# Automatic Quality Assessment for Speech and Beyond

Wen-Chin Huang
Nagoya University, Japan

**Mila Conversational AI reading group**
**2025.5.15**

名古屋大学
NAGOYA UNIVERSITY

# Who am I?


- Assistant professor @ Toda lab, Nagoya Univ., Japan
  - National Taiwan Univ. (B.E.)
    ⇨ Nagoya Univ., Japan (M.S. & Ph.D.)
- Research interest
  - ~ Ph.D.: voice conversion
  - Now: **speech quality assessment**, voice anonymization
- Specialty: organizing challenges & building toolkits
  - Voice conversion challenge: 2020, 2023, 2025
  - VoiceMOS Challenge: 2022, 2023, 2024, 2025
  - seq2seq-vc, s3prl-vc, sheet, jatts
- HP, Google Scholar, Github

# What is speech quality assessment (SQA)?

- **Assess = evaluate** → speech quality evaluation
- What is quality? → an umbrella term!
  - Noisy/clean? Robotic? Native?
  - Take SQA for synthetic speech as an example:
    - 1980s to early 1990s: intelligibility, comprehension
      Mid-1990s and 2000s: naturalness, intelligibility
      2010s to the present: similarity, hard cases, etc.
  - Nowadays: we ask for more than quality! Similarity, diversity, … etc.
- Properties of SQA:
  - *Subjective*: cognitive difference among different people
  - *Relative*: results differ when the reference sample(s) change

# Why did I shift from “synthesis” to “evaluation”?

- The era of “speech synthesis as fundamental research” is over
  - People are seeking for more than naturalness
  - In voice conversion: emotion conversion, accent conversion…
  - Evaluating these dimensions is hard!
- Evaluation is what makes science different from product development
  - (Technically speaking) the goal of product is to satisfy the market & the customers
  - In science we care about “progress” = ”fair evaluation”
- The ability to “evaluate” is the ability to “**appreciate**”
  - Ex., making AI understand movies, music, art…
  - Related to sociology, psychology, …

# Outline of today's talk

1. Speech quality assessment in the era of DNNs
2. Experiences and lessons from the VoiceMOS Challenge Series
3. Ongoing work and unexplored problems

# Speech quality assessment in the era of DNNs

# You might have seen these metrics in papers...

P. 563

PESQ

POLQA

SSNR

WER

NAT

DNSMOS

MCD

STOI

SI-SDR

MOS

ViSQOL

MUSHRA

ABX

SIM

**Let’s try to classify them!**

# Ways to categorize SQA: SQA for synthetic/non-synthetic speech

**(My definition)**

- Synthetic speech: text-to-speech (TTS), voice conversion, …
- Non-synthetic speech: speech that went through **distortion**
  - Think about *telephony*: noise, reverberation, speech coding, clipping, packet loss, etc.
  - Has a longer history

**Exclude speech enhancement, source separation…**

- Observation: in the literature, SQA for synthetic/non-synthetic speech seems to be different research fields. Why?
  - IMO, SQA for non-synthetic speech is "easier" because it has a **ground truth**
  - Synthetic speech: no ground truth because of the "**one-to-many**" nature
    - Consider TTS: <text, speaker> → speech; there are infinite realizations for a given input
    - Natural fluctuation in human speech production

**(My opinion)**

**Different SQA methods are needed to tackle the difference in nature between synthetic and non-synthetic speech**

# Ways to categorize SQA: subjective/objective

- **Subjective measure:** in the form of **listening tests** (i.e., human studies)
  - Subjective is the most "**accurate**" SQA method
    - The end-user of most speech "generation" tasks is **human**
    - (Exceptions: speech enhancement as front end for ASR)

- Objective measure: any "machine-based" method other than listening tests
  - Subjective tests: too costly in terms of time and money

**Main focus of this talk!**

**IMO: for any objective measure to be valid, its correlation with subjective opinions should be first verified**

# Subjective test types

- Most common type nowadays: **mean opinion score (MOS)**
  - Takes the mean of opinion scores from multiple listeners, usually range from 1-5.
  - Falls into the category of absolute category rating (ACR)
  - Critiques: relative to surrounding samples, equal-ranging bias
  - (Sub-optimal) Solutions: provide references (DMOS; MUSHRA: low-pass filtered)

- What the community tries to promote: **pairwise preference (AB) test**
  - Comparing is less noisy than direct scoring
  - The human auditory system can make comparisons rather than absolute judgments
  - Disadvantage: hard to scale up

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., & Wainwright, M. (2014). When is it better to compare than to score?. arXiv preprint arXiv:1406.6618.

M. Goldstein, “Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener,” Speech Communication, vol. 16, no. 3, pp. 225–244, 1995.

# Ways to categorize SQA: intrusive/non-intrusive

- Intrusive = reference-based = double-ended
  Non-intrusive = reference-free = single-ended
- SQA for **non-synthetic speech** usually adopts **intrusive** methods
  - Because there is a clear ground-truth (as mentioned before)
    - Examples: short-time objective intelligibility (**STOI**), Perceptual Evaluation of Speech Quality (**PESQ**), scale invariant signal-to-distortion ratio (**SI-SDR**)
- It is harder to adopt intrusive methods in SQA for synthetic speech
  - However, intrusive methods are sometime used: Mel cesptrum distortion (MCD), speaker similarity tests, ABX preference tests
- Developing non-intrusive methods has been a trend in the past decade.

# Ways to categorize SQA: signal-/model-based

**(My definition)**

- Model-based: **learns from data** to make the prediction
  - Advantage: correlates better with human judgements
  - Disadvantage: generalization issues

**⇨ Gaining attention since late 2010s thanks to DNNs!**

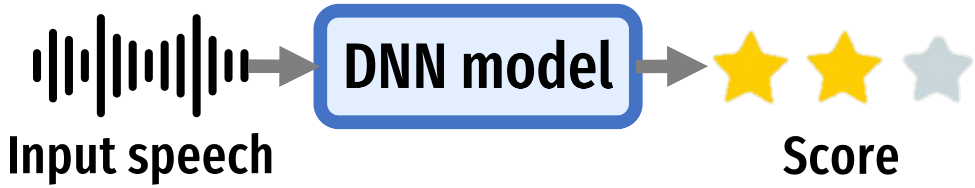
- Signal-based: does not require learning such a model
  - Calculates some pre-defined **distance** between input and reference
  - Advantage: suffers less from generalization
  - Disadvantage: mostly intrusive

# Let's categorize some non DNN-based objective SQA metrics...

| Metric | Evaluation target? | Intrusive? Non-intrusive | Signal-/model-based | What does it measure? |
|---|---|---|---|---|
| **PESQ** | Non-synthetic speech | Intrusive | Model-based! | Perceptual quality |
| **STOI** | Non-synthetic speech | Intrusive | Signal-based | intelligibility |
| **SSNR** | Non-synthetic speech (for speech enhancement) | Intrusive | Signal-based | Signal distortion |
| **SI-SDR** | Non-synthetic speech (for source separation) | Intrusive | Signal-based | Signal distortion |
| **POLQA** | Non-synthetic speech (for telephony) | Intrusive | Signal-based | Perceptual quality |
| **ViSQOL** | Non-synthetic speech (for VoIP, codecs) | Intrusive | Signal-based | Perceptual quality |
| **P.563** | Synthetic speech | Non-intrusive | Signal-based | Perceptual quality |
| **MCD** | Synthetic speech | Intrusive | Signal-based | Spectral distortion |

