Generalization Ability of MOS Prediction Networks

Erica Cooper[1], Wen-Chin Huang[2], Tomoki Toda[2], Junichi Yamagishi[1]

[1]National Institute of Informatics, Japan
[2]Nagoya University, Japan

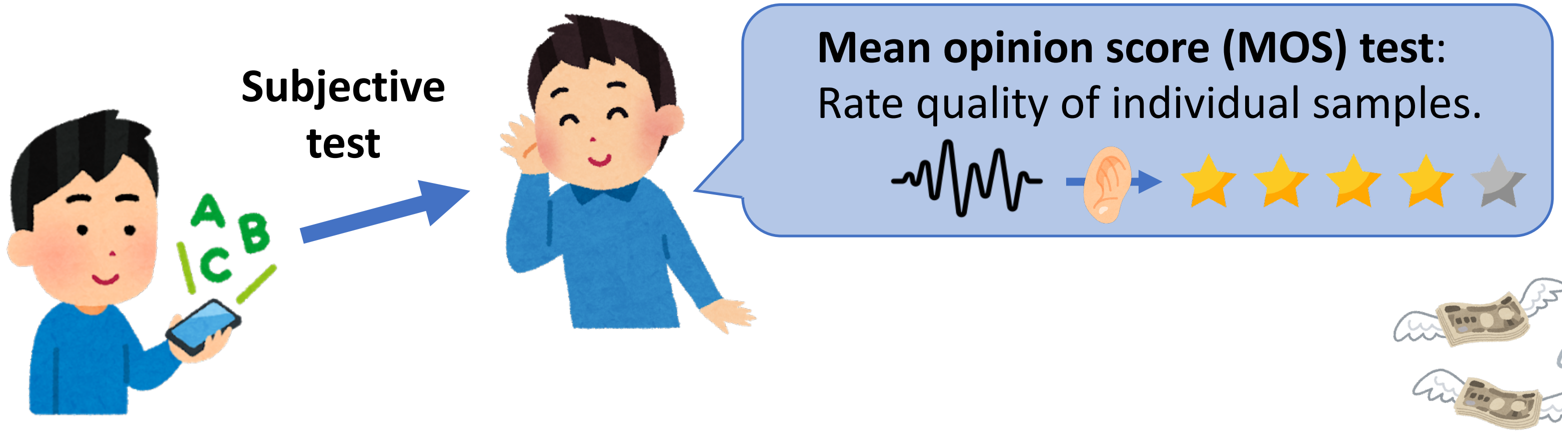
IEEE ICASSP 2022

Outline

- Motivation
- Research questions
- Datasets
- Experiments
- Conclusions

Motivation

Listening tests: evaluate speech synthesis systems, e.g. text-to-speech (TTS), voice conversion (VC).

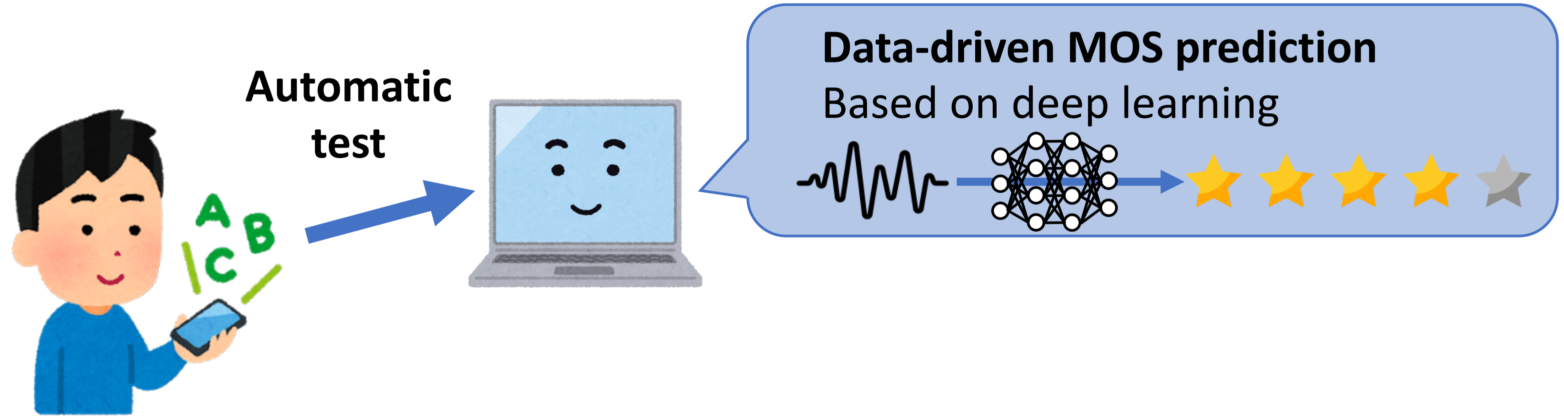


Drawbacks:

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

Motivation

Automatic speech quality assessment: predict human ratings using data.



