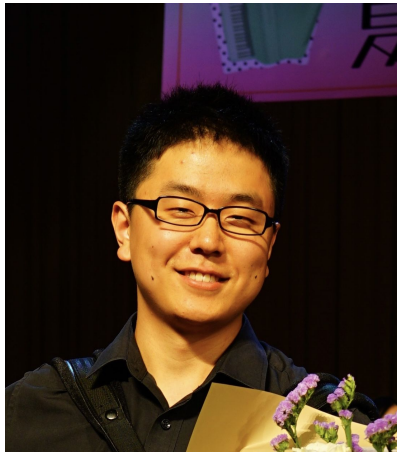
Automatic Quality Assessment for Speech and Beyond

Wen-Chin Huang
Nagoya University, Japan

Erica Cooper
NICT, Japan



Jiatong Shi
CMU, USA

INTERSPEECH 2025 Tutorial 2025.8.17

Outline

- **Part I: Erica Cooper** (45mins presentation + 5mins Q&A)
  - Subjective quality assessment for synthesized speech
  - Objective quality assessment for synthesized speech
- **Part II: Wen-Chin Huang** (45mins presentation + 5mins Q&A)
  - Experiences and lessons from the VoiceMOS Challenge Series
  - Ongoing trends and efforts on quality assessment for speech
- **Part III: Jiatong-Shi** (45mins presentation + 5mins Q&A)
  - Ongoing trends and efforts on quality assessment for music and general audio
  - Applications of quality assessment metrics
- **Part IV: Wen-Chin Huang, Jiatong-Shi** (25mins presentation + 5mins Q&A)
  - Resources: datasets, benchmarks, toolkits

Part I:
Quality Assessment for **Synthesized Speech**

Outline

- Subjective quality assessment for synthesized speech
  - 1980s-1990s: Intelligibility and comprehension
  - 1990s-2000s: Naturalness, intelligibility, and efforts to standardize
  - 2010s-present: Crowdsourcing, MOS, and critiques
  - Current practices and challenges in subjective evaluation
- Objective quality assessment for synthesized speech
  - Motivation and challenges
  - Speech quality assessment metrics from telephony
  - Model-based evaluation of synthesized speech
    - Early machine learning approaches
    - Neural network based approaches
    - SSL-based approaches
    - Unsupervised approaches
    - Beyond MOS prediction

Subjective Quality Assessment (Listening Tests)

1970s-1990s: Intelligibility and comprehension

Early speech synthesizers had a robotic sound, and the first challenge is to generate **intelligible** speech – naturalness would only become a primary consideration later.

- Diagnostic Rhyme Test [1]
- Modified Rhyme Test [2]
  `BAD BACK BAN BASS BAT BATH`
- Word pointing test [3]
- Minimal pairs intelligibility test [3]
- Word and name transcription tasks [3]
- Quality ratings with problem categorizations [3]
- Paired comparisons with certainty ratings [3]
- Transcription of semantically-unpredictable sentences [4]
  `The table walked through the blue truth.`

1970s: formant synthesis

1980s: diphone synthesis

Thanks to Junichi Yamagishi and Simon King for these audio samples which were generated using Festival.

1970s-1990s: Intelligibility and comprehension

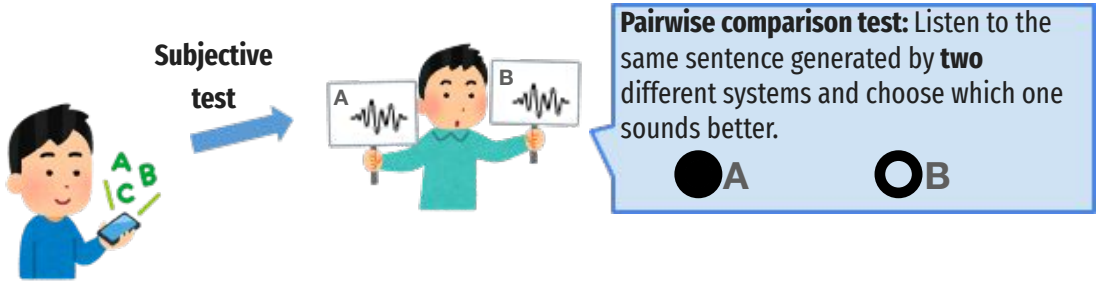
Comprehension: a listener's ability to understand and retain information in the speech.

- Multiple choice listening comprehension test [5]
- Sentence verification task [6]

Although listening comprehension tests have strong **ecological validity,** they are easily **saturated** (meaningful distinctions cannot be found between natural and synthetic speech). For this reason, these kinds of tests have not seen widespread use.

Some popular listening tests

Pairwise comparison tests



Pros:

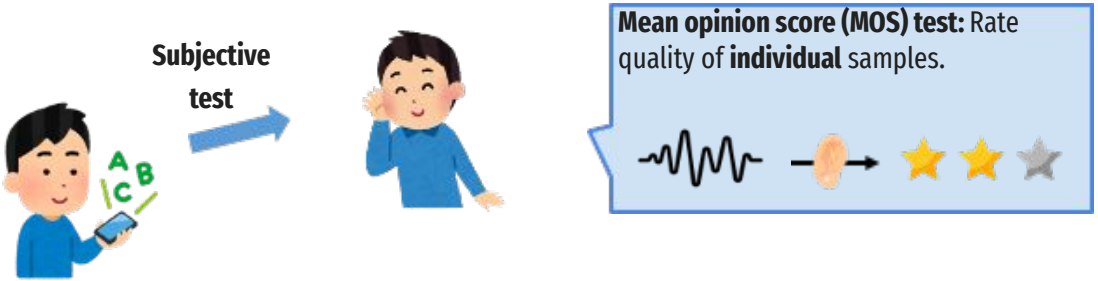
- Direct comparisons between two systems allow fine-grained distinctions to be made

Cons:

- Doesn't scale well to a large number of systems because you have to compare all pairs

Some popular listening tests

MOS (Mean Opinion Score) tests (ITU-T Rec. P800 [7])



Pros:

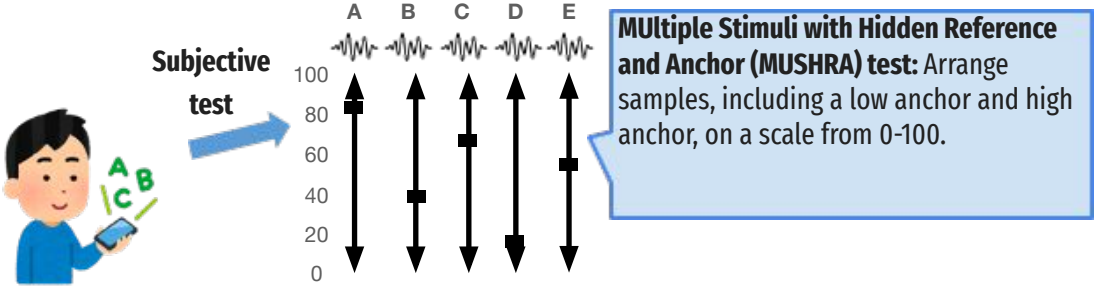
- Can compare more different systems

Cons:

- Comparison between systems is **implicit**, not direct
- Listener ratings will be influenced by the overall range of quality in the entire test

Some popular listening tests

MUSHRA(-like) (ITU-R BS.1534-3 [8])



MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test: Arrange samples, including a low anchor and high anchor, on a scale from 0-100.

Pros:

- Direct and fine-grained comparisons between multiple systems
- Fewer listeners required to find statistically-significant differences between systems

Cons:

- Can't include too many systems
- More effort for listeners

Beyond listening tests

- Reaction time to measure cognitive processing
  - e.g., sentence verification task
  - Paired word choice task – reaction times are approaching or even indistinguishable from those for natural speech [9]
- Pupillometry to measure cognitive load
  - Pupil dilation is sensitive to the quality of synthesized speech [10]
- Brain signals while listening to synthetic speech
  - PhySyQX dataset [11]: multidimensional rating of synthetic speech + fNIRS + EEG data; found to correlate with ratings and can be used for decision tree classification-based prediction [12]
  - Different fMRI patterns when listening to deepfake audio (more activity in the auditory cortex) vs. natural speech (more activity in the ventral striatum) [13]

1990s-2000s: Naturalness, intelligibility, and efforts to standardize

As speech synthesis quality improved, the focus of evaluations shifted more towards **naturalness.**

- ITU-T Rec. P.80 [14]; P.85 [15]: overall impression, comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness, acceptance
  - In 2002, 8 years after it was introduced, P.85 had not seen much adoption [16]
  - Very strong correlations across several of the different P.85 rating scales -> **unnecessary complexity**
  - A later study [17] found opposite results – it depends on the TTS systems.
  - Pairwise test gave same rankings with less variability and more significant differences with the same number of listeners [16]
  - P.85 was found to be unsuitable for measuring intelligibility; SUS tests were more useful [17]

1990s: unit selection 

2000s: HMM synthesis 

Thanks to Junichi Yamagishi and Simon King for these audio samples which were generated using Festival.

1990s-2000s: Naturalness, intelligibility, and efforts to standardize

The Blizzard Challenge (2005 -)

- Shared task for comparing corpus-based text-to-speech synthesis techniques using standardized datasets and evaluations
- MOS, SUS intelligibility (and speaker similarity, later)
- Strong precedent for TTS evaluation
- Listening test results are shared with the community **-> DATA!**

2010s-present: Crowdsourcing, MOS, and critiques

Crowdsourcing: Ask anonymous participants online to do annotation or answer a survey through a web-based platform

Pros: Access to a much larger pool of participants; tasks can be completed very quickly and asynchronously

Cons: Less control over the listening environment

Challenges:

- Inattentive listeners → require entire audio sample to be played; include some "attention check" questions; require some threshold of previously-accepted tasks; check the results
- Unknown listeners → collect some demographic information (but be mindful of privacy)
- Unknown listening environment → require headphones; Huggins pitch [18]

2010s-present: Crowdsourcing, MOS, and critiques

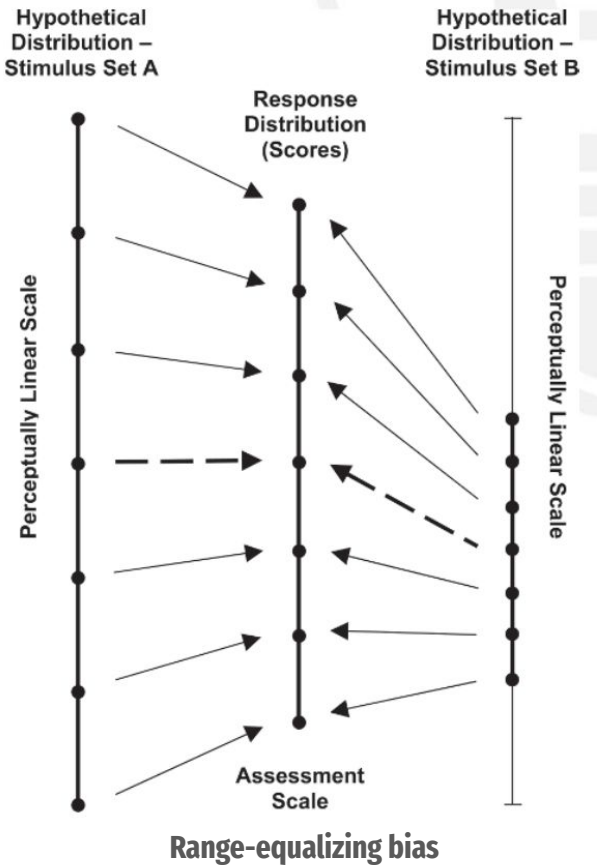
Critiques of MOS

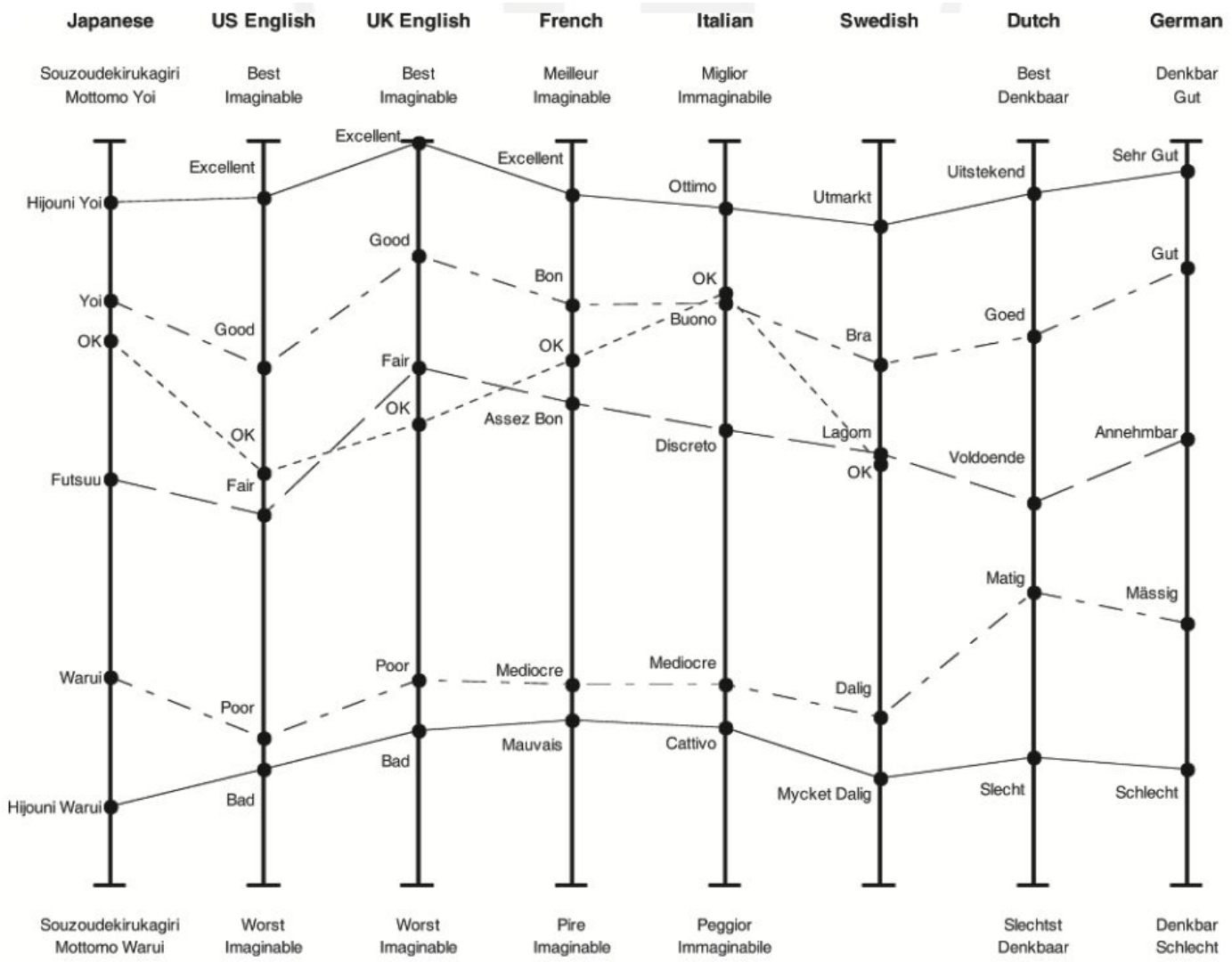
- What even **is** naturalness?? [19,20,21]
- Specific guidelines exist but are often not followed [22]
- Number of listeners, questions asked, etc. are often not reported [22]
- Averaging loses distributional information [23]
- MOS results cannot be meaningfully compared across separate listening tests [23]
- Evaluation of isolated sentences is not ecologically valid [19]
- Instructions given to listeners affect the results [24]
- **So many biases!!!** (this also applies to other kinds of listening tests) [25]

2010s-present: Crowdsourcing, MOS, and critiques

Biases in listening tests [25]

- Biases arising from **affective judgments**
  - Appearance of testing equipment
  - Expectations
  - Personal preferences
  - Emotions and mood
- **Response mapping bias** arising from the test design
  - Stimulus spacing and frequency
  - Perceptually nonlinear scales
  - Range-equalizing bias
- **Interface bias**
  - Layout and appearance of the scale
  - Words chosen for the labels




Japanese
US English
UK English
French
Italian
Swedish
Dutch
German
Souzoudekirukagiri Mottomo Yoi
Best Imaginable
Best Imaginable
Meilleur Imaginable
Miglior Immaginabile
Best Denkbaar
Denkbar Gut
Hijouni Yoi
Yoi
OK
Futsuu
Warui
Hijouni Warui
Excellent
Good
OK
Fair
Poor
Bad
Excellent
Good
OK
Fair
Poor
Bad
Excellent
Bon
OK
Assez Bon
Mediocre
Mauvais
Ottimo
OK
Buono
Discreto
Mediocre
Cattivo
Utmarkt
Bra
Lagom
OK
Dalig
Mycket Dalig
Uitstekend
Goed
Voldoende
Matig
Slecht
Sehr Gut
Gut
Annehmbar
Mässig
Schlecht
Souzoudekirukagiri Mottomo Warui
Worst Imaginable
Worst Imaginable
Pire Imaginable
Peggior Immaginabile
Slechtst Denkbaar
Denkbar Schlecht

Current practices and challenges in subjective evaluation

Do not cross-compare MOS test results!!!