# DNN-based SQA: basic idea & learning target

- Early attempts: intrusive methods; non-intrusive has soon become mainstream




- **Way to categorize DNN-based SQA: learning target**
  1. Some other objective metric: PESQ, STOI, … etc.
     - Motivation: use a non-intrusive network to mimic intrusive metrics
     - Advantage: data is infinite (can be artificially generated)
  2. Human judgement scores ⇨ **subjective speech quality assessment (SSQA)**
     - Collected through listening tests
     - Problem: such dataset is always scarce…

**I am personally more interested in this direction**

# **Subjective SQA datasets** (all with MOS labels)

| Name | Speech type | Language | FS (kHz) | # samples (train/dev) |
|---|---|---|---|---|
| BVCC | **TTS, VC**, natural speech | English | 16 | 4944/1066 |
| SOMOS | **TTS**, natural speech | English | 24 | 14100/3000 |
| SingMOS | **SVS, SVC**, natural singing voice | Mandarin, Japanese | 16 | 2000/544 |
| NISQA | **artificial distorted speech, real distorted speech**, clean speech | English | 48 | 11020/2700 |
| TMHINT-QI | **artificial noisy speech, enhanced speech**, clean speech | Mandarin | 16 | 11644/1293 |
| Tencent | **artificial distorted speech**, clean speech | Mandarin | 16 | 10408/1155 |
| PSTN | **PSTN speech, artificial distorted speech** | English | 8 | 52839/5870 |

W.-C. Huang, E. Cooper, and T. Toda, "MOS-Bench: Benchmarking generalization abilities of subjective speech quality assessment models," arXiv preprint arXiv:2411.03715, 2024

# Evaluation of (DNN-based) SQA methods

- Four commonly used metrics to evaluate SQA
  - **MSE (mean squared error):** Sensitive to large errors; penalizes outliers
  - **LCC (linear correlation coefficient):** measures linear correlation
  - **SRCC (Spearman rank correlation coefficient):** focuses on ordinal ranking
  - **KTAU (Kendall's Tau correlation coefficient):** more robust than SRCC for small datasets

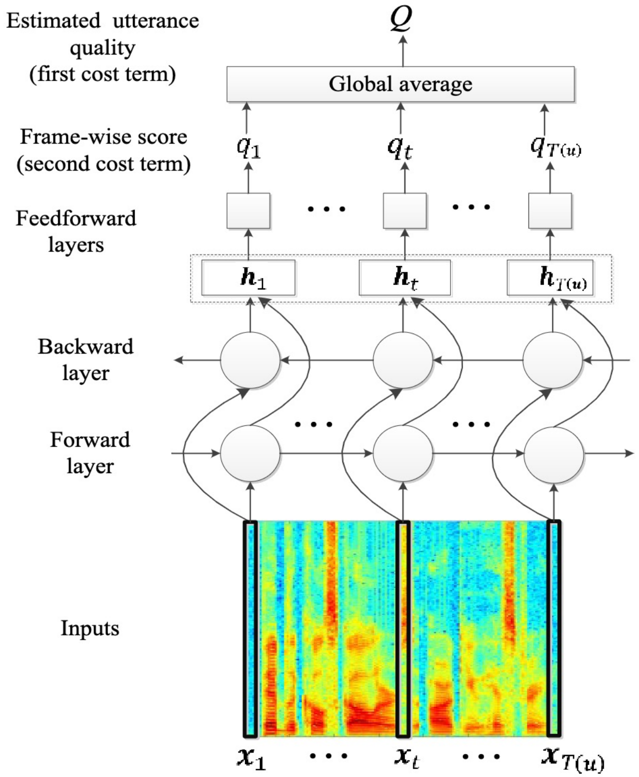
- There are two main usage of SQA methods
  - **Compare a lot of systems** (ex., evaluation in scientific challenges)
    - **Ranking-related** metrics are preferred (LCC, SRCC, KTAU)
  - **Evaluate absolute goodness of a system** (ex., objective function in training)
    - **Numerical** metrics are preferred (MSE)

# SQA for non-synthetic speech: Quality-Net

- Non-intrusive
- Learning target: PESQ
- Evaluation target: noise suppressors
- Model architecture: BLSTM
- Training data: noisy speech
- Contributions: **pioneer work on DNN-based SQA**

Table 3: *Results of Quality-Net and the two-stage model.*

| | MSE | LCC | SRCC |
|---|---|---|---|
| Autoencoder +NN [22] | 0.1529 | 0.8434 | 0.8675 |
| Quality-Net | **0.1266** | **0.8749** | **0.8807** |




S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in Proc. Interspeech, 2018.
Google scholar citations: 209

# SQA for non-synthetic speech: DNSMOS (Deep Noise Suppression MOS?)

- Non-intrusive
- Learning target: human judgement
- Evaluation target: noise suppressors
- Model architecture: CNN
- Training data: noise-suppressed speech
- Contributions: **trained on crowdsourced human preference data; easy-to-use API**

| Layer | Output dimension |
|---|---|
| Input | 900 x 120 x 1 |
| Conv: 32, (3 x 3), 'ReLU' | 900 x 120 x 32 |
| MaxPool: (2 x 2), Dropout(0.3) | 450 x 60 x 32 |
| Conv: 32, (3 x 3), 'ReLU' | 450 x 60 x 32 |
| MaxPool: (2 x 2), Dropout(0.3) | 225 x 30 x 32 |
| Conv: 32, (3 x 3), 'ReLU' | 225 x 30 x 32 |
| MaxPool: (2 x 2), Dropout(0.3) | 112 x 15 x 32 |
| Conv: 64, (3 x 3), 'ReLU' | 112 x 15 x 64 |
| GlobalMaxPool | 1 x 64 |
| Dense: 64, 'ReLU' | 1 x 64 |
| Dense: 64, 'ReLU' | 1 x 64 |
| Dense: 1 | 1 x 1 |

**Table 2**: Correlation of DNSMOS with other widely used objective metrics

| | PESQ | SDR | POLQA | DNSMOS ($M_0$) |
|---|---|---|---|---|
| PCC | 0.78 | 0.23 | 0.79 | 0.93 |
| SRCC | 0.82 | 0.25 | 0.84 | 0.94 |

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in Proc. ICASSP, 2021.
Google scholar citations: 344

# SQA for non-synthetic speech: NISQA (Non-Intrusive Speech Quality Assessment)

- Non-intrusive
- Learning target: human judgement
- Evaluation target: noise suppressors
- Model architecture: CNN
- Training data: 59 distorted speech datasets
- Contributions: **released a large-scale human preference data; released pre-trained model weights and code**