Drawback: **Bad generalization ability** to unseen systems and listening tests.

Research Questions

- What can help to improve generalization ability?
  - **A large and diverse training dataset?**
  - **Data augmentation?**
  - **Self-supervised learning** (SSL) based speech models?
  - **Fine-tuning** on a small amount of target-domain data?
- What types of **unseen conditions** are the most difficult for MOS prediction?

Datasets

- **BVCC:** English TTS and VC systems spanning over a decade, collected from past Blizzard Challenges, Voice Conversion Challenges, and ESPnet-TTS. **Main (in-domain) dataset.**
- **Out-of-domain (OOD) datasets:**
  - **ASV19:** English samples from a variety of state-of-the-art TTS and VC systems for the ASVSpoof Challenge in 2019.
  - **BC19:** Chinese TTS samples from the Blizzard Challenge 2019
  - **COM18:** Japanese TTS combining four different acoustic models and four different vocoders in 2018.

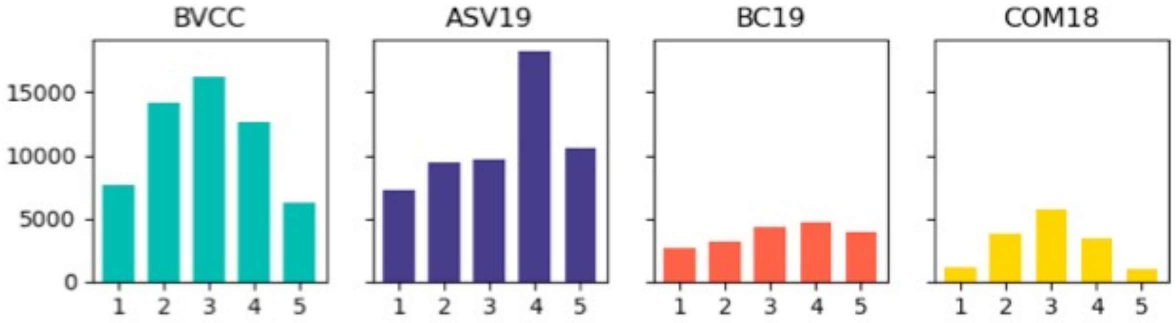


Fig. 1: Distributions of scores for each dataset

Name	samp	ratings per samp	spk	sys
BVCC	7106	8	27	187
ASV2019	18079	1-26	67	14
BC2019	1352	10-17	1	26
COM2018	4760	1-9	1	10

Training/development/testing splits

- Training/development/testing splits were chosen to match the overall distribution of scores as closely as possible for each dataset
- Unseen speakers, systems, listeners, and texts were held out for each development and test set where possible

- BVCC (in-domain): Training/Validation/Test
  70% / 15% / 15%
- OOD: **Fine-tuning**/Validation/Test
  33% / 33% / 33%

Name	unseen spk	unseen sys	unseen listeners	unseen texts
BVCC	1	6	8	5
ASV2019	4	2	10	-
BC2019	-	2	70	2
COM2018	-	1	5	5

Experiments

- MOSNet trained on BVCC
- Fine-tune SSL models for the MOS prediction task using BVCC
- Zero-shot and fine-tuned MOS prediction performance on OOD datasets
- **Evaluation metrics:**
  - Mean squared error (**MSE**): Absolute difference between actual and predicted MOS
  - Linear Correlation Coefficient (**LCC**): Simple correlation
  - Spearman Rank Correlation Coefficient (**SRCC**): Ranking-based correlation
  - Kendall Tau Rank Correlation (**KTAU**): Ranking-based correlation that is more robust to errors

Experiments: MOSNet trained on BVCC

- MOSNet: a CNN-BLSTM architecture for MOS prediction
- Experiments:
  - Zero-shot performance of MOSNet pretrained on VCC2018
  - Train MOSNet from scratch on BVCC
  - Fine-tune pretrained MOSNet on BVCC
  - Fine-tune pretrained MOSNet on BVCC + silence augmentation
  - Fine-tune pretrained MOSNet on BVCC + speed augmentation
  - FIne-tune pretrained MOSNet on BVCC + both augmentations