Report your listening test details!!! (number of listeners, etc.)

Consider ecological validity!

MOS is arguably saturated – consider other tests with more discriminative power.

- Pairwise tests have greater discriminative power, can avoid n^2 with active learning type approaches [26]
- **De-saturate** your listening tests by choosing challenging / well-differentiated test material [27,28,29]

Objective Quality Assessment

Motivation and challenges

Motivation: Listening tests are a time-consuming evaluation paradigm. Reducing the listening effort and time burden can speed up experimentation cycles and also make listening tests more efficient.

Challenges:

- **Listening test context dependence:** All of the biases inherent to listening tests mean that their results are not directly cross-comparable.
- **Subjective listener preferences:** Listeners' ratings may be influenced by their cultural background, personal opinions, familiarity with technology, etc.
- **Generalization to new domains:** Even if MOS ratings for one type of synthesized speech are available, it is not predictable how well a model trained on this data will generalize to other types of synthetic speech (different languages, types of expressivity, use cases, etc.)
- **The "one-to-many" problem:** Given a particular input text, there are many valid and natural-sounding ways to speak it.

MOS is subjective. It is also not absolute but relative.

Evaluation metrics

System-level and **Utterance-level** evaluation metrics:

- **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

Taxonomy of objective evaluation methods

Type of reference data available:

- **Matched reference audio sample** AKA intrusive; double-ended
- **Partially-matched reference audio sample** e.g., speaker-matched but not lexically-matched audio sample to measure speaker similarity
- **Reference text** e.g., ASR WER for intelligibility
- **Non-matched reference audio** e.g., distribution modeling
- **No reference data** AKA non-intrusive; single-ended; reference-free

Evaluation type:

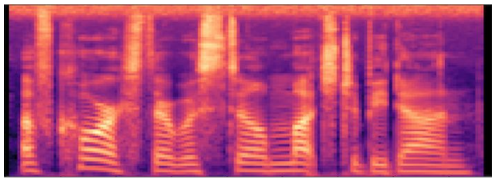
- MOS prediction / quality rating
- A/B pairwise comparison
- Similarity to a reference
- Multi-dimensional evaluation
- .....

Speech quality assessment metrics from telephony

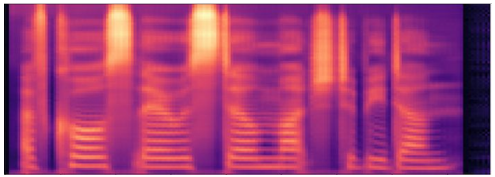
Signal-based intrusive metrics:

- **Mel-cepstral distance (MCD)** [30]: Difference between the Mel cepstra of a reference and a test speech sample. *Adapted for TTS using dynamic time-warping alignment.* [31]
- **Root mean squared error of f0:** Distance between reference and test f0 sequences
- **Correlation of f0:** How well changes in direction of f0 in a test sample match the reference

Real speech



Synthesized speech



Speech quality assessment metrics from telephony

Perceptual Evaluation of Speech Quality (PESQ) [32,33]: an **intrusive** algorithm designed for narrow-band telecommunications applications

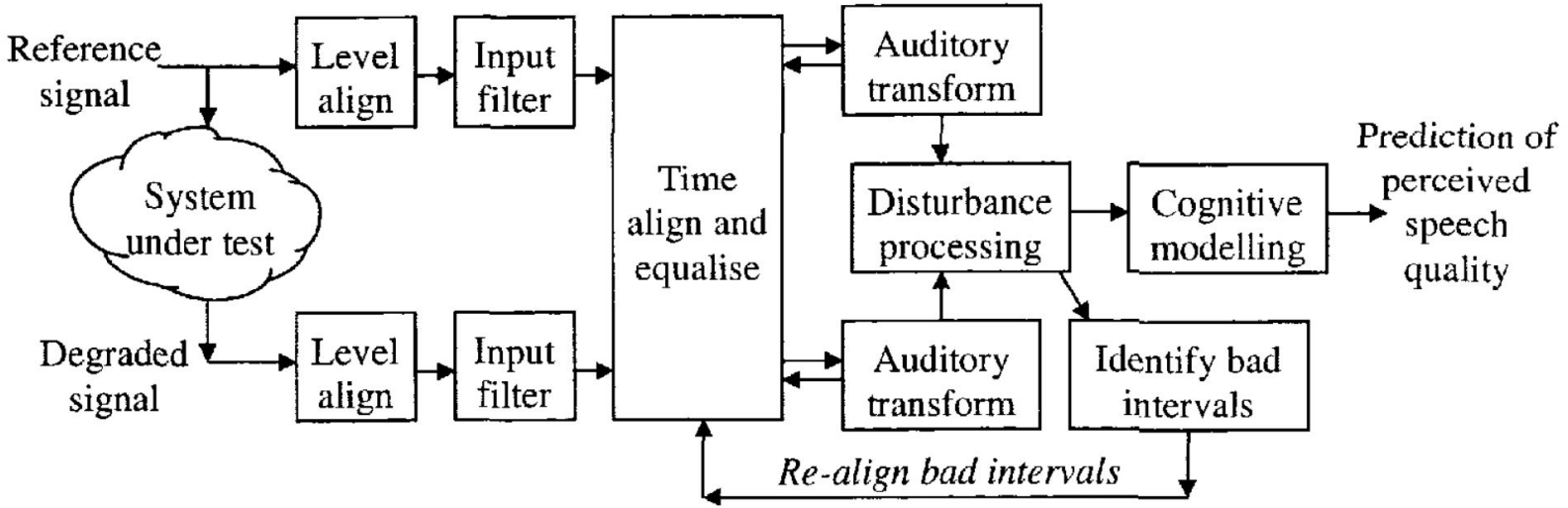
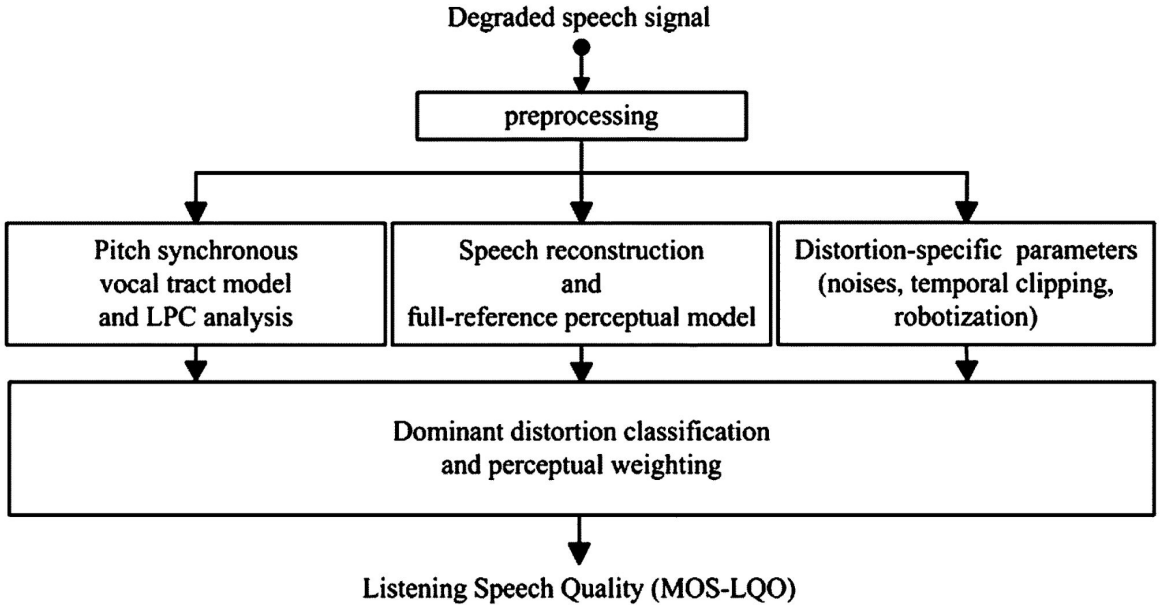


Figure 1: Structure of perceptual evaluation of speech quality (PESQ) model.

Speech quality assessment metrics from telephony

ITU Recommendation P.563 [34,35]: the first **reference-free** metric developed by the ITU for narrow-band telephony.



Model-based evaluation: Early machine learning approaches

- 2008: Decision trees to identify the most useful internal features of P.563 [36]
- 2010: Linear regression models and SVMs to incorporate larger-scale acoustic feature sets related to signal duration, formants, intensity, pitch, and spectrum [37]
  - Evaluated on datasets from Blizzard Challenge 2008 and 2009
  - Correlations in the range of 0.7-0.8
- 2012: Prosodic, micro-prosodic, and MFCC-based features; feature selection and SVM [38]
  - Evaluated on more challenging data from Blizzard 2012 including synthesized paragraphs
  - Combination of features produces the best results
- 2015: Prediction of naturalness, prosody, and intelligibility [39]
  - Large feature sets + SVMs
  - Incorporation of nonlinear modeling: "regular perception range" in which correlations between acoustic features and the quality rating is maximized
  - Correlations upwards of 0.9

Model-based evaluation: Neural network based approaches

- 2016: **Hierarchical approach** to first predict a system-level score and then use that prediction as a feature to predict sample-level scores [40]
  - DNNs were shown to work better than regression models at both stages.
- 2016: **AutoMOS** [41]
  - LSTMs trained on internal MOS test data
- 2019: **MOSNet** [42] for voice-converted speech
  - Based on QualityNet [43] for enhanced speech
  - CNN-BLSTM
  - Could also predict speaker similarity
  - Open source → popular use
- 2020: **NISQA-TTS** [44]
  - CNN-LSTM
  - Pretrained for degraded natural speech; fine-tuned on multilingual MOS datasets of TTS
  - Open source → popular use

Model-based evaluation: Neural network based approaches

Listener modeling: incorporate listener ID and individual ratings into model training.

- 2021: **MBNet** [45]: mean and bias subnets to learn individual listener preferences and averaged scores at the same time
- 2022: **LDNet** [46]: "all listeners" and "mean listener" inference modes
- 2023: **DeePMOS** [47]: MBNet + variance prediction

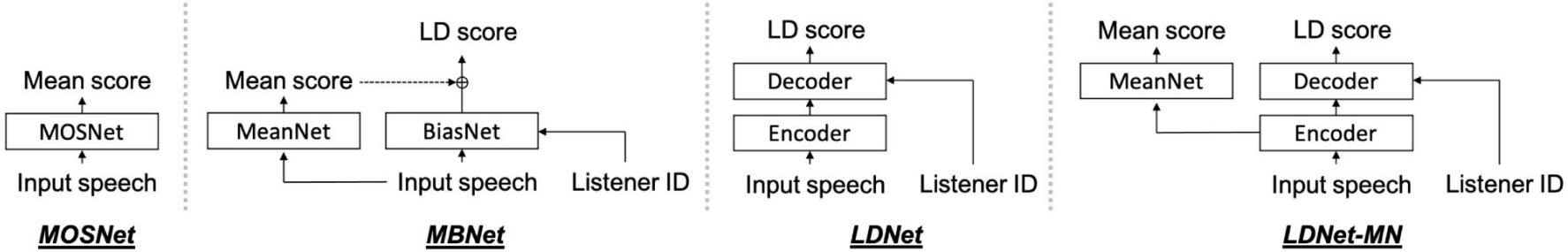
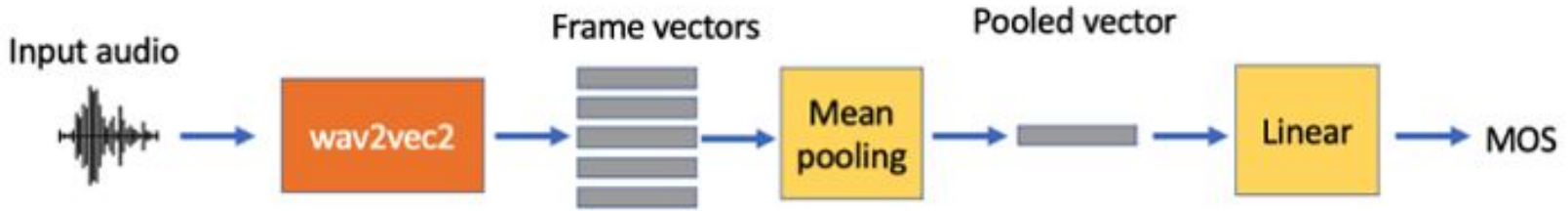


Fig. 1: Illustration of the models described in this work. From lest to right: MOSNet, MBNet, LDNet, LDNet with MeanNet multitask learning (LDNet-MN).

Model-based evaluation: SSL-based approaches

Basic idea: fine-tune a pretrained SSL model on MOS-labeled data.

- Simple approach: SSL-MOS [48] – fine-tune Wav2Vec2 + linear layer for MOS prediction as a regression task using L1 loss



Model-based evaluation: SSL-based approaches

Basic idea: **fine-tune** a pretrained SSL model on MOS-labeled data.

