KyotoMOS: An Automatic MOS Scoring System for Speech Synthesis

Wangjin Zhou
Kyoto University
Kyoto, Japan
zhou@sap.ist.i.kyoto-u.ac.jp

Zhengdong Yang
Kyoto University
Kyoto, Japan
zd-yang@nlp.ist.i.kyoto-u.ac.jp

Sheng Li
NICT
Kyoto, Japan
sheng.li@nict.go.jp

Chenhui Chu
Kyoto University
Kyoto, Japan
chu@i.kyoto-u.ac.jp

ABSTRACT

The Mean Opinion Score (MOS) serves as a subjective measure for assessing the quality of synthesized speech. Nevertheless, the conventional approach to MOS evaluations can be resource-intensive in terms of both time and cost. This article unveils an automatic MOS scoring toolkit that builds upon our success in securing the top position for some metrics in VoiceMos2022 Challenge and emerging as champions in some tracks of the VoiceMos2023 Challenge. We offer a pre-trained MOS scoring tool for English and provide training code for other languages. Our documentation, examples, and source code are available at https://github.com/superphysics/KyotoMOS.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Automatic MOS Prediction, Synthesis Speech Evaluation, Benchmark

1 INTRODUCTION

Subjective evaluation of speech synthesis involves human listeners assessing the quality of synthesized speech. The Mean Opinion Score (MOS) serves as a commonly employed subjective quality assessment metric for synthesized speech. During MOS evaluations, listeners assign quality ratings to speech samples on a scale spanning from 1 (poor) to 5 (excellent). These ratings' collective average is then employed to assess the overall quality of the speech. Nevertheless, conventional MOS assessments often entail significant costs and consume considerable time resources.

Considering these difficulties, there has been an increasing inclination towards utilizing annotated MOS rating datasets to train automatic MOS prediction models. Many of these studies lean towards using neural network-driven frameworks and harness large volumes of synthetic speech during their training phase [1, 13].

It's worth highlighting that The VoiceMOS Challenge [4] has played a central role in driving progress in this field in recent times. This event coordinators introduced three cutting-edge techniques as benchmark solutions [3, 5]. Additionally, they provided an extensive synthetic dataset sourced from previous Blizzard challenges [6–12]. These benchmark systems predominantly rely on utterance-level MOS scores, utilizing either the original speech [5], domain, or latent attributes [5, 16].

We participated in the VoiceMOS2022 and VoiceMOS2023 Challenges. In VoiceMOS2022, we introduced a fusion technique for MOS scores, harnessing multiple auto self-supervised learning (SSL) based MOS models, and achieved top rankings in specific performance metrics. In VoiceMOS2023, we advanced even further, introducing a MOS scoring baseline that incorporates listener information and conducting experiments to amalgamate a variety of acoustic features. Drawing on our insights from the previous two challenges, we expanded our auto MOS models and released them as a benchmark for the Auto MOS research community.

2 PROPOSED MODEL

We proposed KyotoMOS, an automatic MOS predictor employs a multi-model fusion architecture as shown in Figure 1. Inputs for the fusion are derived from two core systems (SSL-MOS and LE-SSL-MOS), each constructed using distinct SSL models. In fusion research, it has been observed that, in comparison to embedding fusion techniques, KyotoMOS achieves superior results through skillfully designed rate fusion.

SSL-MOS: The basic SSL-MOS system is the baseline of VoiceMOS2022 [4], it is constructed by adding a mean pooling layer and a fully connected layer after the feature extractor of the SSL model. Wav2Vec, Hubert, WavLM, MMS have been used to construct different SSL-MOS models.

LE-SSL-MOS: LE-SSL-MOS is an automatic MOS predictor which utilizes pre-trained SSL models and further improves prediction accuracy by utilizing the opinion scores of each utterance in the listener enhancement branch.

Fusion: We provide two core fusion methodologies: one employing MOS embeddings fusion, and the other utilizing MOS rates fusion. Embeddings Fusion combines embeddings produced by various MOS predictors using a linear model, while Rates Fusion employs a bias-free linear layer to create a voting mechanism.

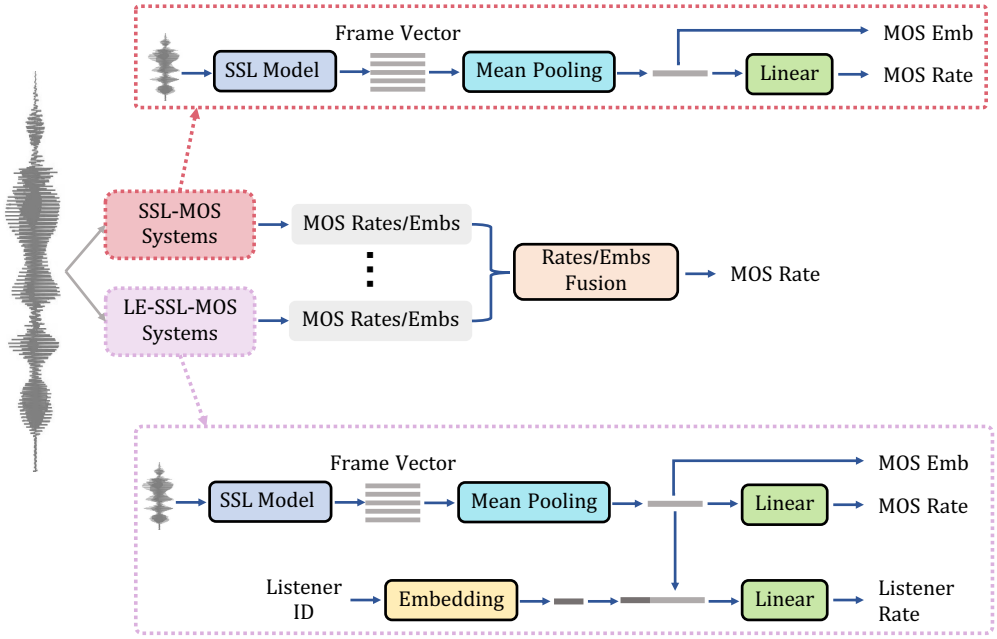

Figure 1: Proposed model structure.