"MOSNet: Deep Learning based Objective Assessment for Voice Conversion" (Lo et al., Interspeech 2019)

	Utterance level				System level			
Model	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
Pretrained	0.831	0.374	0.393	0.275	0.541	0.354	0.352	0.243
From scratch	0.777	0.304	0.261	0.178	0.504	0.239	0.181	0.117
Fine-tuned	0.417	0.715	0.711	0.529	0.162	0.852	0.862	0.663
FT+sil.aug	0.428	0.713	0.709	0.528	0.153	0.854	0.861	0.665
FT+speed aug	0.421	0.716	0.707	0.526	0.176	0.857	0.867	0.672
FT+both aug	**0.305**	**0.796**	**0.791**	**0.604**	**0.096**	**0.905**	**0.912**	**0.737**

Experiments: MOSNet trained on BVCC

- MOSNet: a CNN-BLSTM architecture for MOS prediction
- Experiments:
  - Zero-shot performance of MOSNet pretrained on VCC2018
  - Train MOSNet from scratch on BVCC
  - Fine-tune pretrained MOSNet on BVCC
  - Fine-tune pretrained MOSNet on BVCC + silence augmentation
  - Fine-tune pretrained MOSNet on BVCC + speed augmentation
  - FIne-tune pretrained MOSNet on BVCC + both augmentations

"MOSNet: Deep Learning based Objective Assessment for Voice Conversion" (Lo et al., Interspeech 2019)

Model	Utterance level				System level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
Pretrained	0.831	0.374	0.393	0.275	0.541	0.354	0.352	0.243
From scratch	0.777	0.304	0.261	0.178	0.504	0.239	0.181	0.117
Fine-tuned	0.417	0.715	0.711	0.529	0.162	0.852	0.862	0.663
FT+sil.aug	0.428	0.713	0.709	0.528	0.153	0.854	0.861	0.665
FT+speed aug	0.421	0.716	0.707	0.526	0.176	0.857	0.867	0.672
FT+both aug	**0.305**	**0.796**	**0.791**	**0.604**	**0.096**	**0.905**	**0.912**	**0.737**

- Training from scratch on BVCC was worse than simply using the pretrained model!
- BVCC has a smaller number of audio samples in total than VCC2018 (but with more ratings per sample). -> not enough data!

Experiments: MOSNet trained on BVCC

- MOSNet: a CNN-BLSTM architecture for MOS prediction
- Experiments:
  - Zero-shot performance of MOSNet pretrained on VCC2018
  - Train MOSNet from scratch on BVCC
  - Fine-tune pretrained MOSNet on BVCC
  - Fine-tune pretrained MOSNet on BVCC + silence augmentation
  - Fine-tune pretrained MOSNet on BVCC + speed augmentation
  - FIne-tune pretrained MOSNet on BVCC + both augmentations

"MOSNet: Deep Learning based Objective Assessment for Voice Conversion" (Lo et al., Interspeech 2019)

	Utterance level				System level			
Model	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
Pretrained	0.831	0.374	0.393	0.275	0.541	0.354	0.352	0.243
From scratch	0.777	0.304	0.261	0.178	0.504	0.239	0.181	0.117
Fine-tuned	0.417	0.715	0.711	0.529	0.162	0.852	0.862	0.663
FT+sil.aug	0.428	0.713	0.709	0.528	0.153	0.854	0.861	0.665
FT+speed aug	0.421	0.716	0.707	0.526	0.176	0.857	0.867	0.672
FT+both aug	**0.305**	**0.796**	**0.791**	**0.604**	**0.096**	**0.905**	**0.912**	**0.737**

- Fine-tuning the pretrained model gives a large jump in performance

Experiments: MOSNet trained on BVCC

- MOSNet: a CNN-BLSTM architecture for MOS prediction
- Experiments:
  - Zero-shot performance of MOSNet pretrained on VCC2018
  - Train MOSNet from scratch on BVCC
  - Fine-tune pretrained MOSNet on BVCC
  - Fine-tune pretrained MOSNet on BVCC + silence augmentation
  - Fine-tune pretrained MOSNet on BVCC + speed augmentation
  - FIne-tune pretrained MOSNet on BVCC + both augmentations