- Simple approach: SSL-MOS [48] – fine-tune Wav2Vec2 + linear layer for MOS prediction as a regression task using L1 loss
- Extensions and improvements:
  - UTMOS [49]: ensembling of strong (SSL) and weak (regression) learners; contrastive loss
  - DDOS [50]: domain-adaptive SSL pretraining on synthesized speech; distribution prediction
  - Encoders for prosodic and linguistic features [51]
  - ZevoMOS [52]: incorporate ASR confidence scores and pretrain for the real/fake speech classification task
  - RAMP [53]: non-parametric component based on *k*NN
  - Fusion of 7 SSL models [54]
  - SQuId [55]: massively multilingual training on MOS datasets for 52 language locales

Model-based evaluation: Unsupervised approaches

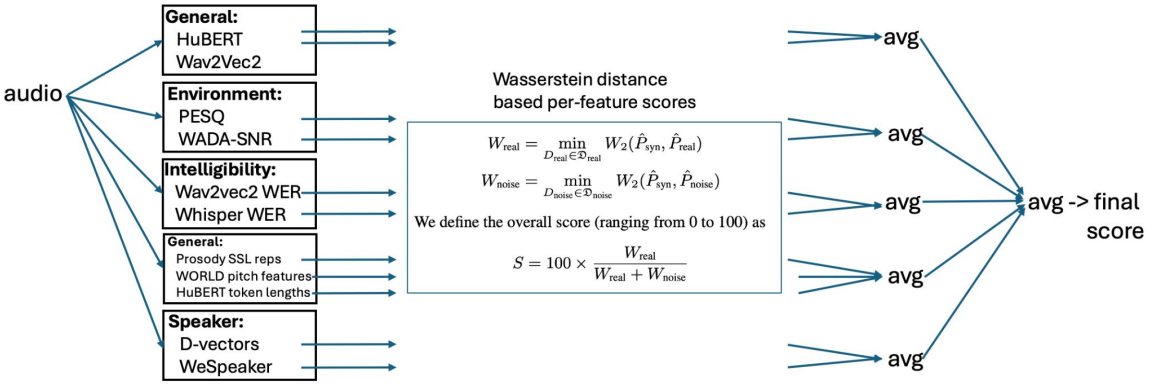
Basic idea: build a **reference model** from natural speech and measure the distributional similarity or difference between synthesized and natural speech.

- Precedent in the HMM synthesis era [56]: Gaussian mixture models of natural and synthesized speech acoustic features
- SpeechLMScore [57]: Token sequences from a pretrained SSL; k-means clustering; perplexity of token cluster sequences is measured wrt. a speech language model trained on natural speech
- UNIQUE [58]: Uses a distribution model of k-means cluster sequences instead of a sequence model
- VQScore [59]: A VQ-VAE trained on natural clean speech; reconstruction error as quality measure
- SpeechBERTScore [60]: **Intrusive** approach – lexically-matched reference samples are required. Embeddings of ground-truth and synthesized speech samples are extracted using an SSL model and cosine similarity is computed.

Model-based evaluation: Unsupervised approaches

Approaches based on Fréchet Inception Distance: proposed in 2017 to evaluate image generation models – compares the distribution of two datasets in an embedding space (from the last pooling layer of the Inception-V3 model) using the 2-Wasserstein distance [61]

- Fréchet Audio Distance [62]: adapted for evaluating music enhancement algorithms by using embeddings from VGGish
- Adapted for evaluating TTS specifically [63] by using embeddings from the last layer of the DeepSpeech2 ASR model
- TTSDS [64]: an ensemble of models from which embeddings are extracted and their distributions are compared
- TTSDS2 [65]: extended to multilingual



Beyond MOS prediction

- Word error rate from automatic speech recognizers (ASR WER) has come to basically replace human intelligibility evaluations, with correlations to human-transcribed WER reported at 0.94 in 2015 [66]
- Cosine similarity between speaker embeddings extracted from synthesized speech and natural speech of the target speaker is often used as an objective measure of speaker similarity, with correlations to human ratings reported at 0.85 using x-vectors in 2020 [67]; 0.75 when using ECAPA embeddings in 2024 [68]

References

[0] 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* 45, no. 4 (2024): 161-183.

[1] W. D. Voiers, “Evaluating processed speech using the diagnostic rhyme test,” *Speech Technol.*, vol. 1, pp. 30–39, 1983.

[2] A. S. House, C. Williams, M. H. Hecker, and K. D. Kryter, “Psychoacoustic speech tests: A modified rhyme test,” *The Journal of the Acoustical Society of America*, vol. 35, pp. 1899–1899, 1963.

[3] J. P. van Santen, “Perceptual experiments for diagnostic testing of text-to-speech systems,” *Computer Speech & Language*, vol. 7, no. 1, pp. 49–100, 1993.

[4] M. Grice, “Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages,” in *Proc. Speech Input/Output Assessment and Speech Databases*, Vol.2, pp. 19–22, 1989.

[5] D. Pisoni and S. Hunnicutt, “Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system,” ICASSP, 1980.

[6] D. Pisoni, L. Manous, and M. Dedina. "Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility." *Computer speech & language* 2.3-4: 303-320, 1987.

[7] Methods for subjective determination of transmission quality,” in *ITU-T Rec. P.800*. International Telecommunication Union (ITU-R)., 1996.

[8] “Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA),” in *Recommendation ITU-R BS.1534-3*. International Telecommunication Union (ITU-R)., 2015.

[9] Z. Malisz,G. E. Henter, C. Valentini Botinhao, O. Watts, J. Beskow, and J. Gustafson. "Modern speech synthesis for phonetic sciences: A discussion and an evaluation." In *19th International Congress of Phonetic Sciences*, pp. 487-491. Australian Speech Science & Technology Association Inc, 2019.

References

[10] A. Govender, A.E. Wagner, S. King. "Using Pupil Dilation to Measure Cognitive Load When Listening to Text-to-Speech in Quiet and in Noise." Proc. Interspeech, 2019.

[11] R. Gupta, H. J. Banville, and T. H. Falk. "PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience." WASPAA, 2015.

[12] R. Gupta. "Physiology-based Quality-of-Experience Assessment for Next Generation Multimedia Technologies." Diss. Institut National de la Recherche Scientifique (Canada), 2016.

[13] C. Roswandowitz, T. Kathiresan, E. Pellegrino, V. Dellwo, and S. Frühholz. "Cortical-striatal brain network distinguishes deepfake from real speaker identity." *Communications Biology* 7, no. 1 (2024): 711.

[14] “Methods for subjective determination of transmission quality,” in *ITU-T Rec. P.80*. International Telecommunication Union (ITU-R)., 1993.

[15] A method for subjective performance assessment of the quality of speech voice output devices,” in *ITU-T Rec. P.85*. International Telecommunication Union (ITU-R)., 1994

[16] Y. V. Alvarez and M. Huckvale, “The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems,” ICSLP 2002.

[17] D. Sityaev, K. Knill, and T. Burrows, “Comparison of the ITU-T P.85 standard to other methods for the evaluation of text-to-speech systems,” Interspeech 2006.

[18] A. E. Milne, R. Bianco, K. C. Poole, S. Zhao, A. J. Oxenham, A. J. Billig, and M. Chait. "An online headphone screening test based on dichotic pitch." *Behavior Research Methods* 53, no. 4 (2021): 1551-1562.

[19] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Éva Székely, C. Tånnander, and J. Voße, “Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program,” ISCA SSW 10, 2019.

References

[20] R. Dall, J. Yamagishi, and S. King, “Rating naturalness in speech synthesis: The effect of style and expectation,” Speech Prosody 2014.

[21] S. Shirali-Shahreza and G. Penn, “Better Replacement for TTS Naturalness Evaluation,” SSW 2023.

[22] M. Wester, C. Valentini-Botinhao, and G. E. Henter, “Are we using enough listeners? No! — An empirically-supported critique of interspeech 2014 TTS evaluations,” Interspeech 2015.

[23] R. C. Streijl, S. Winkler, and D. S. Hands, “Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives,” *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[24] J. O’Mahony, P. O. Gallegos, C. Lai, and S. King, “Factors affecting the evaluation of synthetic speech in context,” ISCA SSW 11, 2021.

[25] S. Zielinski, F. Rumsey, and S. Bech, “On some biases encountered in modern audio quality listening tests-a review,” *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[26] Y. Yasuda and T. Toda, “Automatic design optimization of preference-based subjective evaluation with online learning in crowdsourcing environment,” arXiv:2403.06100, 2024.

[27] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. "Fastspeech: Fast, robust and controllable text to speech." *Advances in neural information processing systems* 32 (2019).

[28] J. Chevelu, D. Lolive, S. Le Maguer, and D. Guennec. "How to compare TTS systems: a new subjective evaluation methodology focused on differences." Interspeech 2015.

[29] O. Perrotin, B. Stephenson, S. Gerber, G. Bailly, and S. King. "Refining the evaluation of speech synthesis: A summary of the Blizzard Challenge 2023." *Computer Speech & Language* 90 (2025): 101747.

References

[30] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in IEEE Pacific Rim Conference on Communications Computers and Signal Processing, 1993.

[31] J. Kominek, T. Schultz, and A. W. Black, “Synthesizer voice quality of new languages calibrated with mean Mel cepstral distortion,” in SLTU, 2008.

[32] “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” ITU-T Recommendation P.862, 2001.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," ICASSP 2001.

[34] Single-ended method for objective speech quality assessment in narrow-band telephony applications,” ITU-T Rec. P.563, 2004.

[35] L. Malfait, J. Berger and M. Kastner, "P.563—The ITU-T Standard for Single-Ended Speech Quality Assessment," ICASSP 2006.

[36] T. H. Falk, S. Möller, V. Karaiskos, and S. King, “Improving instrumental quality prediction performance for the Blizzard Challenge,” in Proc. Blizzard Challenge Workshop, 2008.

[37] F. Hinterleitner, S. Möller, T. H. Falk, and T. Polzehl, “Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: Data from Blizzard Challenges 2008 and 2009,” in Proc. Blizzard Challenge Workshop, vol. 2010, 2010, pp. 48–60.

[38] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, “Towards perceptual quality modeling of synthesized audiobooks – Blizzard Challenge 2012,” in Proc. Blizzard Challenge Workshop, 2012.

[39] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, “Quality prediction of synthesized speech based on perceptual quality dimensions,” *Speech Communication*, 2015.

References

[40] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, “A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks,” Interspeech 2016.

[41] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, “AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech,” arXiv:1611.09207, 2016.

[42] C.-C. Lo, S.-W. Fu, W.-C. Huang, et al., “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion,” Interspeech 2019.

[43] 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,” Interspeech 2018.

[44] G. Mittag and S. Möller, “Deep Learning Based Assessment of Synthetic Speech Naturalness,” Interspeech 2020.

[45] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, “MBNet: MOS prediction for synthesized speech with mean-bias network,” ICASSP 2021.

[46] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, “LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech,” ICASSP 2022.

[47] X. Liang, F. Cumlin, C. Schüldt, and S. Chatterjee, “DeePMOS: Deep Posterior Mean-Opinion-Score of Speech,” Interspeech 2023.

[48] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, “Generalization ability of MOS prediction networks,” ICASSP 2022.

[49] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022,” Interspeech 2022.

References

[50] W.-C. Tseng, W.-T. Kao, and H.-Y. Lee, “DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores,” Interspeech 2022.

[51] A. Vioni, G. Maniati, N. Ellinas, et al., “Investigating Content-Aware Neural Text-to-Speech MOS Prediction Using Prosodic and Linguistic Features,” ICASSP 2023.

[52] A. Stan, “The ZevoMOS entry to VoiceMOS Challenge 2022,” Interspeech 2022.

[53] H. Wang, S. Zhao, X. Zheng, and Y. Qin, “RAMP: Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting,” Interspeech 2023.

[54] Z. Yang, W. Zhou, C. Chu, S. Li, R. Dabre, R. Rubino, and Y. Zhao, “Fusion of Self-supervised Learned Models for MOS Prediction,” Interspeech 2022.

[55] T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh, and J. Riesa, “SQuId: Measuring speech naturalness in many languages,” ICASSP 2023.

[56] S. L. Maguer, N. Barbot, and O. Boeffard, “Evaluation of contextual descriptors for HMM-based speech synthesis in French,” SSW 2013.

[57] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, “SpeechLMScore: Evaluating speech generation using speech language model,” ICASSP 2023.

[58] J. Yoon, W. Ko, S. Um, et al., “UNIQUE: Unsupervised network for integrated speech quality evaluation,” Interspeech 2024.

[59] S.-W. Fu, K.-H. Hung, Y. Tsao, and Y.-C. F. Wang, “Self-supervised speech quality estimation and enhancement using only clean speech,” ICLR 2024.

References

[60] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, “SpeechBERTScore: reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics,” Interspeech 2024.

[61] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” NeurIPS 2017.

[62] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” Interspeech, 2019.

[63] M. Bińkowski, J. Donahue, S. Dieleman, et al., “High fidelity speech synthesis with adversarial networks,” ICLR 2020.

[64] C. Minixhofer, O. Klejch, and P. Bell, “TTSDS - text-to-speech distribution score,” SLT 2024.

[65] C. Minixhofer, O. Klejch, and P. Bell, “TTSDS2: resources and benchmark for evaluating human-quality text to speech systems,” SSW 2025.

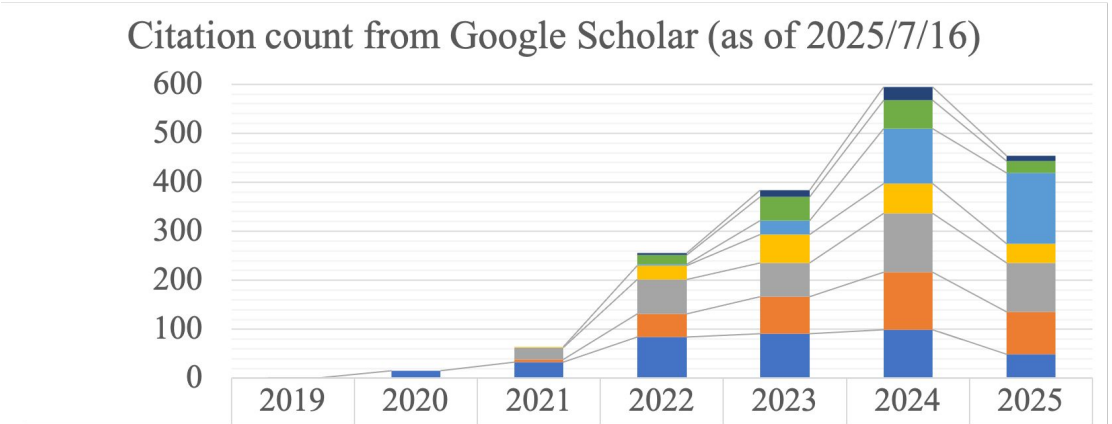
[66] F. Hinterleitner, S. Zander, K.-P. Engelbrecht, and S. Möller, “On the use of automatic speech recognizers for the quality and intelligibility prediction of synthetic speech,” in Konferenz Elektronische Sprachsignalverarbeitung, 2015.

[67] R. K. Das, T. Kinnunen, W.-C. Huang, et al., “Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions,” in Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge, 2020.

[68] J. Ahn, Y. Kim, Y. Choi, et al., “VoxSim: a perceptual voice similarity dataset,” 2024.