| Dataset | Scale | Lang | Con | Files | NISQA | |
|---|---|---|---|---|---|---|
| | | | | | $r$ | RMSE |
| 103_ERICSSON | SWB | se | 54 | 648 | 0.85 | 0.38 |
| 104_ERICSSON | NB | se | 55 | 660 | 0.77 | 0.47 |
| 203_FT_DT | SWB | fr | 54 | 216 | **0.92** | **0.36** |
| 303_OPTICOM | SWB | en | 54 | 216 | 0.92 | 0.33 |
| 403_PSYTECHNICS | SWB | en | 48 | 1152 | 0.91 | 0.36 |
| 404_PSYTECHNICS | NB | en | 48 | 1151 | 0.77 | 0.39 |
| 503_SWISSQUAL | SWB | de | 54 | 216 | 0.92 | 0.34 |
| 504_SWISSQUAL | NB | de | 49 | 196 | **0.92** | **0.37** |
| 603_TNO | SWB | nl | 48 | 192 | 0.89 | 0.44 |
| ERIC_FIELD_GSM_US | NB | en | 372 | 372 | **0.79** | **0.36** |
| HUAWEI_2 | NB | zh | 24 | 576 | **0.98** | **0.21** |
| ITU_SUPPL23_EXP1o | NB | en | 44 | 176 | 0.92 | 0.31 |
| ITU_SUPPL23_EXP3d | NB | ja | 50 | 200 | 0.92 | 0.27 |
| ITU_SUPPL23_EXP3o | NB | en | 50 | 200 | 0.91 | 0.30 |
| TUB_AUS | FB | en | 50 | 600 | **0.91** | **0.21** |
| TUB_LIKE | SWB | de | 8 | 96 | 0.98 | 0.25 |
| NISQA_VAL_LIVE | FB | en | 200 | 200 | **0.82** | **0.40** |
| NISQA_VAL_SIM | FB | en | 2500 | 2500 | **0.90** | **0.48** |
| NISQA_TEST_P501 | FB | en | 60 | 240 | **0.95** | 0.31 |
| NISQA_TEST_NSC | FB | de | 60 | 240 | **0.97** | **0.23** |
| NISQA_TEST_FOR | FB | en | 60 | 240 | **0.95** | **0.26** |
| NISQA_TEST_LIVETALK | FB | de | 58 | 232 | **0.90** | **0.35** |

**Good performance across many telephony datasets**

G. Mittag, B. Naderi, A. Chehadi, and S. M¨oller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in Interspeech, 2021, Google scholar citations: 286

# SQA for non-synthetic speech: TorchAudio-Squim

(TorchAudio-Speech QUality and Intelligibility Measures)

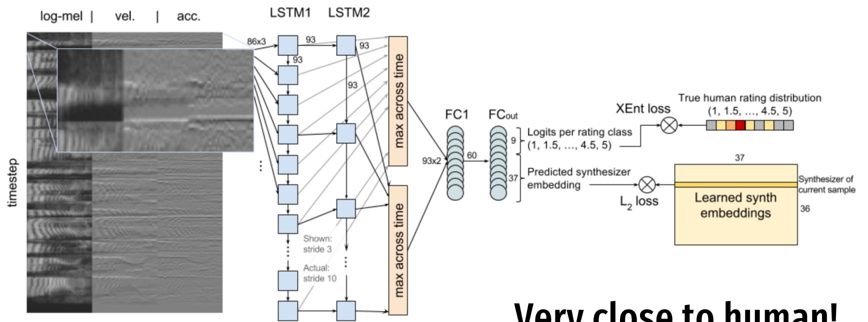
- Non-intrusive & intrusive with human judgement & STOI, PESQ, SI-SDR
- Evaluation target: noise suppressors
- Model architecture: DPRNN & Transformers
- Training data: DNS Challenge 2020
- Contributions: **relatively new (2023); tight integration with TorchAudio**

| Approach | STOI (%) | | | WB-PESQ | | | SI-SDR (dB) | | | # Params | # MAC/5s |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE ↓ | PCC ↑ | SRCC ↑ | MAE ↓ | PCC ↑ | SRCC ↑ | MAE ↓ | PCC ↑ | SRCC ↑ | | |
| Quality-Net [28] | - | - | | 0.396 | 0.845 | 0.849 | - | - | - | 0.30 M | 297.30 K |
| MOSA-Net [17] | 5.254 | 0.900 | 0.864 | 0.335 | 0.904 | 0.914 | 1.990 | 0.965 | 0.958 | 317.19 M | 94.86 G |
| AMSA [13] | 3.498 | 0.913 | 0.826 | 0.207 | 0.932 | 0.938 | 1.562 | 0.968 | 0.964 | 2.96 M | 687.61 M |
| MetricNet [16] | - | - | - | 0.182 | 0.938 | 0.947 | - | - | - | 6.61 M | 2.08 G |
| Ours without MTL | 2.324 | 0.939 | 0.935 | 0.168 | 0.942 | 0.951 | 1.158 | 0.977 | 0.973 | 7.39 M | 40.27 G |
| Ours with MTL | **1.994** | **0.950** | **0.950** | **0.142** | **0.958** | **0.963** | **0.838** | **0.985** | **0.985** | 7.39 M | 40.27 G |

A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in Proc. ICASSP, 2023
Google scholar citations: 52

# SQA for synthetic speech: AutoMOS

- Non-intrusive
- Learning target: human judgement
- Evaluation target: TTS
- Model architecture: LSTM
- Training data:
  36 TTS systems, 168086 scores
- Contributions:
  **very first DNN-based work for synthetic speech**

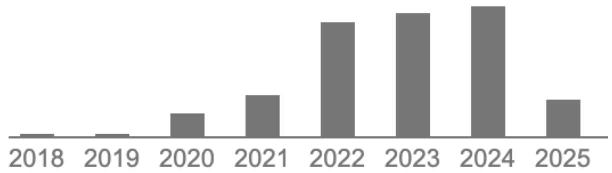



**Very close to human!**

| | Baseline | **AutoMOS** | GT |
|---|---|---|---|
| Utt-RMSE | 0.553 | **0.462** | 0.512 |
| Utt-LCC | 0.454 | **0.668** | 0.764 |
| Utt-SRCC | 0.399 | **0.667** | 0.757 |
| Sys-RMSE | 0.132 | **0.073** | 0.034 |
| Sys-LCC | 0.795 | **0.938** | 0.987 |
| Sys-SRCC | 0.679 | **0.949** | 0.986 |

B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," in NIPS 2016 Workshop.
Google scholar citations: 109

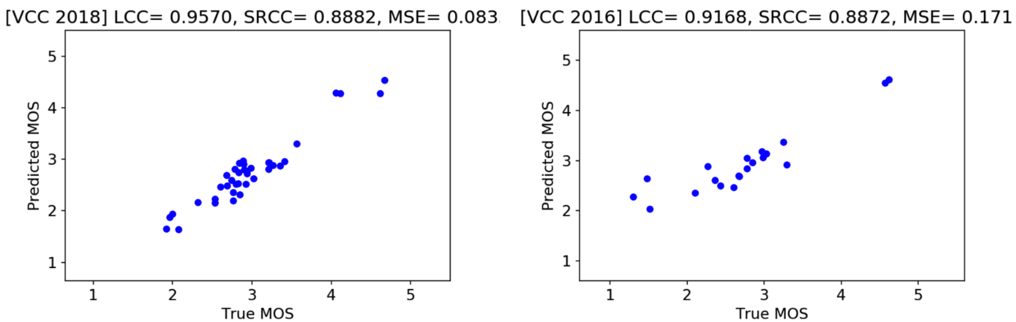
# SQA for synthetic speech: MOSNet




- Non-intrusive
- Learning target: human judgement
- Evaluation target: VC
- Model architecture: CNN & LSTM
- Training data:
  Voice Conversion Challenge 2018
- Contributions:
  **one of the first works with pre-trained model & easy-to-beat performance**

| | *utterance-level* | | | *system-level* | | |
|---|---|---|---|---|---|---|
| **Model$_{batchsize}$** | **LCC** | **SRCC** | **MSE** | **LCC** | **SRCC** | **MSE** |
| BLSTM$_1$ [7] | 0.511 | 0.484 | 0.604 | 0.826 | 0.808 | 0.165 |
| BLSTM$_{16}$ | 0.487 | 0.453 | 0.658 | 0.818 | 0.797 | 0.190 |
| BLSTM$_{64}$ | 0.251 | 0.254 | 0.803 | 0.412 | 0.427 | 0.404 |
| CNN$_1$ | 0.638 | 0.587 | **0.486** | 0.945 | 0.875 | 0.058 |
| CNN$_{16}$ | 0.620 | 0.573 | 0.512 | 0.944 | 0.890 | 0.067 |
| CNN$_{64}$ | 0.624 | 0.585 | 0.522 | 0.946 | 0.872 | 0.057 |
| CNN-BLSTM$_1$ | 0.584 | 0.551 | 0.634 | 0.951 | 0.873 | 0.135 |
| CNN-BLSTM$_{16}$ | 0.607 | 0.569 | 0.540 | 0.944 | **0.897** | **0.055** |
| CNN-BLSTM$_{64}$ | **0.642** | **0.589** | 0.538 | **0.957** | 0.888 | 0.084 |

