Y
NICT, Japan; Kyoto Univ., Japan; Kuaishou Inc., China	Y	Y
Novosibirsk State University	N	Y
Personal?	Y	Y
Princeton University	Y	Y
ReadSpeaker, The Netherlands	Y	N
Sillwood Technologies, UK	Y	Y
Technical University of Cluj-Napoca, Romania	Y	N
The University of Tokyo, Japan	Y	Y
Tsinghua University?	Y	Y
University College Dublin, Ireland	Y	Y
University of West Bohemia, Czech Republic	Y	Y

Overall evaluation results: main track, OOD track



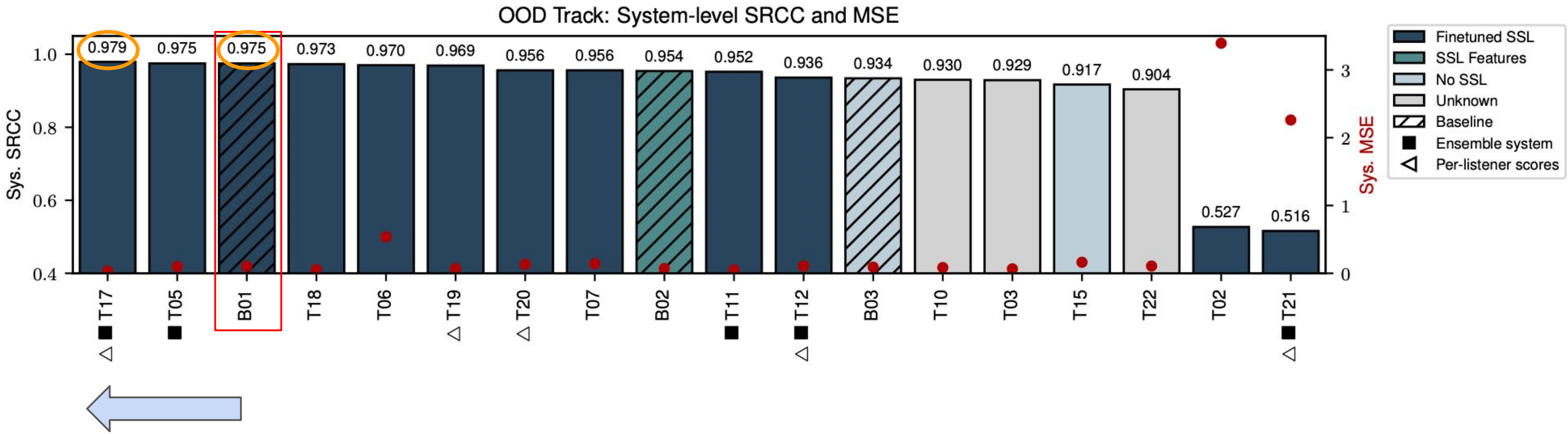
Comparison of baseline systems: main track



In terms of **system-level SRCC,** 11 teams outperformed the best baseline, B01!

However, the gap between the best baseline and the top system is not large...

Comparison of baseline systems: OOD track



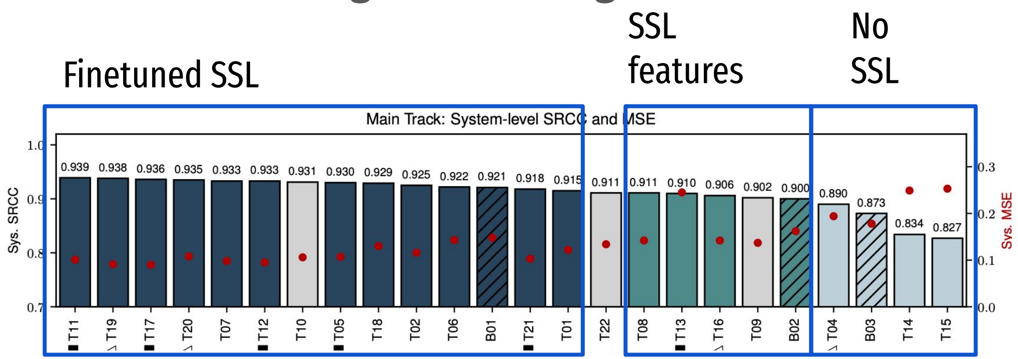
In terms of **system-level SRCC,** only 2 teams outperformed or on par with B01.

The gap is even smaller...

Participant feedback:
"The baseline was too strong! Hard to get improvement!"