Part II-(1): Experiences and lessons from the VoiceMOS Challenge Series

Impact of the VMC series



	2019	2020	2021	2022	2023	2024	2025
CSC2022 **				4	13	27	11
VMC22 **				19	49	57	24
UTMOS *				3	29	112	144
SSL-MOS *			2	29	57	61	40
DNSMOS *			24	69	70	120	100
NISQA *			5	48	75	118	86
MOSNet *	1	16	33	84	91	99	49

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 path of the VMC series

- The VoiceMOS Challenge 2022 @ INTERSPEECH — **22** participants
  - **In-domain** prediction for synthetic speech (TTS, VC)
  - Results: best system achieved **0.939 SRCC**
- The VoiceMOS Challenge 2023 @ ASRU — **10** participants
  - **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 — **8** participants
  - **Diverse tasks**: zoomed-in tests, singing conversion/synthesis, semi-supervised SQA
- The **AudioMOS** Challenge 2025 @ ASRU — **24** participants
  - Expand to **general audio: text-to-speech/audio/music**; different sampling frequencies

Having a baseline toolkit is important

- Each year we provide baseline starter toolkits such that:
  - The baseline is more often than not the **STATE-OF-THE-ART**
- Most important reason: to measure **PROGRESS**
  - One of the goals of scientific challenges is to advance the field
  - If the baseline is SOTA and gets outperformed, then there is progress!
  - Since it's the starter toolkit, it's easier for participants to just focus on improving it
- Other benefit: trigger competition (although it's not the main goal of the challenge)

VMC 2022: tracks

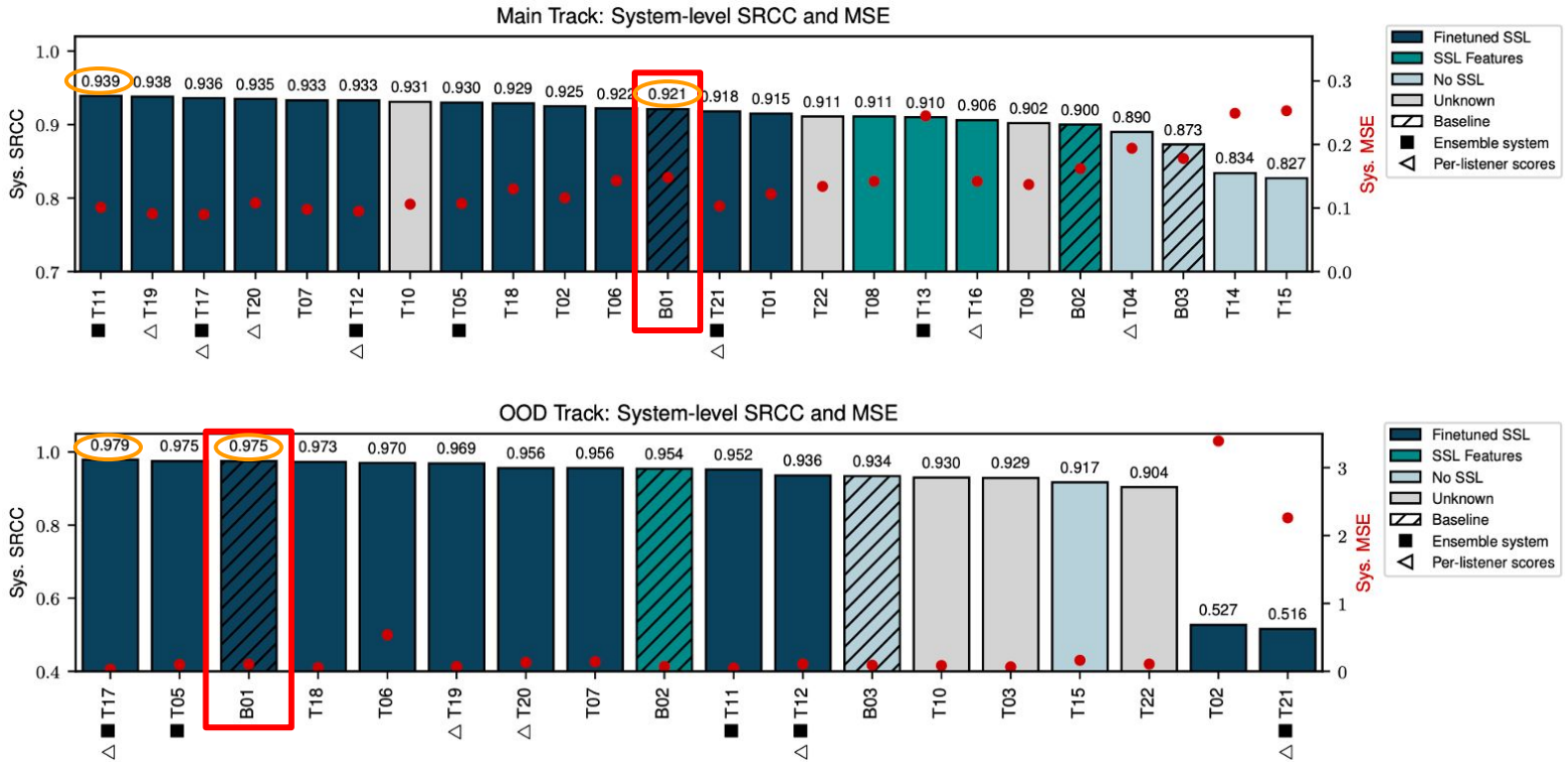
Track	Lang	# Samples			# ratings per sample
		Train	Dev	Test	
Main	Eng	4,974	1,066	1,066	8
OOD	Chi	Label: 136 Unlabel: 540	136	540	10-17

- **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

"OOD track": Probably a bad name...
"limited-data" track might be better

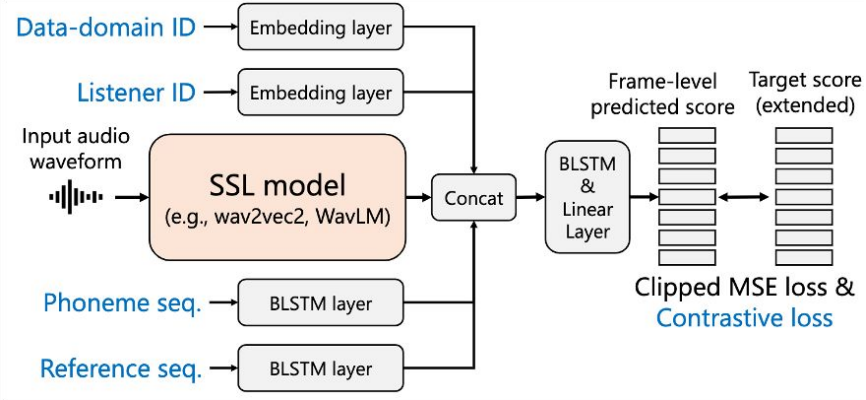
VMC 2022: results

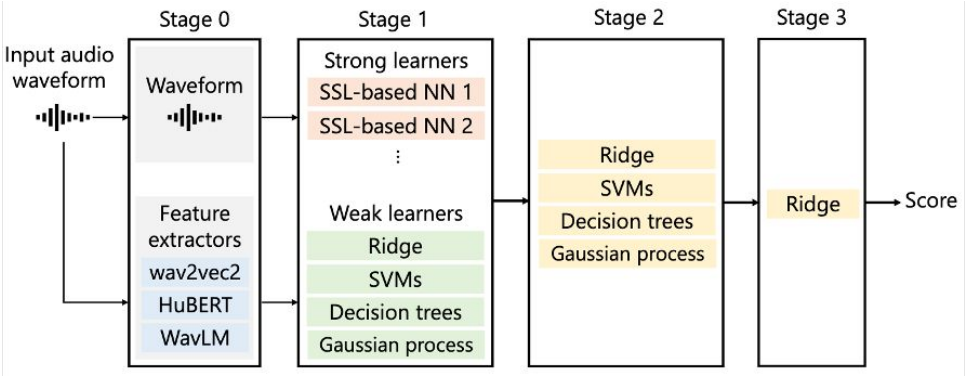
- **Improvements over baseline**
- **Good performance even with 136 samples only**
  ⇨ in-domain is probably "too simple"?



VMC 2022: top system – UTMOS

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





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)

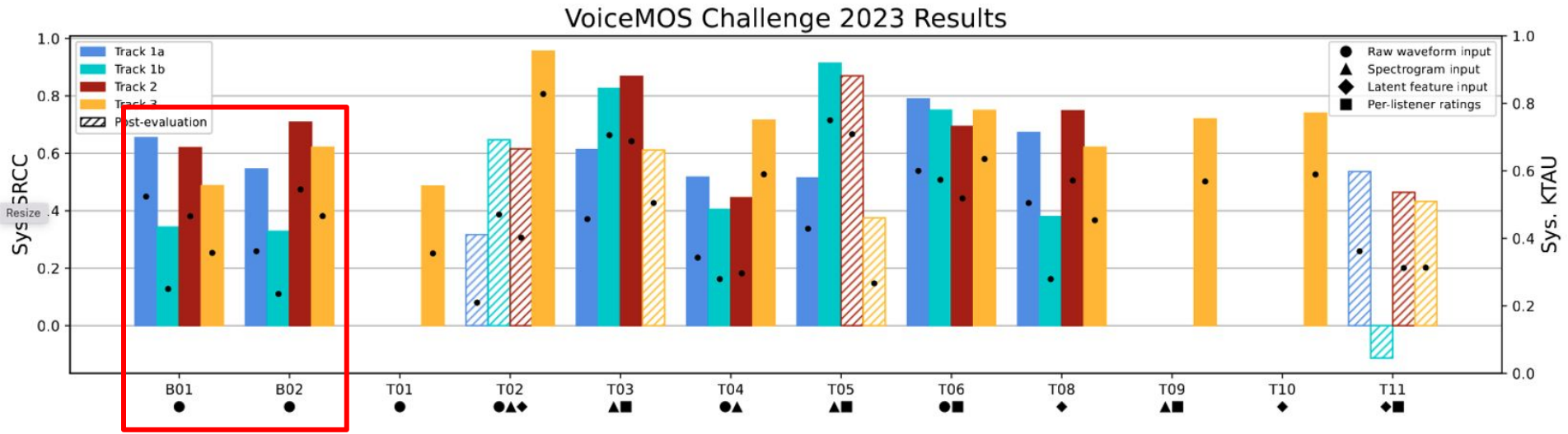
VMC 2023: tracks

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

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

No official training data
= Complete out-of-domain!

VMC 2023: results



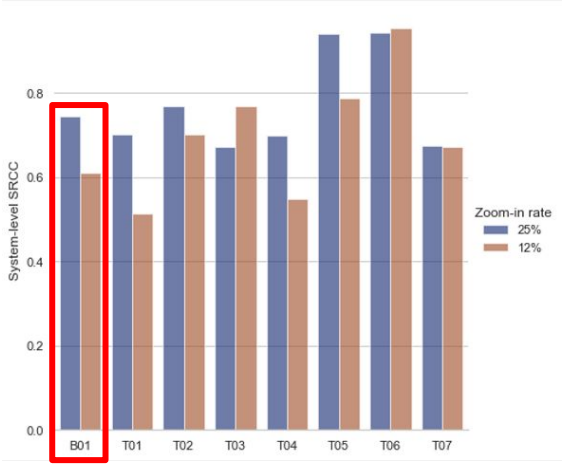
- **Improvements over baseline**
- **Good performance even with 136 samples only**
  ⇨ in-domain is probably “too simple”?

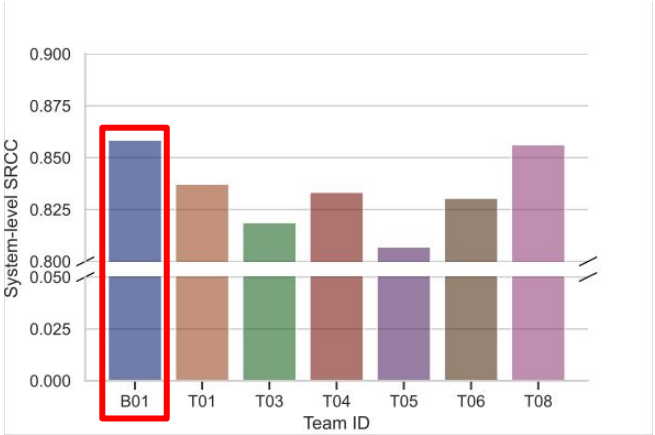
VMC 2024: tracks

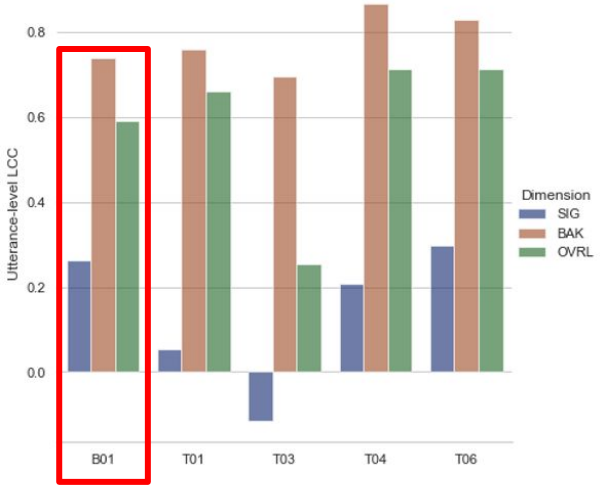
More diverse tasks!

- **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



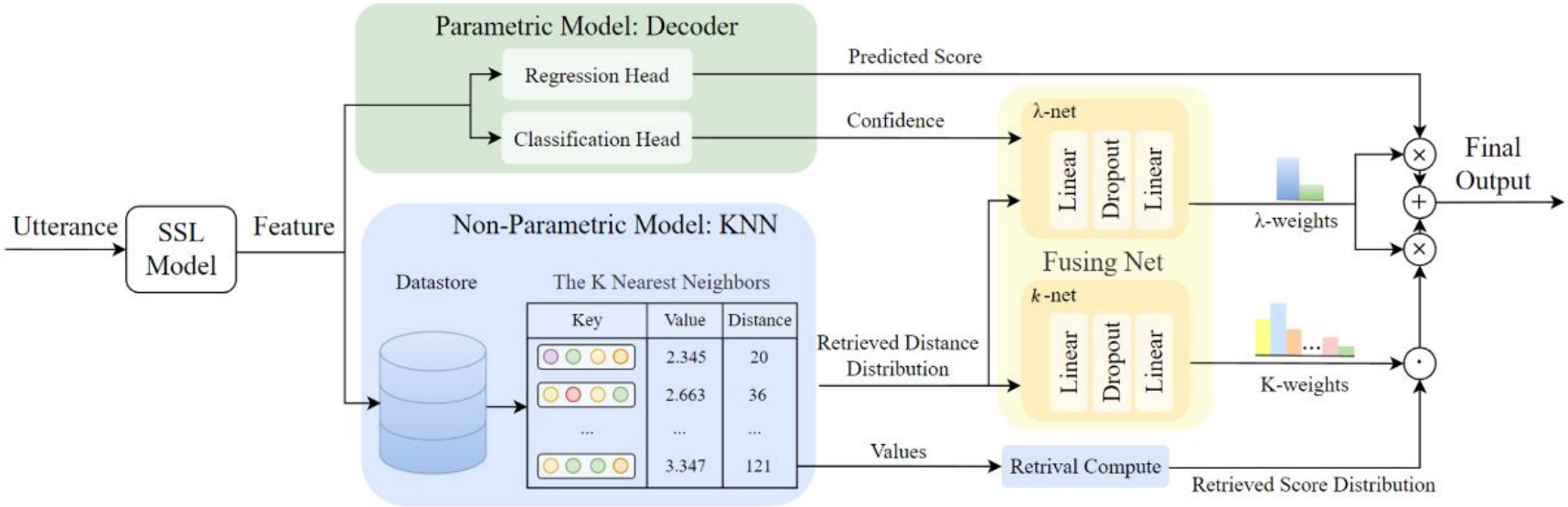




- **Some systems beat the baselines**
- **We had less participants this year, thus less insights & feedbacks...**