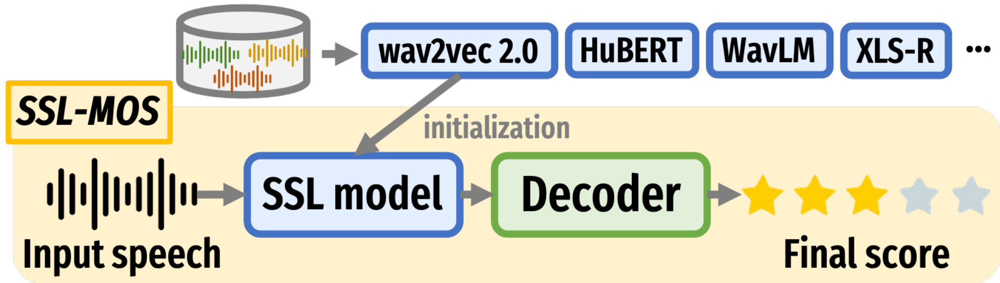
**Decent correlation!**




C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Ya- magishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Con- version," in Proc. Interspeech 2019, 2019, pp. 1541– 1545.
Google scholar citations: 352

# SQA for synthetic speech: SSL-MOS

- Non-intrusive
- Learning target: human judgement
- Evaluation target: BVCC
- Model architecture: SSL (wav2vec 2.0)
- Training data: BVCC
- Contributions:
  **one of the first SSL-based SQA works with pre-trained model**

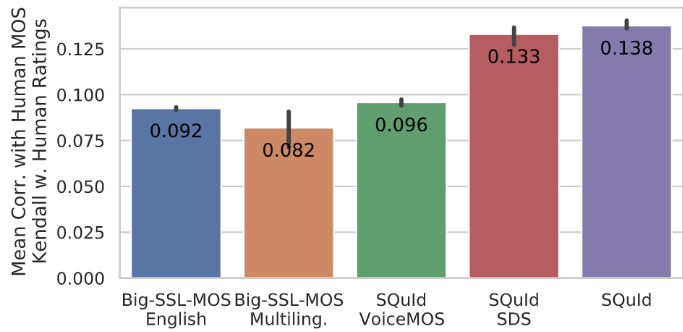



| | Test set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Utterance level | | | | System level | | | |
| Base model | MSE | LCC | SRCC | KTAU | MSE | LCC | SRCC | KTAU |
| w2v_small | 0.227 | **0.868** | **0.866** | **0.690** | 0.121 | 0.938 | 0.942 | 0.790 |
| libri960_big | 0.342 | 0.823 | 0.820 | 0.635 | 0.136 | 0.901 | 0.901 | 0.730 |
| w2v_vox_new | 0.342 | 0.767 | 0.753 | 0.570 | 0.112 | 0.903 | 0.900 | 0.721 |
| w2v_large | **0.220** | **0.868** | 0.865 | **0.690** | **0.059** | **0.948** | **0.944** | **0.803** |
| xlsr_53_56k | 0.281 | 0.821 | 0.816 | 0.633 | 0.107 | 0.902 | 0.894 | 0.730 |
| hubert_base_ls960 | 0.318 | 0.842 | 0.837 | 0.655 | 0.213 | 0.919 | 0.915 | 0.745 |
| hubert_large_ll60k | 0.444 | 0.696 | 0.687 | 0.507 | 0.184 | 0.812 | 0.805 | 0.620 |

E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in Proc. ICASSP, 2022
Google scholar citations: 175

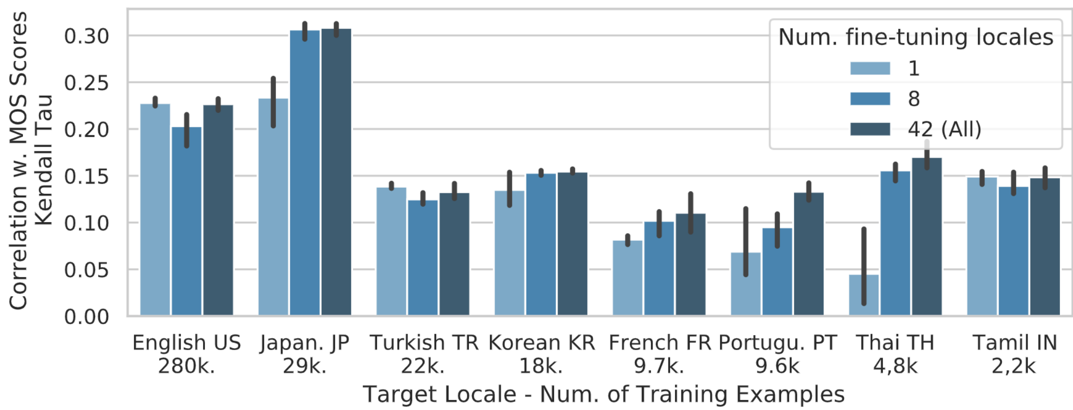
# SQA for synthetic speech: SQuID (Speech Quality IDentification)

- Non-intrusive
- Learning target: human judgement
- Evaluation target: internal dataset
- Model architecture: SSL (mSLAM)
- Training data: internal TTS samples (~1M samples, 1476 systems)
- Contributions: **first massive multi-lingual subjective SQA work**




⇦ Results on SQuID dataset: Using SQuID boosts!

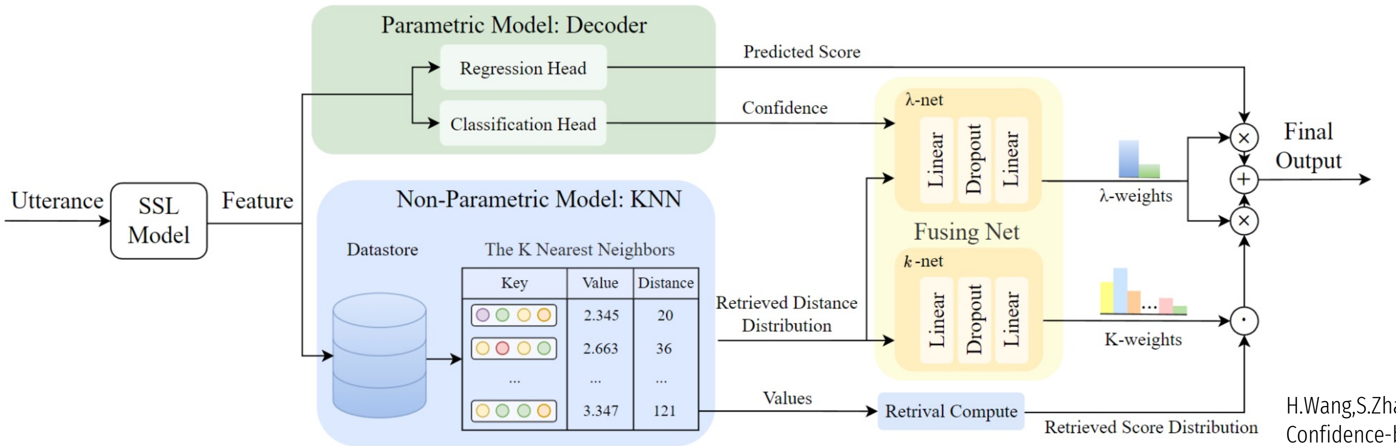
⇩Broken down by locale




T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh, and J. Riesa, "SQuId: Measuring speech naturalness in many languages," in Proc. ICASSP, 2023
Google scholar citations: 26

# SQA for synthetic speech: RAMP (Retrieval Augmented MOS Prediction)

- Non-intrusive. Learning target: human judgement
- Training data: BVCC; Evaluation target: BVCC, SOMOS
- Model architecture: SSL + retrieval
- Contributions: **top-performing system in VoiceMOS Challenge 2023, 2024**




H.Wang,S.Zhao,X.Zheng,andY.Qin,"RAMP:Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting," in Proc. Interspeech, 2023, pp. 1095–1099.