Analysis of approaches used

- Main track: Finetuning SSL > using SSL features > not using SSL



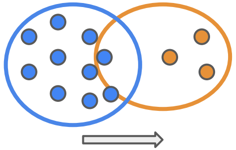
- OOD track: finetuned SSL models were both the best and worst systems
- Popular approaches:
  - **Ensembling** (top team in main track; top 2 teams in OOD track)
  - **Multi-task learning**
  - **Use of speech recognizers** (top team in OOD track)

Analysis of approaches used

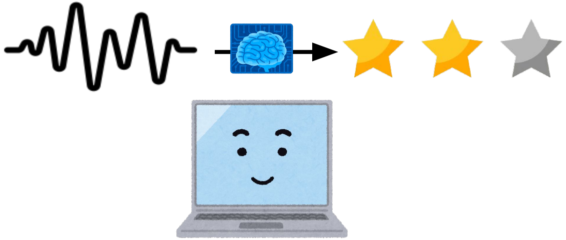
- 7 teams used **per-listener ratings**
- No teams used listener demographics
  - One team used “listener group”
- OOD track: only 3 teams used the **unlabeled data:**

Conducted their own listening test (top team)

Task-adaptive pretraining



“Pseudo-label” the unlabeled data using trained model



Sources of difficulty

Are unseen categories **more difficult?**

Category	**Main track**	**OOD track**
Unseen systems	no	**yes** (6 teams)
Unseen speakers	**yes** (7 teams)	N/A
Unseen listeners	**yes** (17 teams)	no

Sources of difficulty

Systems with **large differences between their training and test set distributions** are harder to predict.

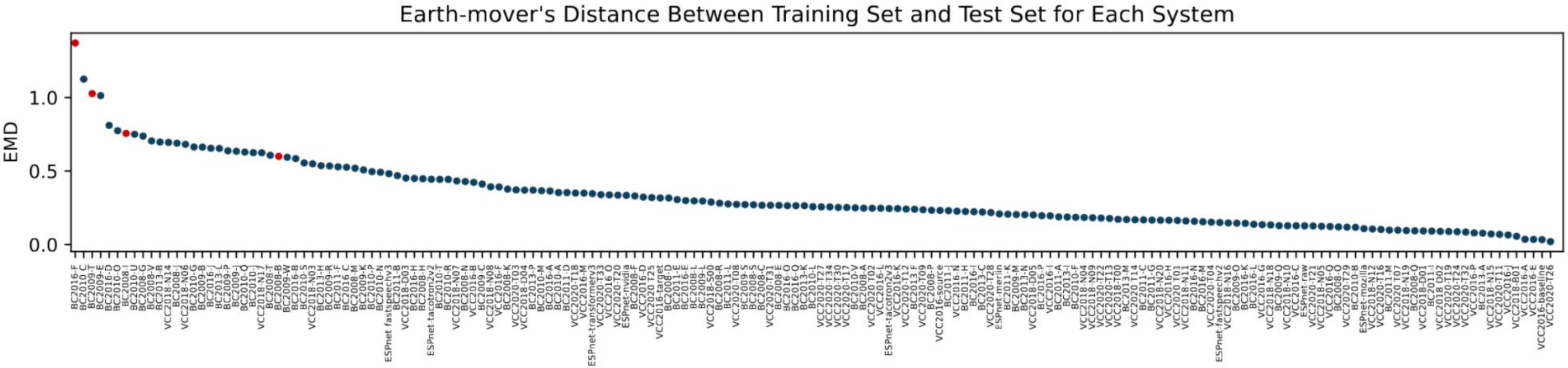


Figure 4: *Difference in distributions between training and test data. The four most difficult systems to predict (red) are in the top EMD range of this figure (left side), indicating that large differences in the distributions of the training and test data contribute to prediction difficulty.*

Sources of difficulty

Low-quality systems are easy to predict.

Middle and high quality systems are harder to predict.