VMC 2023 & 2024: common top system – RAMP

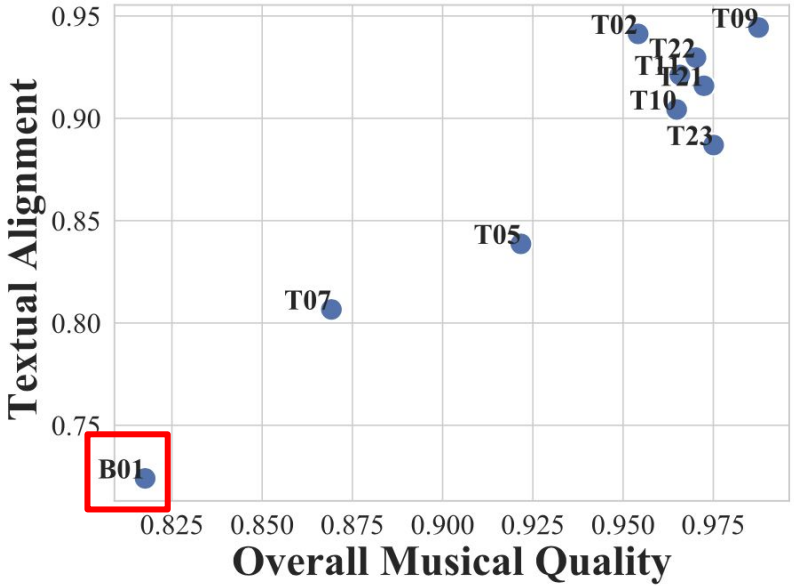
- Augments SSL-MOS with a non-parametric, retrieval head
  - Maintains a “datastore”, which has <SSL feature, score> pairs
  - Given an input sample, retrieve from the datastore the closest sample and its score

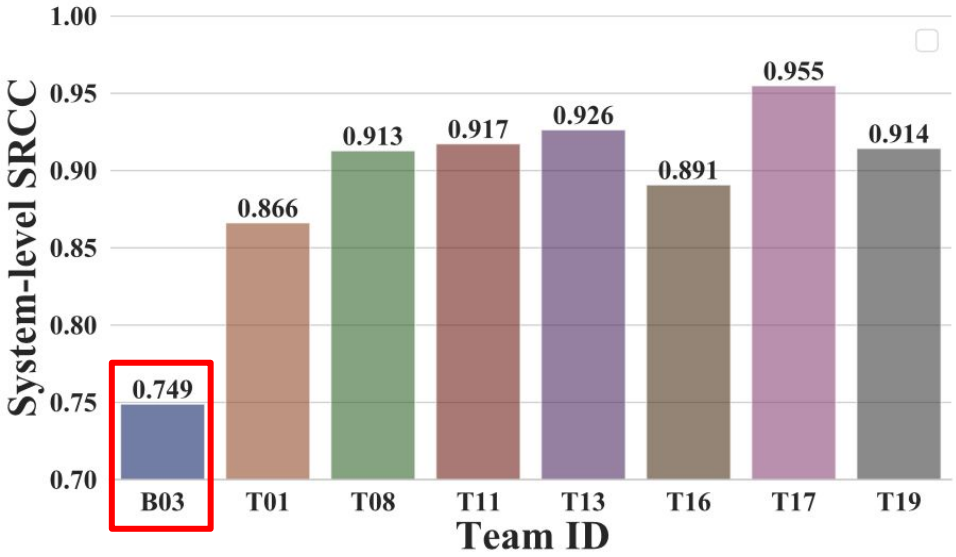


AMC 2025: tracks

- Why from VoiceMOS to AudioMOS?
  - Rapid development of audio generation (speech, singing voice, music, etc.)
  - We felt automatic quality assessment other than speech was lagging
- **Track 1: MOS prediction for text-to-music samples, based on MusicEval**
  - MusicEval: music clips from 31 TTM systems, rated by music experts
  - Two axes: overall musical quality, and alignment with the text prompt
- **Track 2: Predict Audiobox Aesthetics axes**
  - Audiobox Aesthetics: unified assessment methods for speech, music, and sound
  - Four dimensions: production quality (PQ), production complexity (PC), content enjoyment (CE), content usefulness (CU)
  - Training set: natural speech, music, and sound samples
  - Testing set: TTS, TTA, and TTM samples
- **Track 3: MOS prediction for different sampling frequencies**
  - Training data: listening tests in 16/24/48 kHz
  - Testing data: “mixed” listening test with all 16/24/48 kHz

AMC 2025: results



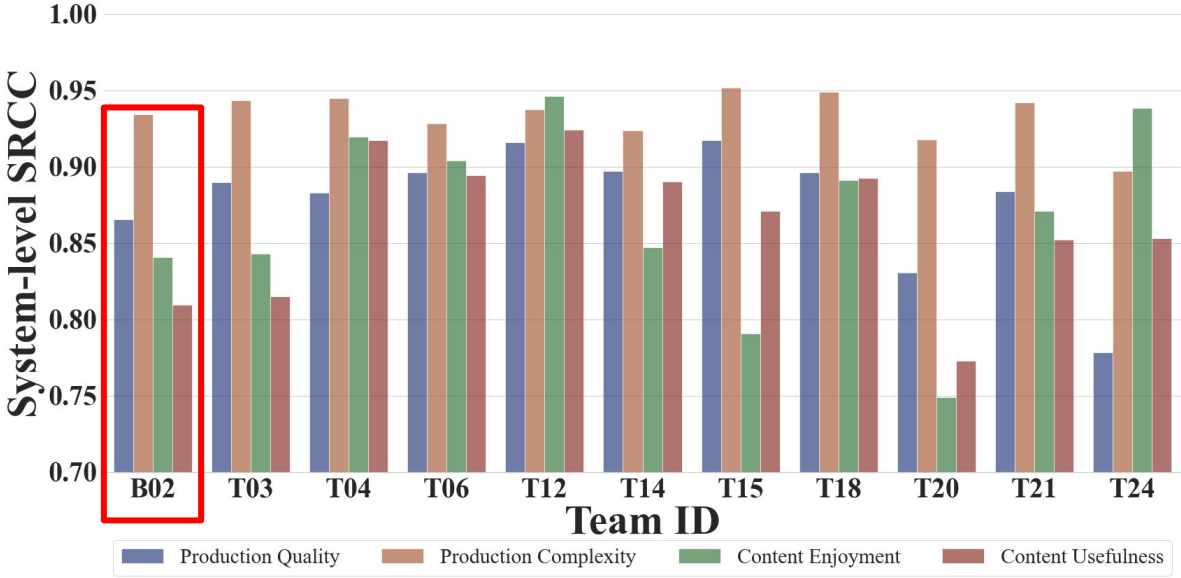


Textual alignment was harder than overall musical quality

Left: track 1; right: track 3.

- **Baseline was outperformed by all teams!**
- **SRCC >0.9 ⇨ probably task setting was too easy...**

AMC 2025: results



Track 2

- **Baseline ranked 9/7/9/10 in PQ/PC/CE/CU ⇨ was outperformed by most teams**
- **Difficult to perform well in all axes**

AMC 2025: feedbacks

- More **speech types**: multilingual speech, expressive TTS, prompt-based TTS, non-verbal speech
- **Multi-dimensional** music evaluation: music rhythm, music theory, music diversity, music style similarity
- **Other audio types**: ambient sound and video-to-audio generation
- Preference score estimation, real-time evaluation...

Key factors to the success of a challenge

Well-defined task

- Ex., “zoomed-in” “different sampling rate” were frequently requested tasks, but task setting is difficult ⇨ **few participants**

User-friendly baseline

- Ex., in VMC 2022 and AMC 2025, we dedicated more effort to developing comprehensive and easy-to-use baseline implementations ⇨ lowered the entry barrier for new participants ⇨ more participants

Marketing and advertisement (most critical!)

- Ex., in VMC 2023 and 2024, we did not actively promote the challenge through mailing lists and social media ⇨ contributed to lower participation

References

[VMC'22] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in Proc. Interspeech, 2022, pp. 4536–4540.

[VMC'23] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2023: Zero-Shot Subjective Speech Quality Prediction for Multiple Domains," in Proc. ASRU, 2023, pp. 1–7.

[VMC'24] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.M. Wang, J. Yamagishi, and Y. Tsao, "The voiceMOS challenge 2024: Beyond speech quality prediction," in Proc. SLT, 2024.

[AMC'25] W.-C. Huang, H. Wang, C. Liu, Y.-C. Wu, A. Tjandra, W.-N. Hsu, E. Cooper, Y. Qin, and T. Toda, "The AudioMOS Challenge 2025," to appear in Proc. ASRU, 2025

[BVCC] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" In Proc. 11th ISCA Speech Synthesis Workshop (SSW 11), 2021, pp. 183–188.

[UTMOS] 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.

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

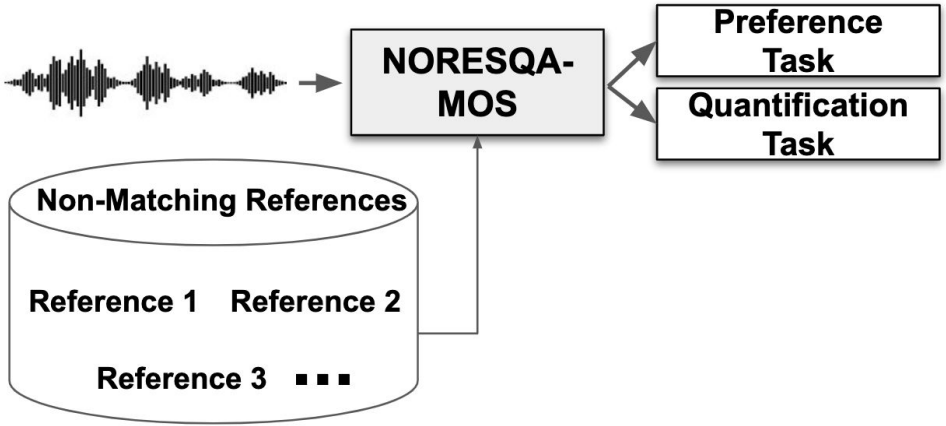
[MusicEval] C. Liu, H. Wang, J. Zhao, S. Zhao, H. Bu, X. Xu, J. Zhou, H. Sun, and Y. Qin, "MusicEval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation," in Proc. ICASSP, 2025.

[Audiobox aesthetics] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu, "Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound," 2025. [Online]. Available: https://arxiv.org/abs/2502.05139

Part II-(2):
Ongoing trends and efforts on quality assessment for **speech**

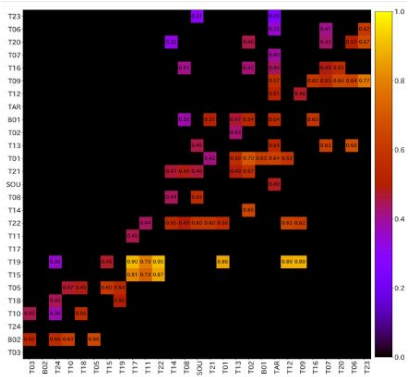
Consider the relative nature of MOS

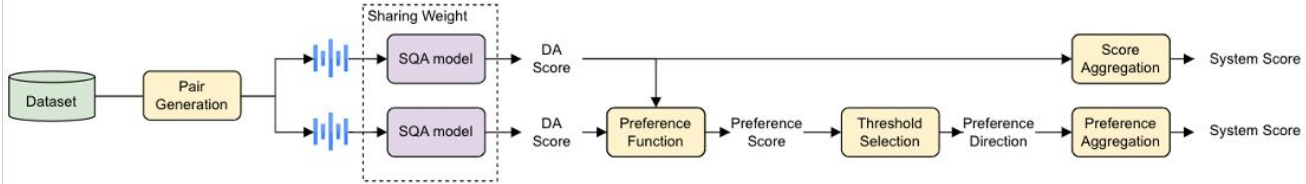
- Most models: trained with L1/L2 loss in the **score space**
- How can we consider the **RANKING** relationship between samples?
- A representative approach: [NORESQA] (NOn-matching REference SQA)
  - The model predicts a **RELATIVE** quality score of the input speech w.r.t. a reference sample
  - In NORESQA, the reference can be (1) another sample in the same batch (2) a clean sample with MOS of 5
  - Methods like UTMOS, [NOMAD], [SCOREQ] extended this idea



Promoting the use of preference tests

- Even leveraging the relative nature, MOS tests still have many flaws...
- Preference tests allow us to compare scores across listening tests!
  - The amount of samples needed is also fewer
- Preference test can be costly... but can be speeded up with online learning [pref]
  - Automatically stops comparing systems that are obviously different in quality
- Learning from preference data led to better generalization ability [E2EPref]





Beyond quality: similarity

- Mainstream: cosine similarity between speaker embeddings (ex., x-vectors)
- Few have attempted to learn from subjective speaker similarity data
  - Dataset: [VoxSim] (derived from VoxCeleb; 41k utterance pairs, nearly 70k ratings)
  - Model: [SVSNet]

Model	LCC ↑	SRCC ↑
ECAPA-TDNN		
pt. speaker recogniser	0.768	0.758
↳ ft. on individual scores	0.827 ±0.002	0.824 ±0.004
↳ ft. on average scores	**0.829** ±0.001	**0.828** ±0.001
WavLM-ECAPA		
pt. speaker recogniser	0.752	0.736
↳ ft. on individual scores	0.833 ±0.001	0.835 ±0.000
↳ ft. on average scores	**0.835** ±0.002	**0.836** ±0.001
SVSNet		
train on individual scores	**0.758** ±0.001	**0.753** ±0.002
train on average scores	0.747 ±0.006	0.742 ±0.006

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**
- **...**

Beyond quality: one model, multiple dimensions

- Why do we want multiple dimensions?
  - Reason 1: interpretability (a single score does not give much insights)
  - Reason 2: increase annotator confidence (quality is subjective, not well-defined ⇨ noisy)
- Ex. 1: [NISQA]
  - Outputs: noisiness, coloration, discontinuity, loudness
- Ex. 2: audiobox aesthetics
  - Outputs: production quality (PQ), production complexity (PC), content enjoyment (CE), content usefulness (CU)

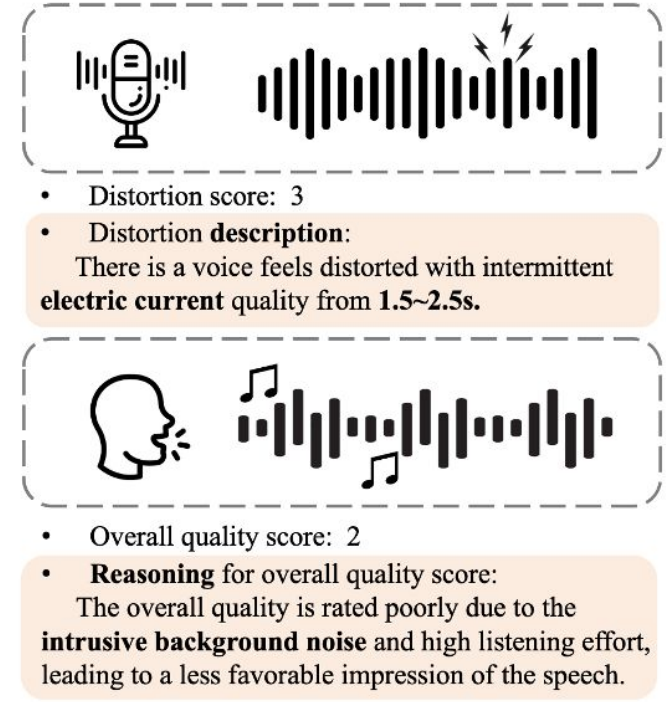
It would be ideal but difficult to have the dimensions orthogonal to each other...

