ID: P4-28-SS05

The T05 System for The VoiceMOS Challenge 2024: Transfer Learning from Deep Image Classifier to Naturalness MOS Prediction of High-Quality Synthetic Speech

Our codes
https://github.com/sarulab-speech/UTMOSv2

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari (The University of Tokyo, Japan)

Introduction

Automatic MOS Prediction

Machine learning system that predicts Mean Opinion Score (MOS) of synthetic speech (e.g., UTMOS [1])

- ✓ Reducing the costs of human-based subjective evaluations
- ✓ Achieving highly reproducible evaluation
- ✗ Suffering from the bias observed in the training data

RQ: Can we develop a MOS predictor suitable for high-quality synthetic speech?

Our Contributions: The Development of UTMOSv2

MOS predictor designed for comparing high-quality synthetic speech

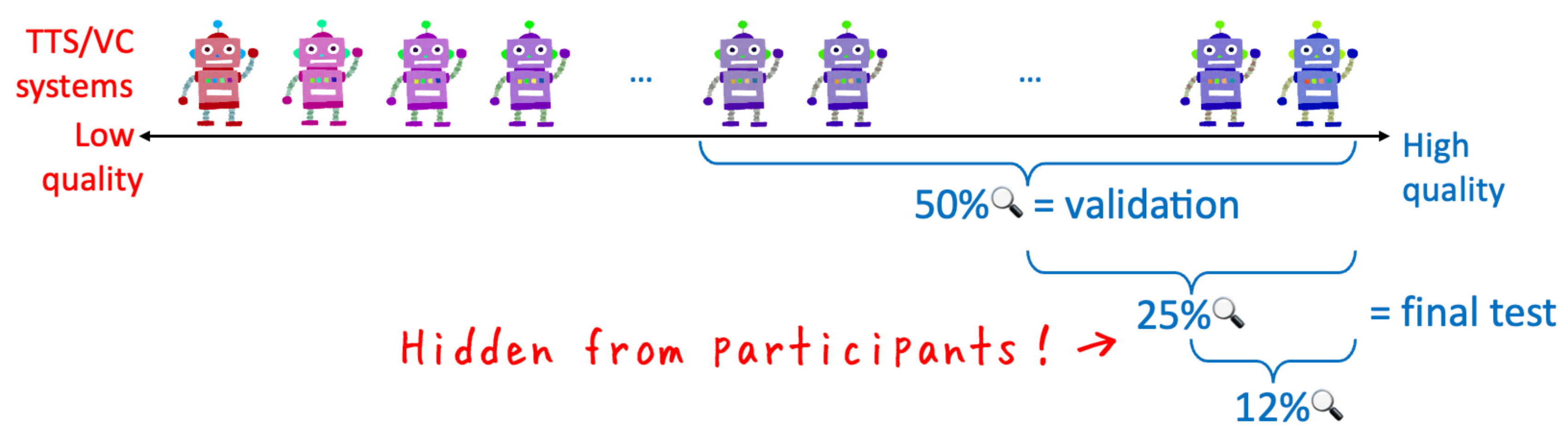
✨ UTMOSv2 achieved 1st place in 7/16 evaluation metrics ✨
✨🏆 and 2nd place in the remaining 9 metrics 🏆✨
✨ in the VoiceMOS Challenge (VMC) 2024 Track 1 [2] ✨

Publicly available on GitHub (scan the QR code above)

Our UTMOSv2 for The VMC 2024 Track 1

The VMC 2024 Track 1

Dataset: Zoomed-in MOS test results of BVCC [3]