# Experiences and lessons from the VoiceMOS Challenge Series

# The goal of the VoiceMOS challenge (VMC) series
## (or any scientific challenge)

VoiceMOS Challenge

https://sites.google.com/view/voicemos-challenge


**Advertise** the research of automatic data-driven MOS prediction for speech

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

Promote **discussion** about the future of this research field

# The whole VMC series is about generalization

- **In-domain (ID)** & **out-of-domain (OOD)** generalization: test & train data are of the same/different distribution
- In practical situations for SQA, **we should always assume it's OOD**
  - Synthetic speech: different TTS system, different listening test, …
  - Non-synthetic speech: different distortion types, levels, combinations, …
- Ultimate goal: **an "almighty" system** that excels in all speech types

# The history of VMC

- The VoiceMOS Challenge 2022 @ INTERSPEECH
  - **In-domain** prediction for synthetic speech (TTS, VC)
  - Results: best system achieved **0.939 SRCC**
- The VoiceMOS Challenge 2023 @ ASRU
  - Fully **out-of-domain** setting on singing voice conversion, French TTS, noisy speech
  - Results: reconfirmed that **OOD generalization is an issue**
- The VoiceMOS Challenge 2024 @ SLT
  - Zoomed-in tests, singing conversion/synthesis, semi-supervised SQA
- The **AudioMOS** Challenge 2025 @ ASRU **ongoing!**
  - Expand to **general audio**: text-to-speech/audio/music; different speech frequencies

# VMC 2022: tracks

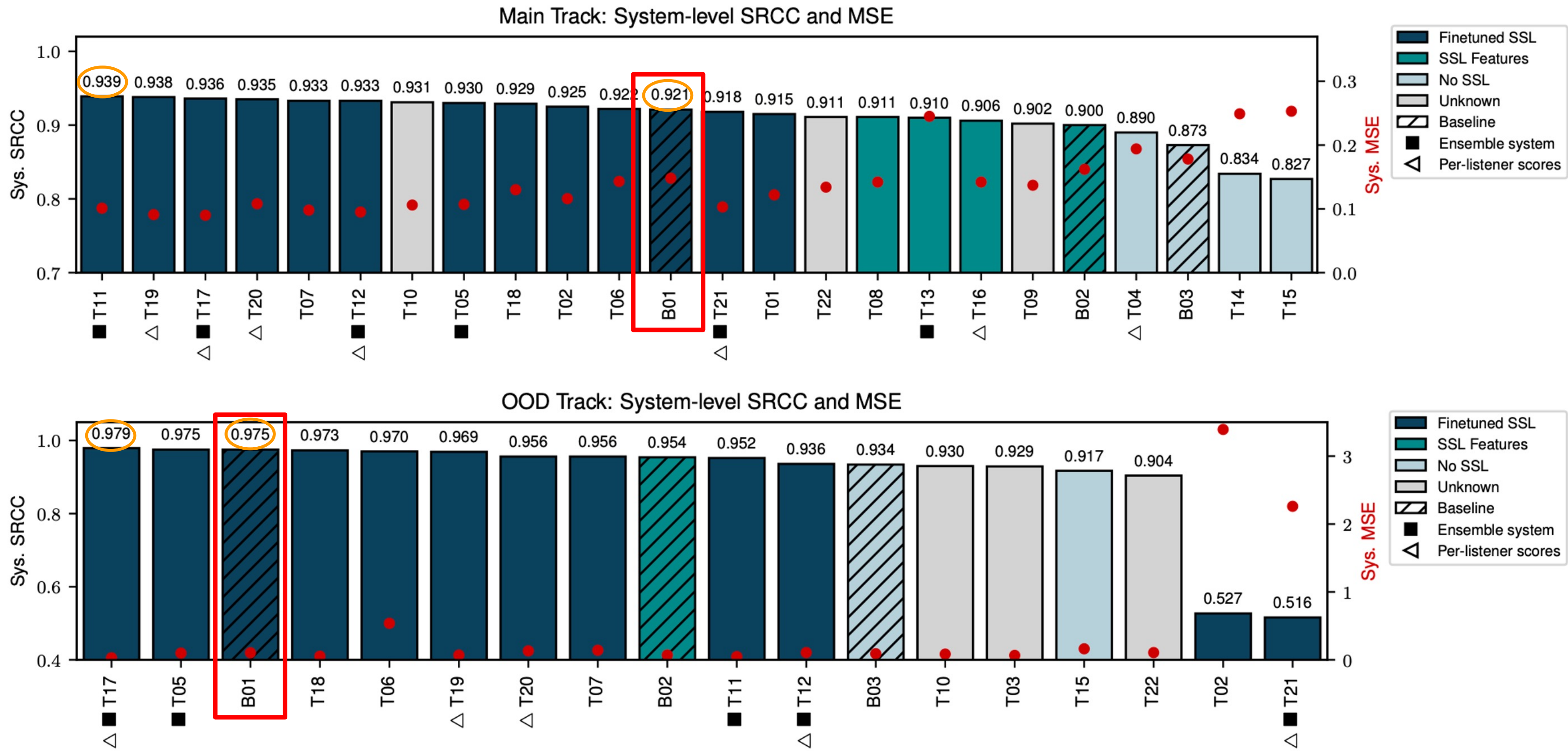
| Track | Lang | # Samples | | | # ratings per sample |
|---|---|---|---|---|---|
| | | Train | Dev | Test | |
| Main | Eng | 4,974 | 1,066 | 1,066 | 8 |
| OOD | Chi | Label: 136<br>Unlabel: 540 | 136 | 540 | 10-17 |

- **Main track: BVCC**
  - Samples from **187** different systems all rated together in one listening test
    - Past Blizzard Challenges (for TTS) 2008 - 2018
    - Past Voice Conversion Challenges (for voice conversion) 2016 - 2020
    - ESPnet-TTS (implementations of modern TTS systems), 2020
  - Test set is split from the training set ⇨ **in-domain**
    - Contains some unseen systems/listeners/speakers
- **OOD track: Blizzard Challenge 2019**
  - Chinese TTS samples from systems submitted to the 2019 Blizzard Challenge
  - Test set is split from the training set ⇨ **in-domain**
    - Contains unseen systems/listeners