It is however difficult to make sure all annotators use a common scale to make the judgments (ex., Meta "trained" their annotators)

Beyond quality: interpretable/explainable SQA

IMO: the ultimate goal

- 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!
- Problem 1: dataset scaricity
- Problem 2: evaluation
  - Natural language description
    = larger variance compared to scores



[QualiSpeech, LLM-SQA, ALLD]

References

[NORESQA P. Manocha, B. Xu, and A. Kumar, “NORESQA: A framework for speech quality assessment using non-matching references,” in Proc. NeruIPS, 2021, pp. 22363–22378.

[NOMAD} A. Ragano, J. Skoglund and A. Hines, "NOMAD: Unsupervised Learning of Perceptual Embeddings For Speech Enhancement and Non-Matching Reference Audio Quality Assessment," in Proc. ICASSP, 2024, pp. 1011-1015.

[SCOREQ] A. Ragano, J. Skoglund, and A. Hines. "SCOREQ: Speech quality assessment with contrastive regression," in Proc. NeruIPS, 2024, pp. 105702-105729.

[Pref] 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).

[E2EPref] 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

[VoxSim] 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.

[SVSNet] 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.

[NISQA] 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 Proc. Interspeech, 2021, pp. 2127–2131.

[QualiSpeech] 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.

[LLM-SQA] 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

[ALLD] 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

Part III:
Quality Assessment for Music and General Audio
Application of Quality Assessment Metrics

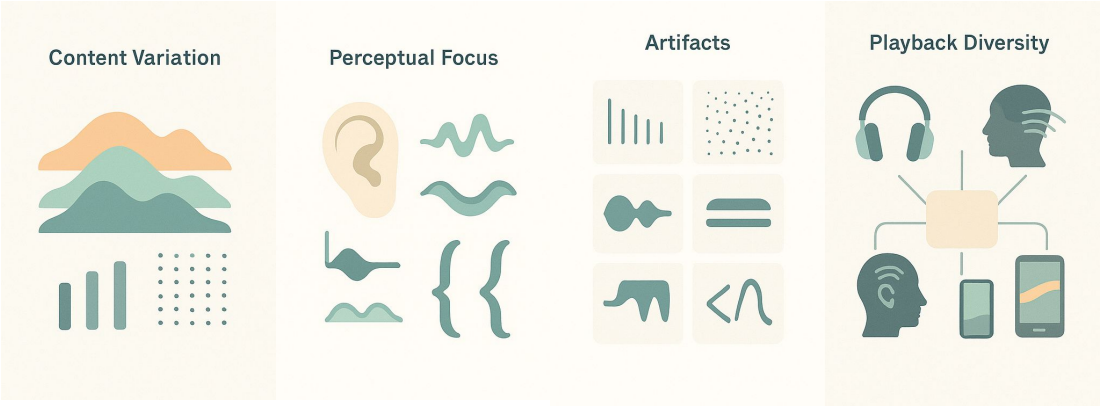
Outline (Part III)

- Quality Assessment for Music and General Audio
  - Why Music/General Audio Quality Assessment is Different
  - Subjective Foundations for Music and General Audio
  - Objective Quality Assessment for Music and General Audio
  - Future Directions of Quality Assessment for Music and General Audio
- Applications of Quality Assessment Metrics
  - Design and Development
  - Runtime Monitoring / Data Selection
  - User-Centric Personalization
  - Optimization and Learning

Part III-(1):
Quality Assessment for **Music** and **General Audio**

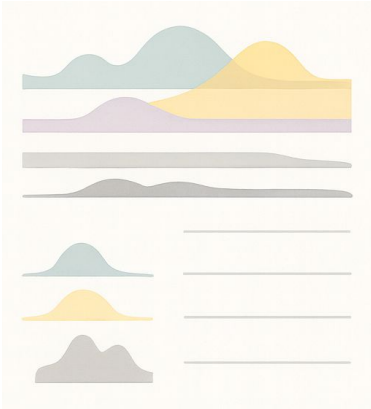
Why Music/General Audio Quality Assessment is Different

- Content variability
- Perceptual focus
- Artifacts
- Playback diversity



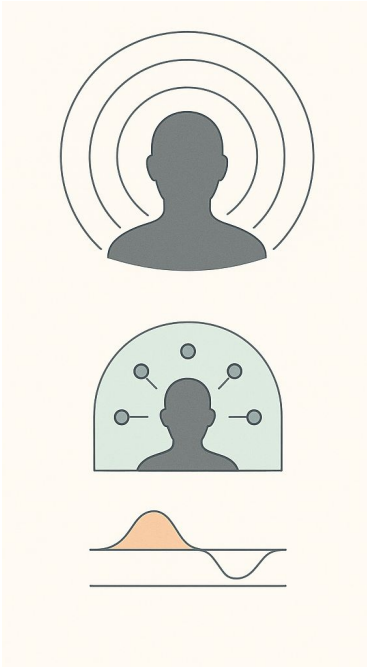
Why Music/General Audio Quality Assessment is Different (Content)

- Content variability
  - Polyphony scenarios
    - Many simultaneous sources: **voice** versus **a mix with vocals/guitars/bass/drums/synth pads**
    - Psychoacoustic masking -> distortions can be **masked** by adjacent frequency bands [psychoacoustic masking]
  - Wide dynamic range
    - **Volume shifts** happen a lot
  - Dense ambiences
    - Broadband, pseudo-random textures hide tonal artifacts, but reveal **pumping** and **spectral breathing** from noise suppression etc.
  - Short alerts/UX sounds (alarms, notifications)
    - Extremely short; **any latency, onset smear, or overshoot** reduces detectability and perceived sharpness
  - ...



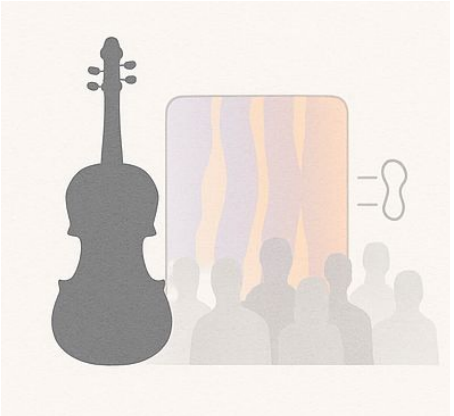
Why Music/General Audio Quality Assessment is Different (Perceptual)

- Perceptual Focus
    - Timbre fidelity
        - Piano vs. bass
            - (attach brightness without metallic ringing versus weight without mud)
        - **Wider space** compared to speaker differences
    - Spatial envelopment and localization
        - Hall width in orchestral recordings or a game scene with discrete sources around listeners
        - Common failure -> **collapsed width after downmix**



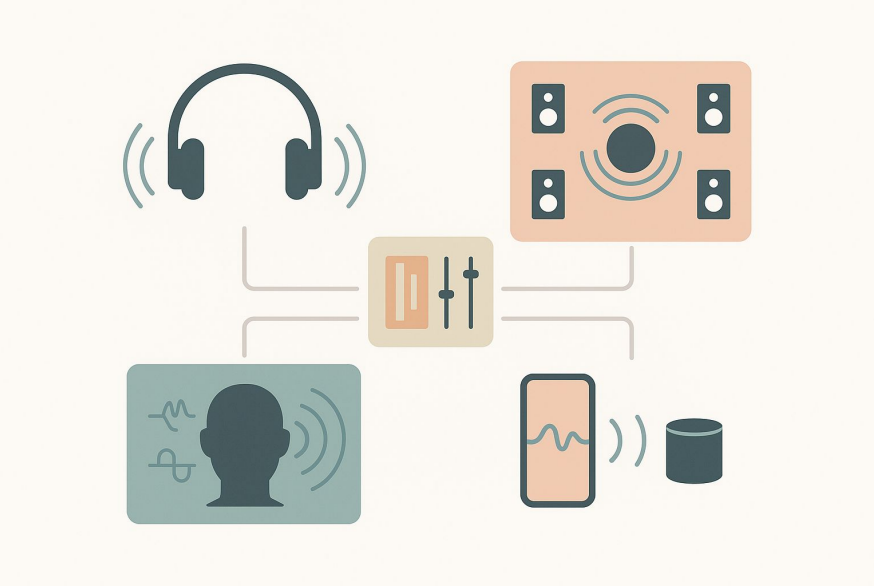
Why Music/General Audio Quality Assessment is Different (Artifacts)

- Artifacts
  - Different types of artifacts
    - Codec pre-echo / denoiser musical noise / clipping / loop seams / spatial fold-down errors etc.
  - How to define “**noise**” in music/general audio evaluation is difficult!



Why Music/General Audio Quality Assessment is Different (Playback)

- Playback diversity
  - Same processing can be judged differently on **different playback chains**
  - Headphones
  - Multichannel speakers
  - Binaural / HRTF renderers
  - Small devices (e.g., phones / smart speakers)

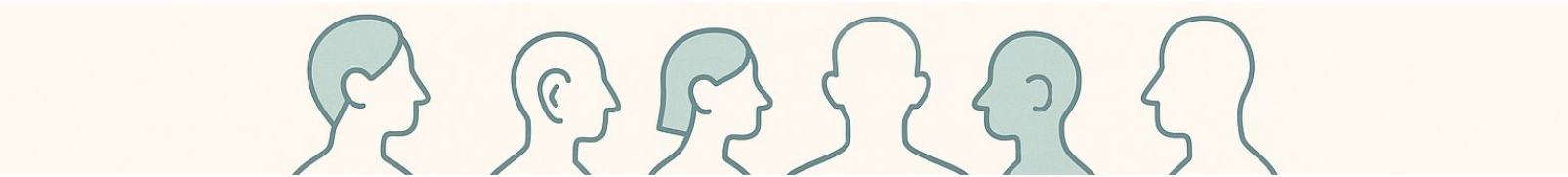


Subjective Foundations for Music and General Audio

- **Perceptual quality ≠ signal-level fidelity**; human listeners integrate complex, context-dependent cues.

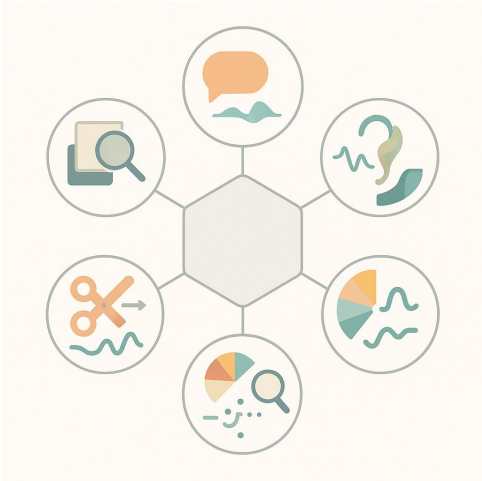
- Music and general audio have different content statistics and listener expectations than speech, so subjective evaluation must account for those.
  - speech listeners prioritize intelligibility and naturalness, music listeners prioritize timbre and musical coherence, general audio listeners may value event salience or plausibility.

- Stil, the **only golden standard** for evaluation



Subjective Foundations for Music and General Audio

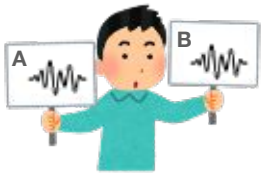
- Base standard with explicit audio evaluation -> MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test
- More recent -> Multi-factored analyses of audio signals
    - Step1: **Scheme setup**
        - Semantic / prompt alignment [MusicGen]
        - Perceptual quality / naturalness [MusicGen]
        - Creativity / diversity for music [CFG Distillation]
        - Concept fidelity (style / mood adherence) [JEN-1]
        - Edit consistency for controllable edits [InstructME]
        - Explainability / interpretability [PAGURI]
    - Step2: **Evaluation formats + Calibration**
        - Pairwise comparison test
        - MOS test
        - MUSHRA test



Same statistical backbone (panels, CI analysis) but different stimuli ranges and scoring anchors

Subjective Foundations for Music and General Audio

- New Protocol in Evaluation (Arena-style evaluation)
  - Multiple systems generate candidates in shared context for direct comparative judgment.
  - Use of **pairwise**-sytle preference aggregation to derive ranking.
  - Advantages:
    - efficient relative scoring
    - exposes subtle quality differences
    - supports adaptive sampling.

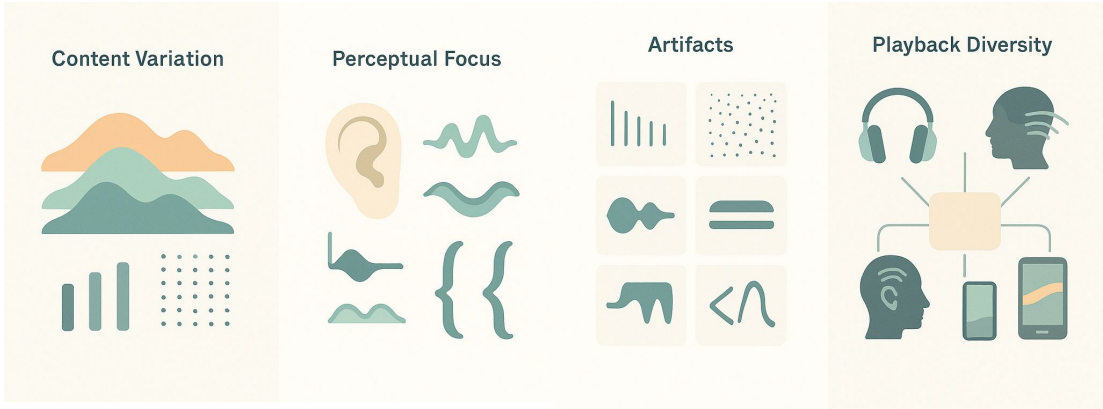


  - Design elements: hidden anchors, expert + crowd panels, real-time selection of hard comparisons.
- Recent music arena [MusicArena]
  - https://beta.music-arena.org/
  - https://huggingface.co/spaces/ArtificialAnalysis/Music-Arena-Leaderboard

Objective Quality Assessment for Music and General Audio

- Same Motivation as speech:
  - Listening tests are a **time-consuming** evaluation paradigm. Reducing the listening effort and time burden can **speed up** experimentation cycles and also make listening tests **more efficient.**

- Challenges compared to speech:
  - Content variability
  - Perceptual focus
  - Artifacts
  - Playback diversity



Objective Quality Assessment for Music and General Audio

Type of reference data available:

- **Matched reference audio sample** AKA intrusive; matching; double-ended
- **Reference text** e.g., audio prompt caption
- **Non-matched reference audio** e.g., distribution modeling
- **No reference data** AKA non-intrusive; single-ended; reference-free

Evaluation type:

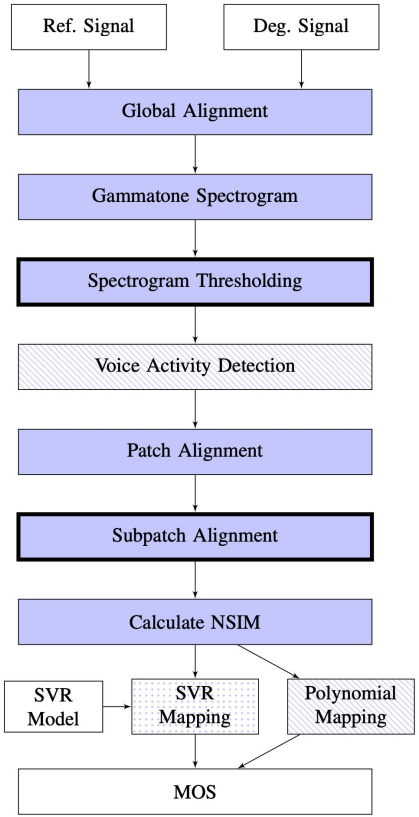
- MOS prediction / quality rating
- A/B pairwise comparison
- Similarity to a reference
- Multi-dimensional evaluation
- .....