3 SYSTEM DESCRIPTION

KyotoMOS consists of multiple components, including Predictor, Trainer, and several auxiliary tools.

Predictor: We provide pre-trained MOS predictor for your convenience. With the assistance of the scripts we offer, you have the capability to seamlessly retrieve pre-trained checkpoints and effortlessly apply them in your MOS scoring tasks.

Trainer: Users have the flexibility to train their own automatic MOS predictor using our supplied training scripts, whether utilizing a single GPU or multiple GPUs, simply by preparing the data in the desired format. In the case of non-English target languages, we provide two training options. One involves fine-tuning on our pre-trained models, while the other necessitates users to source a pre-trained model in the target language and subsequently re-run our training scripts based on that foundation.

Audio Converter: For this system to work correctly, audio files must be sampled at 16 kHz. In case your audio files have a different sampling rate, we offer a conversion tool. Keep in mind that both bandwidth compression and expansion can influence the MOS rating of the original audio.

Noise Reduction: This system is intended for MOS scoring of clean, noise-free speech. The presence of noise can influence the MOS rating. To address this, we offer an enhancement tool that aims to eliminate noise while preserving the original speech to the greatest extent possible.

Performances of Our System: According to the VoiceMOS2022 official analysis results [4], our system has achieved the highest LCC, SRCC, and KTAU scores at the system level on main track, as well as the best performance on the LCC, SRCC, and KTAU evaluation metrics at the utterance level on OOD track. Compared with the basic SSL models, the prediction accuracy of the fused system has been largely improved, especially on OOD sub-track. More detailed descriptions can be found in our previous work [15].

According to the VoiceMOS2023 official analysis results [2], our system performs better than the baseline. Our fusion system achieved an absolute improvement of 13% over LE-SSL-MOS on the noisy and enhanced speech track. And our system ranked 1st and 2nd respectively in the French speech synthesis track and the noisy and enhanced speech track of the challenge. Our system has the most consistent performance across tracks. More detailed descriptions can be found in our previous work [14].

4 CONCLUSION

Drawing on our extensive experience gained from participating in VoiceMOS Challenges over the years, we have unveiled KyotoMOS, an automatic MOS Predictor powered by SSL models. Moreover, we offer users a comprehensive toolkit for tasks such as inference, training, and beyond. For more detailed documentation and access to the source code, please visit our GitHub repository https://github.com/superphysics/KyotoMOS.

REFERENCES

[1] Yeunju Choi, Youngmoon Jung, and Hoirin Kim. 2021. Neural mos prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 462–469.

[2] E. Cooper and et al. 2023. The VoiceMOS Challenge 2023: zero-shot subjective speech quality prediction for multiple domains. In *Proc. IEEE-ASRU*, Vol. 2023.

[3] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2021. Generalization ability of MOS prediction networks. *arXiv preprint arXiv:2110.02635* (2021).

[4] Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022. The VoiceMOS Challenge 2022. *arXiv preprint arXiv:2203.11389* (2022).

[5] Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. 2021. LDNet: Unified Listener Dependent Modeling in MOS Prediction for Synthetic Speech. *arXiv preprint arXiv:2110.09103* (2021).

[6] Vasilis Karaiskos, Simon King, Robert AJ Clark, and Catherine Mayo. 2008. The blizzard challenge 2008. In *Proc. Blizzard Challenge Workshop, Brisbane, Australia*. Citeseer.

[7] Simon King and Vasilis Karaiskos. 2010. The Blizzard Challenge 2010.

[8] Simon King and Vasilis Karaiskos. 2011. The Blizzard Challenge 2011.

[9] Simon King and Vasilis Karaiskos. 2012. The Blizzard Challenge 2012.

[10] Simon King and Vasilis Karaiskos. 2013. The Blizzard Challenge 2013.

[11] Simon King and Vasilis Karaiskos. 2016. The Blizzard Challenge 2016.

[12] Simon King and Vasilis Karaiskosb. 2009. The blizzard challenge 2009. In *The Blizzard Challenge 2009 Workshop*.

[13] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352* (2019).

[14] Z. Qi, X. Hu, W. Zhou, S. Li, H. Wu, J. Lu, and X. Xu. 2023. LE-SSL-MOS: Self-Supervised Learning MOS Prediction with Listener Enhancement. In *Proc. IEEE-ASRU*, Vol. 2023.

[15] Zhengdong Yang, Wangjin Zhou, Chenhui Chu, Sheng Li, Raj Dabre, Raphael Rubino, and Yi Zhao. [n. d.]. Fusion of Self-supervised Learned Models for MOS Prediction. In *Proc. Interspeech 2022*. 5443–5447. https://doi.org/10.21437/Interspeech.2022-10262

[16] Ryandhimas E Zezario, Szu-Wei Fu, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao. 2021. Deep Learning-based Non-Intrusive Multi-Objective Speech Assessment Model with Cross-Domain Features. *arXiv preprint arXiv:2111.02363* (2021).