**Probably a bad naming…**
**"limited-data" track might be better ☹**

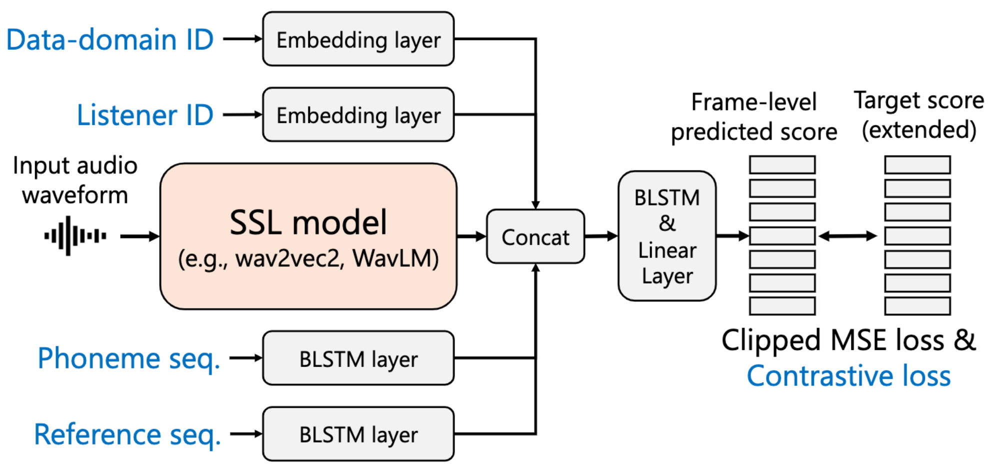
# VMC 2022: results

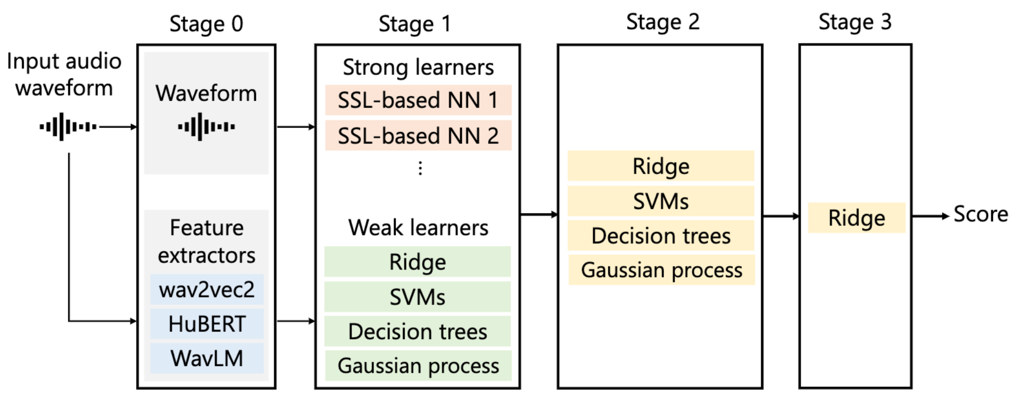
☺ Improvements over baseline
☺ Good performance even with 136 samples only
⇨ in-domain is probably "too simple"?



# VMC 2022 top system: UTMOS

- Main track system: "slightly improved SSL-MOS" (according to 1st author)
- OOD track: ensemble of weak learners using stacking




T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in Proc. Interspeech, 2022, pp. 4521–4525.
Google scholar citations: 229

# VMC 2022: feedback

- About the dataset
  - Test set is too small
  - Is the number of samples per system enough? (T06)

- What do you want to see in the next challenge?
  - Other **speech types**
    - Telephone, conference, speech coding (low bitrate, neural coding), noisy speech (most requested)
    - Music, dialogue TTS, high-quality TTS, speaker similarity, confidence
  - More **languages** (4 participants)
  - Other **listening test types** (A/B preference tests, MUSHRA tests, or simply predict the ranking)
  - Higher **sampling rate** (16000 Hz is too low, at least 22050/24000)

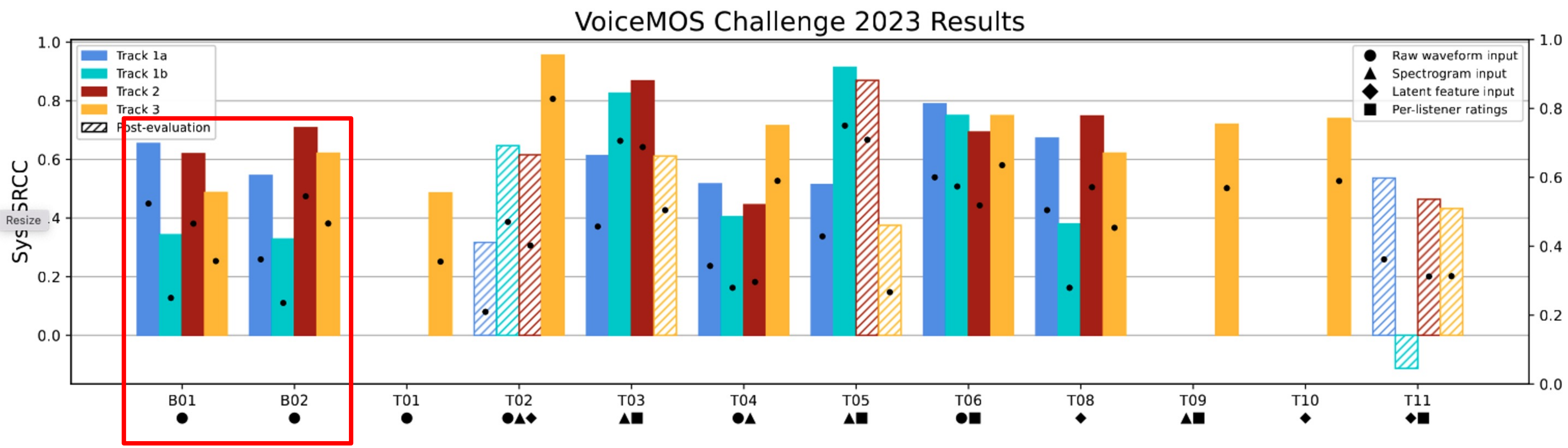
# VMC 2023: tracks

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

- Track 1: Blizzard Challenge 2023 - French TTS
- Track 2: Singing Voice Conversion Challenge - singing voice conversion
- Track 3: Mandarin noisy & enhanced speech

- **Real-world** and challenging MOS prediction in collaboration with ongoing synthesis competitions.
  - Teams submit their predictions before the actual listening test results have been collected.
  - Thus, **no official training data**!

# VMC 2023: results

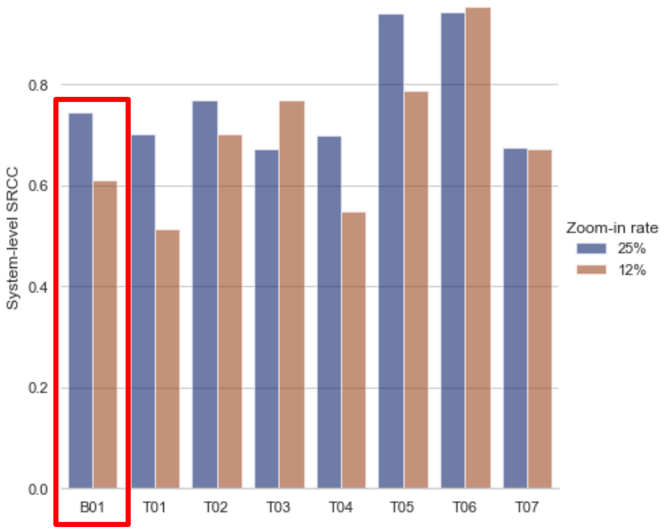


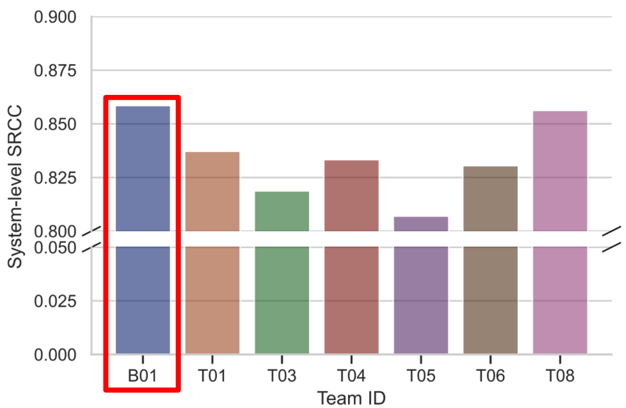

☺ Some systems beat the baselines
☹ Difficult to predict all domains well with a single system