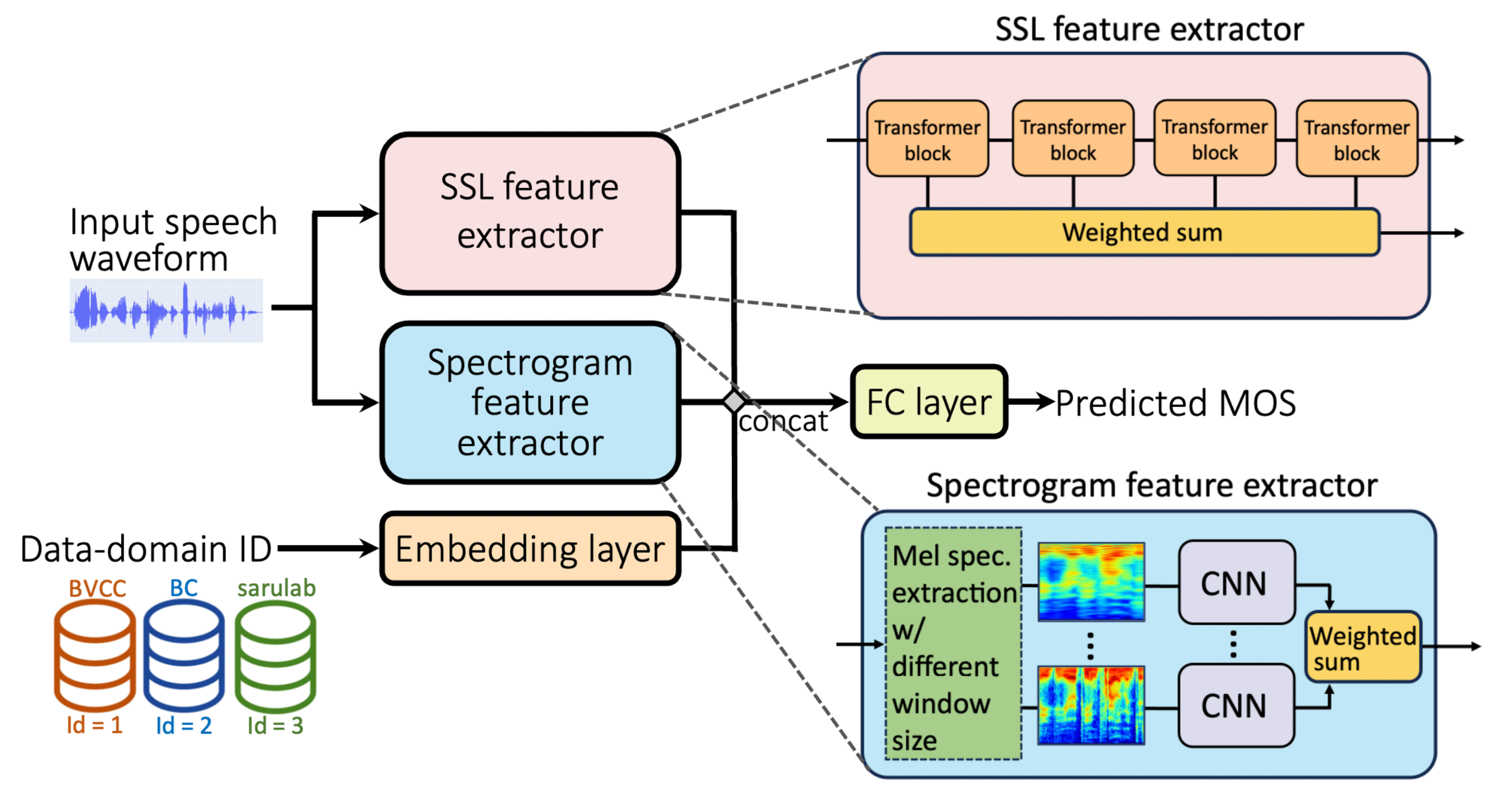
- No official training dataset for zoomed-in MOS tests
  → Our team conducted 50% zoomed-in MOS test using BVCC & published the results as "sarulab-data" (sarulab).