Same as SQA!

Objective Quality Assessment for Music and General Audio

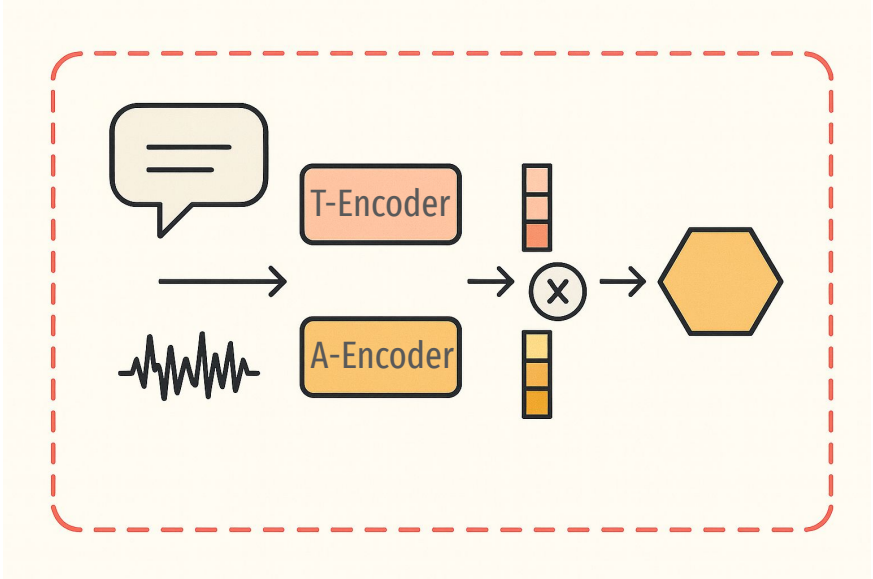
- Quality rating with matched reference audio sample

- Examples
  - Perceptual Evaluation of Audio Quality [PEAQ]
    - An Audio Version of PESQ
  - Virtual Speech Quality Objective Listener (VISQOL)
    - V1 -> neurogram similarity index measure (NSIM) to MOS scores from a set of spectral features [VISQOLv1]
    - V2 -> a support vector regressor over NSIM features [VISQOLv2]
    - V3 -> extended data for VISQOL training [VISQOLv3] (right figure)



Objective Quality Assessment for Music and General Audio

- Quality rating with reference text
- Main setup: Text encoder <-> Audio encoder
- Examples
  - [CLAP Score]
  - [PAM]
  - Refined CLAP with human subjective ratings
    - [MusicEval]
    - [Human-CLAP]



Objective Quality Assessment for Music and General Audio

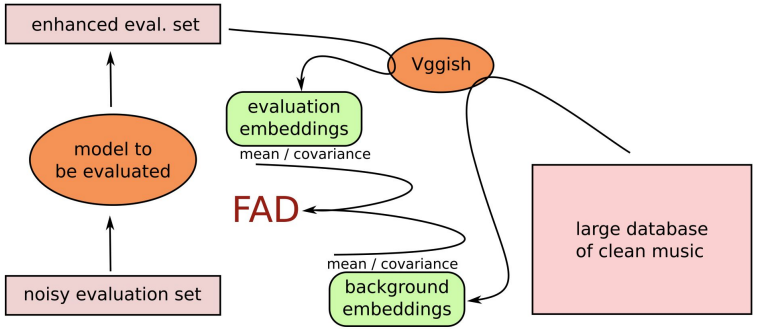
- Quality rating with no reference data

- Examples
  - Analogy to speech MOS score in music domain
    - [MusicEval]
    - [SingMOS]

Objective Quality Assessment for Music and General Audio

- Similarity of references from non-matched reference audio
  - Compare the distribution of representations between synthesized and reference data

- Examples
  - Fréchet Audio Distance [FAD]
  - Kernel Audio Distance [KAD]
    - An efficient substitution of FAD with Hilbert Space
  - MAUVE Audio Distance [MAD]
    - Meta evaluation with synthesized data

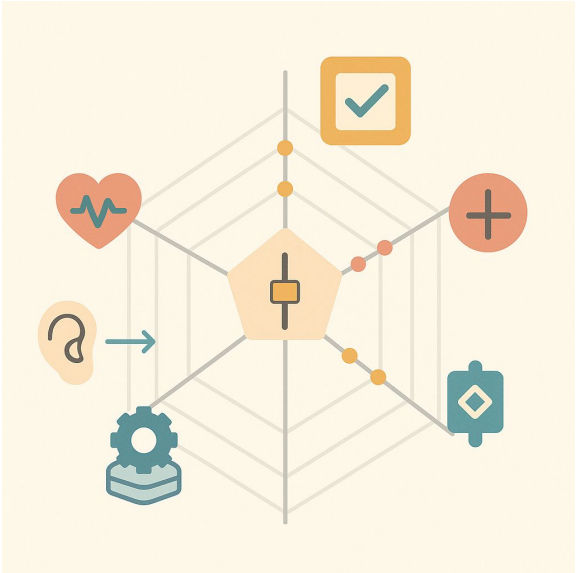


[FAD] Illustration of FAD Calculation Between Synthesized (enhanced) and Noisy Data

Objective Quality Assessment for Music and General Audio

- Multi-dimensional evaluation with no reference audio

- Examples
  - [Audiobox aesthetics]
  - [SongEval] (aesthetics for song)
    - Notable for songs with longer context



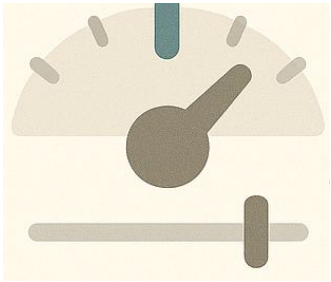
Metric Design for Music and General Audio (Difference from Speech)

Dimensions	Speech	Music and General Audio
Feature Representation	Phonetic clarity, prosody, speaker	Polyphony, timbre richness
Reference Types	Matched lexical or clean reference in common	Often no clear reference, use distributional or prompt-based
Artifact Sensitivity	Voice formant distortion, temporal glitches	Masking effects, genre-dependent noise tolerance, spatial collapse
Scoring and Dimensions	Mostly single-score MOS prediction, utterance-based	Usually multi-criteria, distributional-based

Future Directions of Quality Assessment for Music and General Audio

- Main theme: Maximize subjectiveness from objectiveness

- Detailed directions
  - **Multi-modal integration:** integrating to video context, music score, motion, long text
  - **Robustness to content diversity:** stable evaluation across polyphony, genre shifts (inclusive for genre types), and diverse acoustic environments (e.g., spatial scenario)
  - **Explainable & diagnostic evaluation:** generate interpretable quality breakdowns (localization + multi-dimension break down, especially for long audio/music)



References

- [psychoacoustic masking] Painter, Ted et al. "Perceptual coding of digital audio." Proceedings of the IEEE, 2022.
- [MusicGen] Copet, Jade, et al. "Simple and controllable music generation." NeurIPS, 2023.
- [CFG Distillation] Cideron, Geoffrey, et al. "Diversity-Rewarded CFG Distillation." ICLR, 2025.
- [JEN-1] Li, Peike Patrick, et al. "JEN-1: Text-Guided Universal Music Generation with Omnidirectional Diffusion Models." IEEE CAI, 2024.
- [InstructME] Han, Bing, et al. "InstructME: An Instruction Guided Music Edit Framework with Latent Diffusion Models." IJCAI, 2024.
- [PAGURI] Ronchini, Francesca, et al. "PAGURI: A User Experience Study of Creative Interaction with Text-to-Music Models." ISMIR, 2024.
- [CLAP Score] Xiao, Feiyang, et al. "A Reference-Free Metric for Language-Queried Audio Source Separation using Contrastive Language-Audio Pretraining." DCASE Workshop, 2024.
- [MusicEval] Liu, Cheng, et al. "Musiceval: A Generative Music Dataset with Expert Ratings for Automatic Text-to-Music Evaluation." ICASSP, 2025.

Reference (Cont' d)

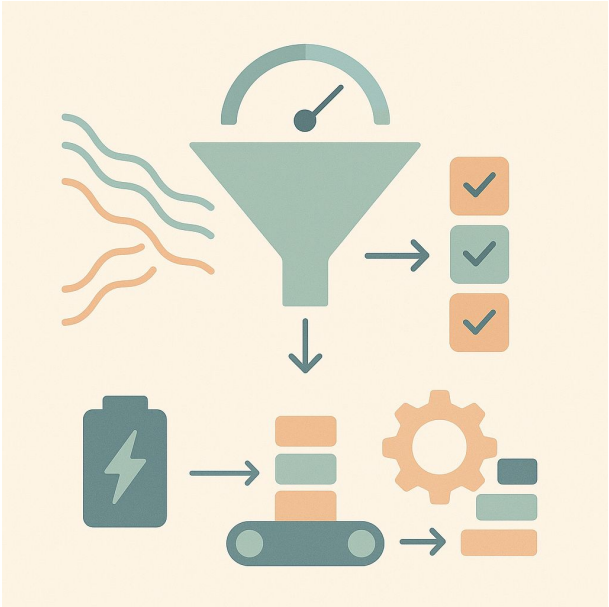
- [Human-CLAP] Takano, Taisei, et al. "Human-CLAP: Human-Perception-Based Contrastive Language-Audio Pretraining." arXiv, 2025.
- [Music Arena] Kim, Yonghyun, et al. "Music Arena: Live Evaluation for Text-to-Music." arXiv, 2025.
- [PEAQ] International Telecommunication Union — Radiocommunication Sector. *"Perceptual Evaluation of Audio Quality (ITU-R Recommendation BS.1387-1)."* ITU-R BS.1387-1, 2001 (Updated version of BS.1387-2, 2023).
- [ViSQOLv3] Chinen, Michael, et al. "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric." QoMEX, 2020.
- [ViSQOLv1] Hines, Andrew, et al. "ViSQOL: An Objective Speech Quality Model." EURASIP Journal on Audio, Speech, and Music Processing, 2015.
- [ViSQOLv2] Hines, Andrew, et al. "Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio." IEEE Transactions on Broadcasting, 2017.
- [SingMOS] Tang, Yuxun, et al. "SingMOS: An Extensive Open-Source Singing Voice Dataset for MOS Prediction." arXiv, 2024.
- [PAM] Deshmukh, Soham, et al. "PAM: Prompting Audio-Language Models for Audio Quality Assessment." Interspeech, 2024.

Part III-(2): Application of Quality Assessment Metrics

Recall the initial motivation

Listening tests are a **time-consuming** evaluation paradigm. Reducing the listening effort and time burden can **speed up** experimentation cycles and also make listening tests **more efficient.**

Can we further extend the assessment model for additional purposes?



Straightforward Application -> Design and Development

- Basic goal: select or improve algorithms

- More and more recent studies start to use objective speech quality assessment for their model comparison:
  - Notable usages:
    - PESQ, STOI, DNSMOS for speech enhancement applications
    - UTMOS, speaker embedding similarity for TTS or VC applications

Everyone still has the consensus that:
Human subjective evaluation is the golden standard for speech/audio quality assessment

Citation plot of [STOI]

Citation plot of [UTMOS]

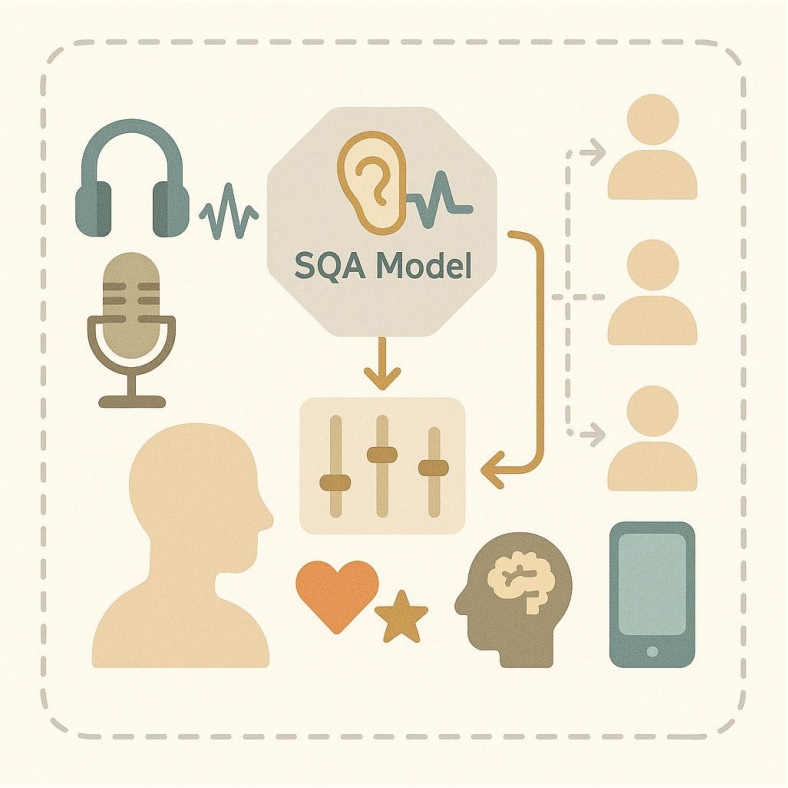
Straightforward Application -> Runtime Monitoring / Data Selection

- Basic goal: observe quality for monitoring / data selection

- Industrial Examples:
  - Cisco's VoIP monitoring of MOS estimation [ThousandEyes-Monitor]
  - Audiocodes' voice call quality monitoring [Audiocodes-Monitor]
  - dotcom-monitor's VoIP Monitoring Tools [Dotcom-Monitor]
- Academic Examples
  - Evaluation-in-the-loop data selection [Dark-Data TTS]
  - URGENT challenge data preparation [URGENT Challenge]
  - [Emilia] dataset collection in the wild

Emerging Usage: User-centric Personalization

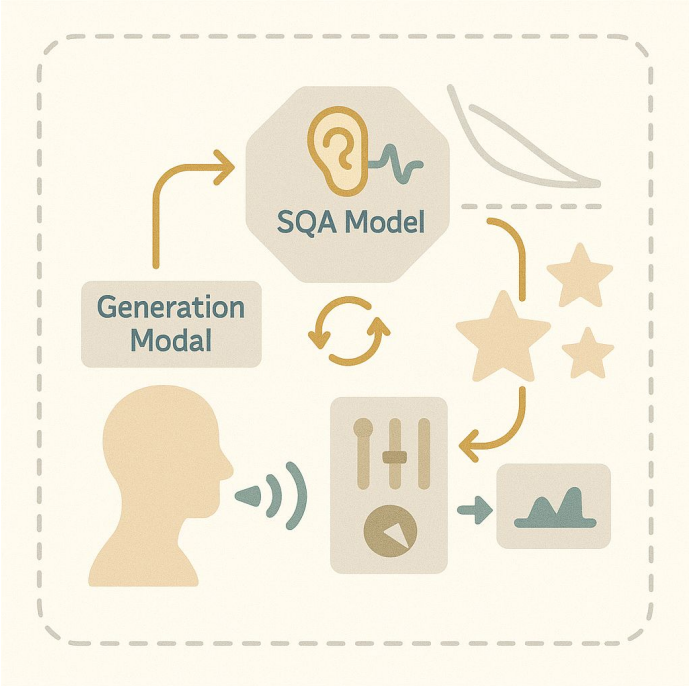
- Basic goal: tailor to individual needs
- Example usage:
  - Hearing-impaired Listeners
    - Speech intelligibility and quality for different hearing condition [HASA-Net+]
  - Speaker-aware SQA [EMDSQA]