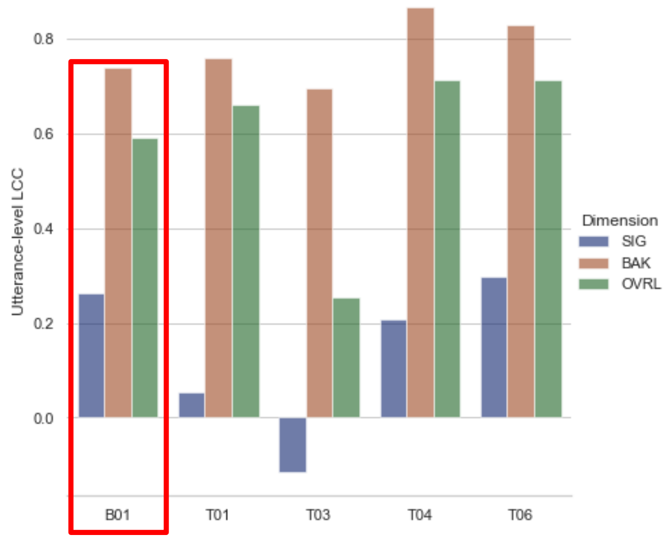
# VMC 2024: tracks

- **Track 1: MOS prediction for "zoomed-in" systems**
  - Motivation: evaluate synthetic systems of high-quality
- **Track 2: MOS prediction for singing voice**
  - Using the SingMOS dataset: natural singing voices, vocoder analysis-synthesis, singing voice synthesis/conversion samples
- **Track 3: semi-supervised MOS prediction for clean/noisy/enhanced speech**
  - Setting: very limited amount of training data & zero-shot setting
  - Beyond quality: speech signal quality (SIG), background intrusiveness (BAK), overall quality (OVRL)

# VMC 2024: results




☺ Some systems beat the baselines

☹ We had less participants this year, thus less insights...

# AMC 2025: tracks

https://sites.google.com/view/voicemos-challenge/audiomos-challenge-2025

- **Track 1: MOS prediction for text-to-music systems**
  - Based on the MusicEval dataset: clips from 31 TTM systems
  - Ratings from music experts
  - Two evaluation axes: overall musical impression, textual alignment
- **Track 2: Audiobox-aesthetics-style prediction for text-to-speech, text-to-audio and text-to-music systems**
  - Based on the Meta Audiobox Aesthetics
  - Train data: natural speec/audio/music samples; test data: TTS/TTA/TTM samples
- **Track 3: MOS prediction for speech in high sampling frequencies**
  - Speech samples from 16/24/48 kHz

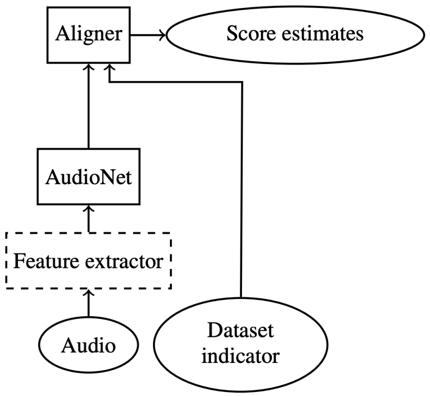
**Stay tuned for the challenge summary!**

# Ongoing work and unexplored problems

# Towards zero-shot, general purpose SQA

- Common idea: how about we **combine** multiple datasets (and their scores)?
- Problem: **"corpus effect"**
  - Same type of speech can receive different scores on different listening tests
  - Stems from the "relative" nature of listening tests like MOS

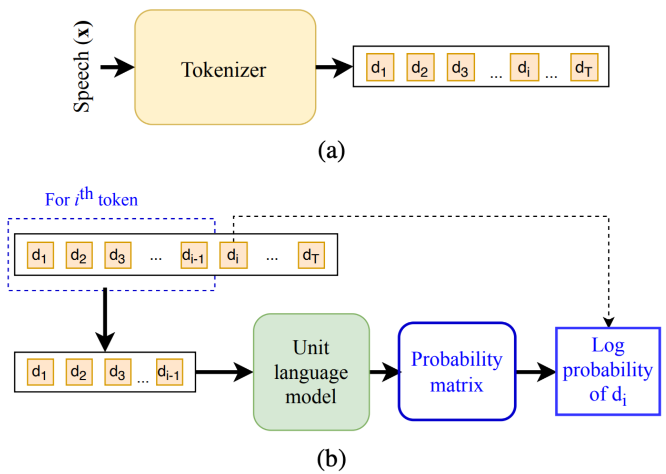
- Recent representative work: **AlignNet**
  - Use a dataset embedding (indicator) to learn the bias in each dataset




J. Pieper and S. Voran, "Alignnet: Learning dataset score alignment functions to enable better training of speech quality estimators," in Proc. Interspeech, 2024, pp. 82–86.

# Alternative solution 1: unsupervised SQA

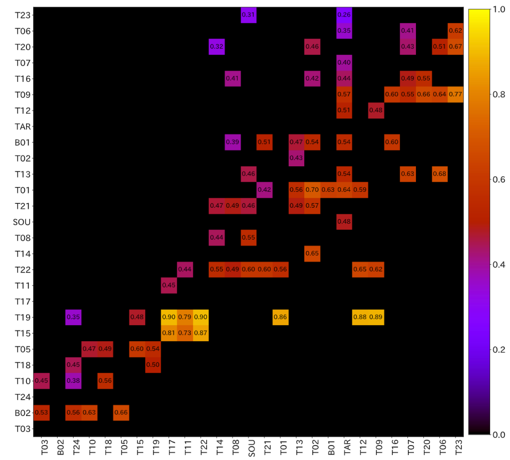
- SQA models are usually supervised: need to be trained with <speech, score>
  - New speech type ⇨ human label needed. Costly!
- Popular idea: **learn a prior model** with the concept of “natural speech”
- Representative work: SpeechLMScore
  - Perplexity of an input speech in the discrete speech token space
- What’s the advantage?
  - **No training = no overfitting = better generalization!**



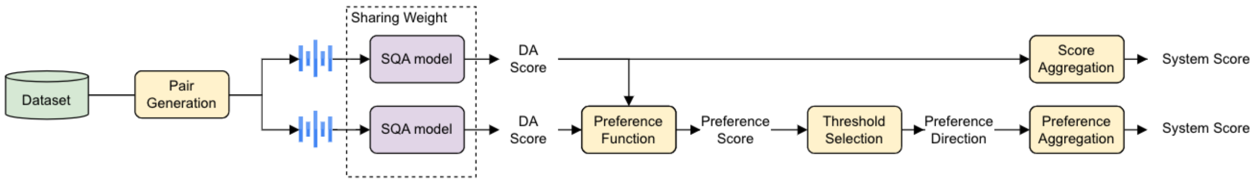

S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, “SpeechLMScore: Evaluating speech generation us- ing speech language model,” in Proc. ICASSP, 2023

# Alternative solution 2: learn from preference data

- Preference test can be speeded up with online learning
  - Automatically stops comparing systems that are obviously different in quality




- Learning from preference data alleviates biases in MOS
  - Listener preference bias, equal-ranging bias
  - Result: **better generalization ability** (both in-domain and OOD!)