Evaluation Metrics:

- **MSE (Mean Squared Error)**
- **SRCC (Spearman's Rank Correlation Coefficient)**
- LCC (Linear Correlation Coefficient)
- KTAU (Kendall's Tau)

- Evaluation for each utterance (Utterance-level)
- Evaluation for each speech synthesis system (System-level)

UTMOSv2



① Fusion of SSL/spectrogram features → See Exp. 1
- Using pretrained speech SSL model / image classification model as powerful feature extractors for MOS prediction

② Multi-stage learning strategy → See Exp. 2
1) Pretrain each feature extractor independently
2) Train the last FC layer
3) Fine-tune the whole system

③ Data-domain encoding → See Exp. 3
- Condition the MOS predictor on the data domain ID

Experimental Evaluation with 12% Zoomed-in BVCC

Experimental Setup (See our paper for more details)

Traning Dataset	BVCC [3], Blizzard Challenge 2008〜2011 [4,5,6,7], SOMOS [8], sarulab (Mixup [15] was used as the data augmentation)
Feature Extractor	SSL: wav2vec2.0-base [9] (pretrained on LibriSpeech [10]) Spectrogram: EfficientNetV2 [11] (pretrained on ImageNet [12])
Optimizer	AdamW [13] w/ cosine annealing scheduler [14] (The learning rates were tuned for each experiment)
Training Objective	Minimizing contrastive loss [1] + MSE
Checkpoint Selection	5-fold cross-validation based on the average system-level SRCC (The primal metric for the VMC 2024 Track 1)

Exp. 1: Effects of Feature Fusion

	Utterance-level		System-level	
	MSE ↓	SRCC ↑	MSE ↓	SRCC ↑
UTMOSv2	0.459	**0.579**	0.288	**0.854**
w/o SSL	**0.357**	0.516	**0.188**	0.770
w/o Spec.	0.673	0.529	0.497	0.793
SSL-MOS [16]	0.741	0.417	0.589	0.609
UTMOS [1]	0.541	0.300	0.378	0.367

✓ ①The fusion improved SRCC

✓ ②Achieved higher performance than the baselines (SSL-MOS, UTMOS)

Exp. 2: Comparison of Multi-Stage Learning

	Utterance-level		System-level	
	MSE ↓	SRCC ↑	MSE ↓	SRCC ↑
UTMOSv2	0.459	**0.579**	0.288	**0.854**
w/o Stage 1	0.342	0.505	0.108	0.816
w/o Stage 2	**0.293**	0.423	**0.097**	0.672

✓ The multi-stage learning process improved SRCC

【Reference】
[1] Saeki et al., INTERSPEECH 2022, [2] Huang et al., SLT 2024, [3] Huang et al., INTERSPEECH 2022,
[4] Karaiskos et al., BC Workshop 2008, [5] Black et al., BC Workshop 2009,
[6] Black et al., BC Workshop 2010, [7] King et al., BC Workshop 2011,
[8] Maniati et al., INTERSPEECH 2022, [9] Baevski et al., NeurIPS 2020,
[10] Panayotov et al., ICASSP 2015, [11] Tan et al., ICML 2021, [12] Deng et al., CVPR 2009,
[13] Loshchilov et al., ICLR 2019, [14] Loshchilov et al., ICLR 2017, [15] Zhang et al., ICLR 2018,
[16] Cooper et al., ICASSP 2022.

Exp. 3: Investigation on Training Dataset

The data domain ID specified for MOS prediction

	BVCC		BC		SOMOS		sarulab	
System-level	MSE ↓	SRCC ↑	MSE ↓	SRCC ↑	MSE ↓	SRCC ↑	MSE ↓	SRCC ↑
All datasets	0.288	**0.854**	**0.088**	**0.851**	**0.056**	0.844	**0.058**	**0.838**
w/o BVCC	-	-	0.343	0.832	0.128	**0.846**	0.101	0.836
w/o BC	**0.145**	0.819	-	-	0.069	0.823	0.122	0.805
w/o SOMOS	0.224	0.696	0.221	0.682	-	-	0.221	0.700
w/o sarulab	0.282	0.647	0.102	0.661	0.186	0.690	-	-

✓ Training on all datasets generally yielded the best performance

✓ Removing SOMOS or sarulab degraded the performance
→ SOMOS/sarulab datasets were important for the zoomed-in MOS prediction