Emerging Usage: Optimization and Learning

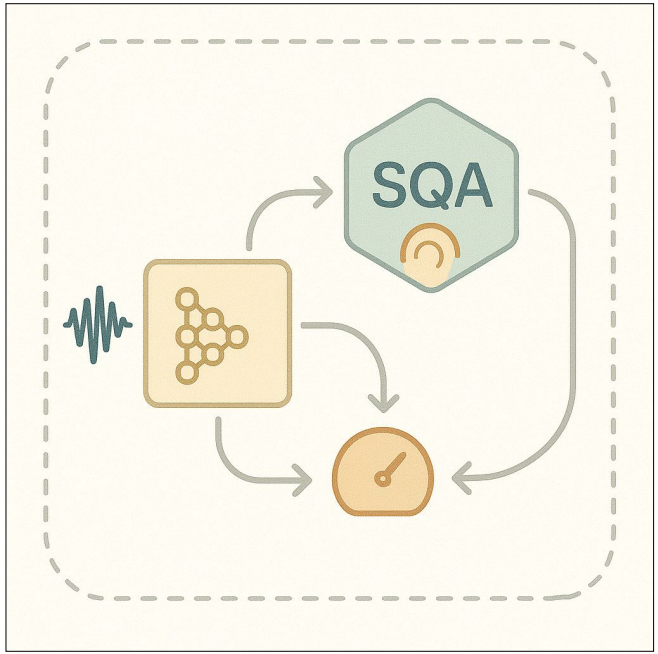
- Basic goal: drive the model toward perceptual quality

- Methodologies
  - Direct loss term
  - Reward modeling / RL



Emerging Usage: Optimization and Learning - Direct Loss Term

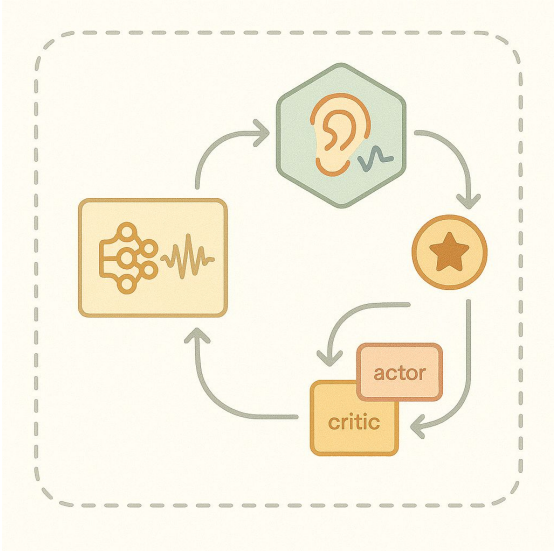
- A differentiable proxy of an SQA metric integrated into the loss function
- Earlier work:
  - Use MOS prediction as loss term [Perceptual-guided TTS]
- Recent works
  - Use a DNSMOS estimator for enhancement [UDASE-CMGAN++]
  - Use UniVERSA-Ext (a multi-metric estimator) for enhancement [Multi-Metric SQA-SE]



Emerging Usage: Optimization and Learning - Reword modeling /RL

- SQA models serve as reward functions in RL setups.

- Examples
  - UTMOS + DPO [DPO-TTS]
  - MOSnet + PPO [UNO]



Reference

- [STOI] Taal, Cees H., et al. "Short-Time Objective Intelligibility (STOI): A Speech Intelligibility Metric." Interspeech (2010). Widely used to assess speech intelligibility by comparing processed audio against a clean reference and demonstrating strong correlation with human perception.
- [UTMOS] Saeki, Takaaki, et al. "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022." Interspeech, 2022. A neural-based MOS (Mean Opinion Score) prediction model using ensemble learning of strong (SSL-fine-tuned) and weak learners, achieving top performance in both in-domain and out-of-domain tracks.
- [Dotcom-Monitor] Dotcom-Monitor. "VOIP Monitoring." Dotcom-Monitor, https://www.dotcom-monitor.com/products/voip-monitoring.
- [ThousandEyes-Monitor] ThousandEyes. "Monitoring VOIP & RTP in the Enterprise WAN." ThousandEyes, https://www.thousandeyes.com/blog/monitoring-voip-rtp-enterprise-wan.
- [AudioCodes-Monitor] AudioCodes. "Management and Monitoring – Contact Center Innovation." AudioCodes, https://www.audiocodes.com/solutions-products/solutions/contact-center-innovation/management-and-monitoring.
- [HASA-Net+] Chiang, Hsin-Tien, et al. "Multi-objective Non-intrusive Hearing-aid Speech Assessment Model." The Journal of the Acoustical Society of America, vol. 156, no. 5, 2024, pp. 3574–3587.
- [EMDSQA] Hao, Yiya, et al. "EMDSQA: A neural speech quality assessment model with speaker embedding." IEEE Signal Processing Letters, 2024

Reference (Cont' D)

- [Dark-Data TTS] Seki, Kentaro, et al. "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection." ICASSP, 2023
- [URGENT Challenge] Wangyou Zhang, et al. "URGENT Challenge: Universality, Robustness, and Generalizability for Speech Enhancement." Interspeech 2024, 4868–4872.
- [Emilia] He, Haorui, et al. "Emilia: A Large-Scale, Extensive, Multilingual, and Diverse Dataset for Speech Generation." arXiv, 2025.
- [Perceptual-Guided TTS] Choi, Yeunju, Youngmoon Jung, Youngjoo Suh, and Hoirin Kim. "Learning to Maximize Speech Quality Directly Using MOS Prediction for Neural Text-to-Speech." IEEE Access, 2022.
- [UDASE-CMGAN++] Close, George, William Ravenscroft, Thomas Hain, and Stefan Goetze. "The University of Sheffield CHiME-7 UDASE Challenge Speech Enhancement System." Proceedings of the 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023), 2023.
- [Multi-Metric SQA-SE] Wang, Wei, et al. "Improving Speech Enhancement with Multi-Metric Supervision from Learned Quality Assessment." arXiv, 2025.
- [UNO] Chen, Chen, et al. "Enhancing Zero-shot Text-to-Speech Synthesis with Human Feedback." ArXiv, 2025.
- [DPO-TTS] Tian, Jinchuan, et al. "Enhancing LM-Based Text-to-Speech with Preference Alignment." ICASSP, 2025.

Part IV:
Resources: Datasets, Benchmarks, Toolkits

MOS-Bench

- A comprehensive benchmark that focuses on the generalization ability of SQA models
- Currently: **7** Train&dev datasets, **12** testing sets (still growing!)
- Multiple domains:
  - Synthetic speech: TTS, VC, SVS, SVC, etc.
  - Distorted speech: noise, reverb, VoIP, transmission, codec, replay, etc.
- Multilingual: English, Chinese, Japanese, France, Dutch
- Multiple sampling rates: 8000 ~ 48000 Hz
- Some training sets come with listener labels

MOS-Bench: train&dev sets

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

MOS-Bench: testing sets (1)

Name	Speech type	Language	FS (kHz)	# samples
BVCC test	TTS, VC, natural speech	English	16	1066
SOMOS test	TTS, natural speech	English	24	3000
NISQA TEST FOR	artificial distorted speech, VoIP	English	48	240
NISQA TEST P501	artificial distorted speech, VoIP	English	48	240
NISQA TEST LIVETALK	real-world distorted speech, VoIP	Dutch	48	232
TMHINT-QI test	artificial noisy speech, enhanced speech, clean speech	Chinese	16	1978

MOS-Bench: testing sets (2)

Name	Speech type	Language	FS (kHz)	# samples
BC19	TTS, natural speech	Chinese	16	540
BC23 Hub	TTS, natural speech	France	22	882
BC23 Spoke	TTS, natural speech	France	22	578
SVCC23	SVC, natural singing voice	English	24	4040
TMHINT-QI(S)	artificial noisy speech, enhanced speech, clean speech	Chinese	16	1960
SingMOS test	SVS, SVC, natural singing voice	Chinese, Japanese	16	645

SHEET

Come for more details on 8/19 (Tue.) 16:00-18:00,
Area5-Oral5 – Speech Quality Assessment

Usage ①: for newcomers of SQA
⇨ All-in-one training & evaluation recipes (data preparation, model training, model evaluation, metric calculation)

Usage ②: for researchers who want to evaluate their SQA model on MOS-Bench
⇨ Easy-to-use scripts to for evaluation only (data preparation, model evaluation, metric calculation)

Usage ③: for researchers who only want to use off-the-shelf models
⇨ User-friendly interfaces via `torch.hub.load` or HuggingFace

Public Available Metrics Hubs - Speech Focus

Name	Task Focus	Example metrics	Metric Count
[ESPnet]	TTS, VC, SE/SS, Codec	MCD, F0-CORR, WER, STOI, UTMOS, SPK-SIM…	22
[SpeechMOS]	SE/SS, PLC	AECMOS, PLCMOS, DNSMOS	3
[ClearerVoice]	SE/SS	DNSMOS, PESQ, MCD, SNR…	14
[Pysepm]	SE/SS	fwSNRseg, LLR, PESQ…	10
[Amphion]	TTS, VC	F0-CORR, Energy-RMSE, SPK-SIM, WER, FAD, …	16

(Recorded at 2025.08)

Public Available Metrics Hubs - Music/Audio Focus

Name	Task Focus	Example metrics	Metric Count
[AudioLDM-Eval]	TTA	FAD, KID, KLD, SSIM, …	9
[Stable-Audio-Metrics]	TTA	FAD, KLD, CLAP score	3
[FADTK]	TTA, TTM	FAD	11
[AudioCraft]	TTA, TTM	VISQOL, FAD, KLD…	6
[SonyCSL-Audio-Metrics]	TTA, TTM	FAD, APA, Density and Coverage, …	4
[AQUA-TK]	TTA	PEAQ, FAD, KID, …	9
[Amphion]	TTA, TTM	FAD	1

(Recorded at 2025.08)

VERSA: Born from Challenges

- Plenty of difficulties exist in evaluation for generated speech/audio quality:
    - Collecting subjective evaluation is not easy
    - Implementing objective evaluation is not easy
    - Increasing need to support multi-domain larger-scale evaluation

VERSA

Versatile Evaluation of Speech and Audio

Introducing VERSA

11492/11692/18495
Speech Technology for Conversational AI

- VERSA (Versatile Evaluation for Speech and Audio)
  - Targets a general interface for speech and audio evaluation
  - A collection of conventional/recent automatic quality evaluation metrics
  - Highly integration to toolkits / challenges

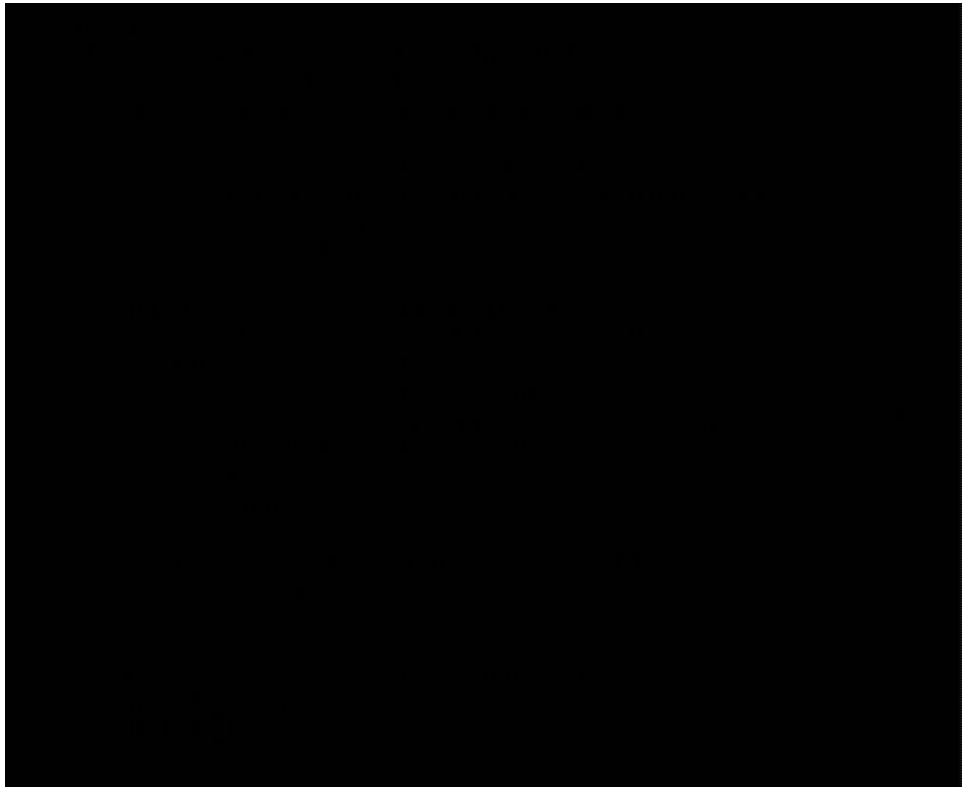
CHiME CHALLENGE

ESPnet

Muskit

VERSA Framework

- Flexible I/O Interface
- Easy Manageable Configuration
- Strict Dependency Control
- Job-scheduling Support
- Continuous Integration Tests
- Efficient Resource Management
  - (cache control)
- Community-driven Effort



VERSA (Cont'D)

VERSA
Versatile Evaluation of Speech and Audio

- Up to 90 metrics in speech and audio evaluation supported currently

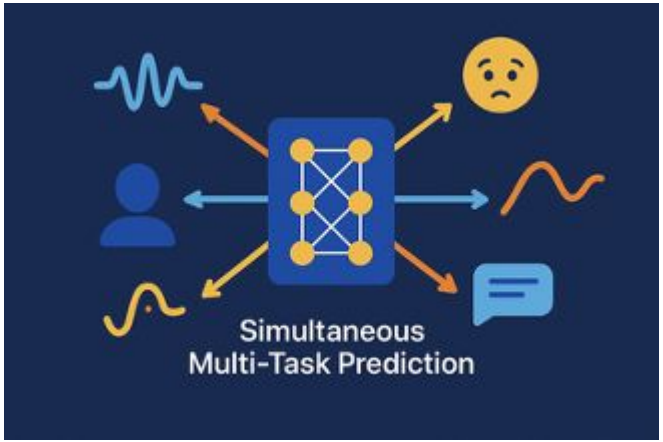
- Community-driven efforts for future applications

Links and Demos

GitHub

Colab Demo

Uni-VERSA: VERSA in a Unified Framework



- Using Uni-VERSA to enjoy up to **109x speed up** over computing same metrics from VERSA

- Using Uni-VERSA as a loss term for generation tasks
  - [Multi-Metric SQA-SE]

- Join the detailed oral presentation by Prof. Watanabe
  - **Area12-Oral3**
    Speech Assessment (Tue 2:10 - 2:30pm)

Colab Demo

Huggingface Pre-trained Models

Stay tuned with VERSA team

We are working further on additional topics on multi-metric speech quality assessment

In the upcoming ASRU2025, we will present

- VERSA-v2 (upcoming at ASRU2025)
- UniVERSA-Ext (upcoming at ASRU2025)

VERSA
Versatile Evaluation
of Speech and Audio

Concluding remarks