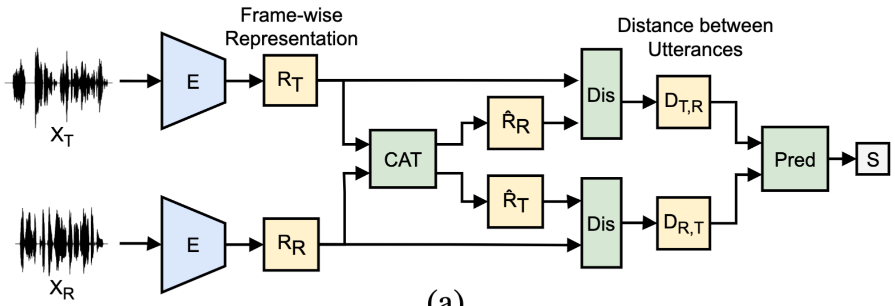



Y. Yasuda, and T. Toda. "Automatic design optimization of preference-based subjective evaluation with online learning in crowdsourcing environment," arXiv preprint arXiv:2403.06100 (2024).

C.-H. Hu, Y. Yasuda, and T. Toda. "E2EPref: An end-to-end preference-based framework for speech quality assessment to alleviate bias in direct assessment scores," Computer Speech & Language, vol. 93, 2025

# Evaluation dimensions beyond general quality

- Many attempts to learn from **subjective speaker similarity** data
- Dataset: VoxSim
  - Derived from VoxCeleb; 41k utterance pairs, nearly 70k ratings
- Model: SVSNet




(a)

**☹ Current results are not significantly better than simple cosine similarity of speaker embeddings (ex., x-vectors)**

**What about other dimensions**
**⇨ Emotion, expressiveness, accent, non-verbal content…**

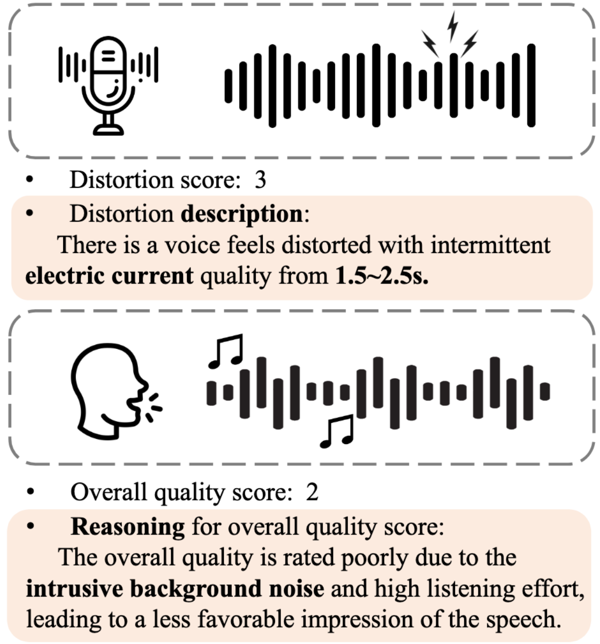

C.-H. Hu, Y.-H. Peng, J. Yamagishi, Y. Tsao, and H.- M. Wang, "SVSNet: An End-to-End Speaker Voice Similarity Assessment Model," IEEE Signal Processing Letters, vol. 29, pp. 767–771, 2022.

J. Ahn, Y. Kim, Y. Choi, D. Kwak, J.-H. Kim, S. Mun, and J. S. Chung, "VoxSim: A perceptual voice similarity dataset," in Proc. Interspeech, 2024.

# Interpretable/explainable SQA

**IMO: the ultimate goal in SQA**

- A recent trend: use LLMs for SQA
  - "Audio captioning" but focusing on quality
  - More than just "another LLM application"!
- Provide "**explanations"** beyond just "scores"
  - Localized evaluation (when & where)
  - Attributed evaluation (what & how)
  - ⇨ No extinction between synthetic/non-synthetic speech!
- Evaluation is the problem
  - Natural language description
    = larger variance compared to scores





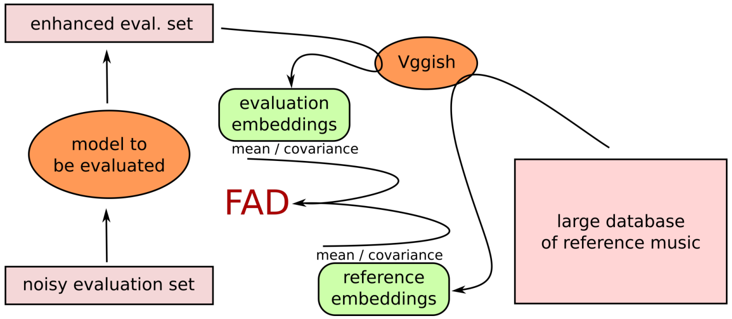
Wang, S., Yu, W., Chen, X., Tian, X., Zhang, J., Tsao, Y., ... & Zhang, C, "QualiSpeech: A Speech Quality Assessment Dataset with Natural Language Reasoning and Descriptions." arXiv preprint arXiv:2503.20290.

S. Wang, W. Yu, Y. Yang, C. Tang, Y. Li, J. Zhuang, X. Chen, X. Tian, J. Zhang, G. Sun, et al, "Enabling auditory large language models for automatic speech quality evaluation," in Proc. ICASSP, 2025

C. Chen, Y. Hu, S. Wang, H. Wang, Z. Chen, C. Zhang, C.-H. Huck Yang, and E. S. Chng, "Audio large language models can be descriptive speech quality evaluators," in Proc. ICLR, 2025

# Status quo in text-to-audio/text-to-music evaluation is mostly objective

- **Fréchet audio distance (FAD)**: evaluates general audio fidelity
  - Set-wise comparison (not sample-wise); calculates statistics in an embedding space
  - Critiques: embedding-sensitive; sample size-sensitive; correlates poorly with perception
  - Improved attempts: KAD, MMD
- **CLAP score**: evaluates alignment between audio and text prompt
  - Cosine similarity between text embedding and audio embedding
  - Critique: **correlates poorly with perception**




**Trend: more and more articles criticizing the inconsistency of these metrics**
**⇨ not completely the metrics' fault… the "one-to-many" problem is just too difficult!**

# Concluding remarks

- Taxonomy in SQA
  - Evaluation target: synthetic speech / non-synthetic speech
  - Subjective / objective
  - Intrusive / non-intrusive
  - Signal-based / model-based
- Long-standing challenge: **out-of-domain generalization** (= all-purpose)
  - Important theme of the Voice/AudioMOS Challenge series

- ***Sooooo many unsolved (and interesting!) problems, even beyond speech!***

# Advertisements

- I have co-authored a review paper on SQA for synthetic speech
  - Mostly done with the amazing Erica Cooper (NICT, Japan)
  - E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, “A review on subjective and objective evaluation of synthetic speech,” Acoustical Science and Technology, vol. 45, no. 4, pp. 161–183, 2024.

- I will co-present a tutorial in INTERSPEECH 2025, also on the title “"Automatic Quality Assessment for Speech and Beyond”
  - With Erica Cooper and Jiatong Shi (CMU, USA)