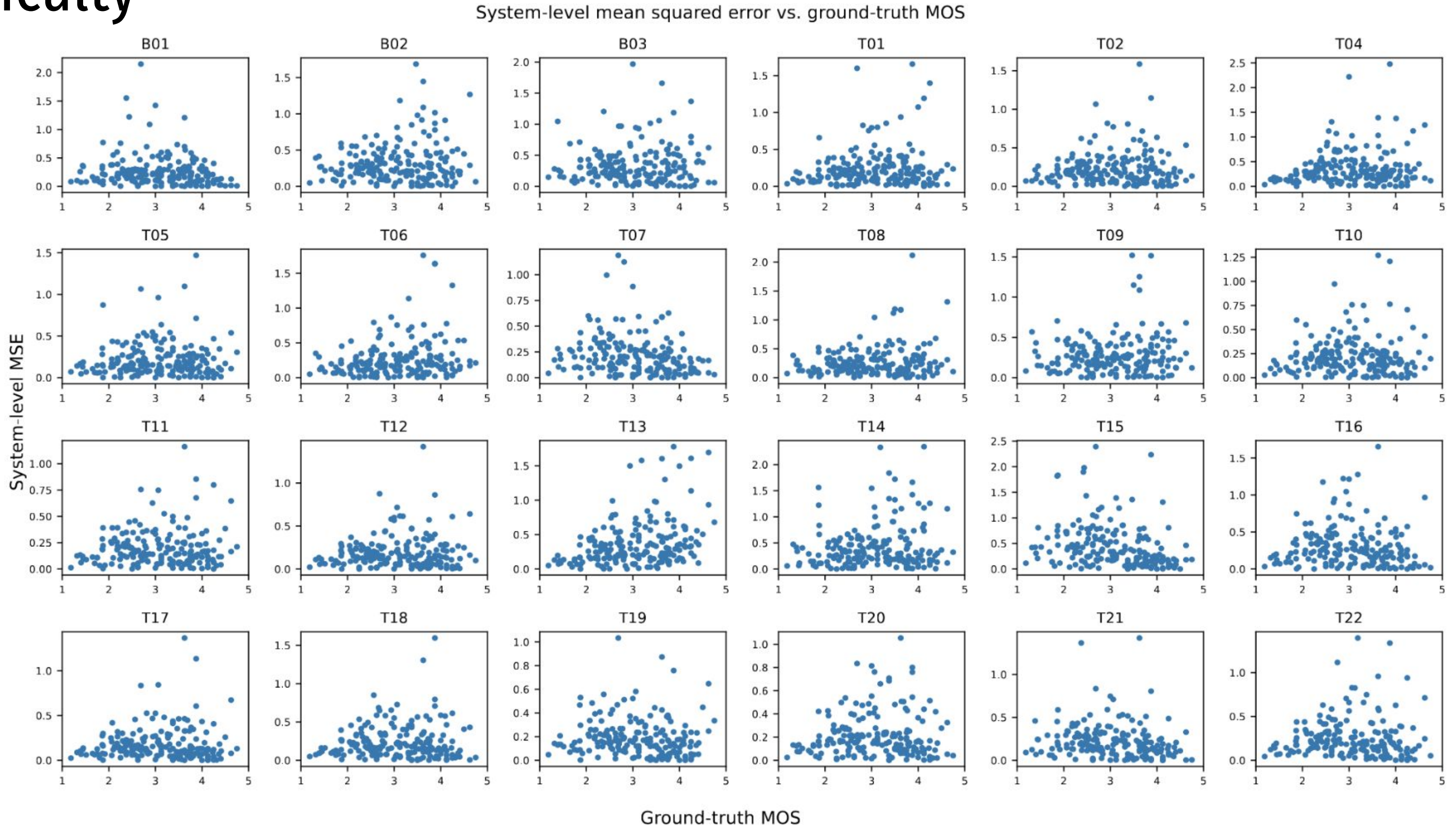


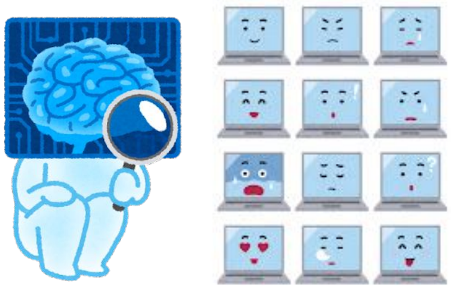
Figure 5: *System-level mean squared error vs. ground-truth system-level MOS. All teams had low errors for low-scoring systems. Higher errors tend to appear for middle- and high-scoring systems.*

Analysis of metrics

Why did you use system-level SRCC as main metric?

Why are there 4 metrics?

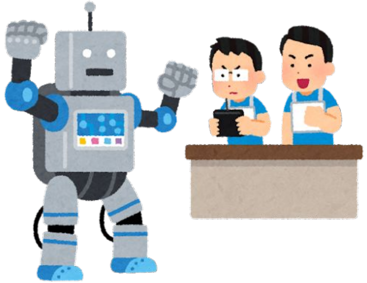
There are two main usage of MOS prediction models:



Compare a lot of systems

Ex. evaluation in scientific challenges.

→ **Ranking**-related metrics are preferred (LCC, SRCC, KTAU)



Evaluate absolute goodness of a system

Ex. use as objective function in training.

→ **Numerical** metrics are preferred (MSE)

Analysis of metrics

We calculated the **linear correlation coefficients** between all metrics using main track results.

	MSE	LCC	SRCC	KTAU
MSE	1.00	-.875	-.862	-.870
LCC	-	1.00	.997	.994
SRCC	-	-	1.00	.994
KTAU	-	-	-	1.00

Correlation coefficients between ranking based metrics are close to 1.

MSE is different from the other three metrics.

⇒ Future researchers can consider just **reporting 2 metrics**: MSE + {LCC, SRCC, KTAU}.

⇒ It is still of significant importance to **develop a general metric** that reflects both aspects.

Conclusions

Conclusions

The goals of the VoiceMOS challenge:

⇒ Attracted more than 20 participant teams.

⇒ SSL is very powerful in this task.

⇒ Generalizing to a different listening test is still very hard.

⇒ **There will be a 2nd, 3rd, 4th,... version!!**

Useful materials

The CodaLab challenge page is still open! Datasets are free to download. VoiceMOS Challenge

The baseline systems are open-sourced!

- https://github.com/nii-yamagishilab/mos-finetune-ssl
- https://github.com/dhimasryan/MOSA-Net-Cross-Domain
- https://github.com/unilight/LDNet

The arXiv paper is available!

- https://arxiv.org/abs/2203.11389