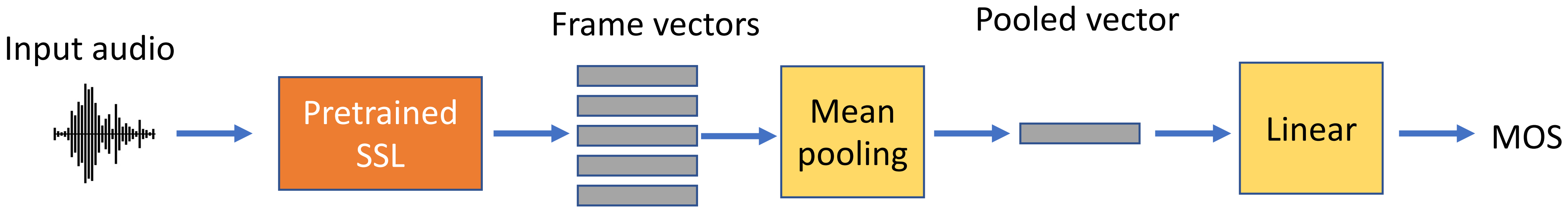
"MOSNet: Deep Learning based Objective Assessment for Voice Conversion" (Lo et al., Interspeech 2019)

	Utterance level				System level			
Model	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
Pretrained	0.831	0.374	0.393	0.275	0.541	0.354	0.352	0.243
From scratch	0.777	0.304	0.261	0.178	0.504	0.239	0.181	0.117
Fine-tuned	0.417	0.715	0.711	0.529	0.162	0.852	0.862	0.663
FT+sil.aug	0.428	0.713	0.709	0.528	0.153	0.854	0.861	0.665
FT+speed aug	0.421	0.716	0.707	0.526	0.176	0.857	0.867	0.672
FT+both aug	**0.305**	**0.796**	**0.791**	**0.604**	**0.096**	**0.905**	**0.912**	**0.737**

- Using both kinds of data augmentation improves even more

Experiments: Fine-tune SSL using BVCC

- Really simple fine-tuning of pretrained Fairseq models for the MOS prediction task



w2v_small
libri960_big
w2v_vox_new
w2v_large
xlsr_53_56k
hubert_base_ls960
hubert_large_ll60k

https://github.com/pytorch/fairseq

Experiments: Fine-tune SSL using BVCC

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

Best results

Experiments: Fine-tune SSL using BVCC

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

Third best model on the dev set.
XLSR was pretrained on **multilingual** data.

Experiments: OOD

- We picked the best and most interesting models trained/fine-tuned on BVCC in the previous experiments and analyzed their generalization ability
  - **MOSNet:**
    - Pretrained on VCC2018
    - Fine-tuned to BVCC
    - Fine-tuned to BVCC + two kinds of data augmentation
  - **SSL fine-tuned to BVCC:**
    - w2v_small
    - w2v_large
    - xlsr (multilingual)
- 2 generalization scenarios:
  - **Zero-shot:** OOD data is completely unseen
  - **Fine-tuned** using the fine-tuning set of the OOD data

Experiments: OOD

MN PT = MOSNet pretrained

MN FT-BVCC = pretrained MOSNet fine-tuned on BVCC

MN FT + aug = pretrained MOSNet fine-tuned on data-augmented BVCC

Table 5: Out-of-domain utterance-level results

Model	Zero-shot MSE	Zero-shot LCC	Zero-shot SRCC	Zero-shot KTAU	Fine-tune MSE	Fine-tune LCC	Fine-tune SRCC	Fine-tune KTAU
ASV2019								
MN PT	1.912	0.142	0.159	0.112	1.217	0.379	0.386	0.273
MN FT-BVCC	1.641	0.218	0.219	0.154	1.249	0.386	0.401	0.286
MN FT+aug	1.617	0.199	0.218	0.153	1.240	0.368	0.377	0.268
w2v_small	1.498	**0.470**	**0.491**	**0.352**	1.073	0.541	**0.558**	**0.405**
w2v_large	1.589	0.453	0.478	0.344	**1.065**	**0.548**	0.557	0.404
xlsr	**1.371**	0.409	0.423	0.301	1.192	0.518	0.525	0.377
BC2019								
MN PT	**0.823**	0.432	0.402	0.276	0.443	0.738	0.690	0.514
MN FT-BVCC	1.328	0.444	0.470	0.321	0.444	0.743	0.692	0.517
MN FT+aug	2.202	0.407	0.488	0.334	0.406	0.770	0.705	0.526
w2v_small	3.672	0.553	0.559	0.409	0.356	0.878	0.840	0.651
w2v_large	3.023	0.575	**0.618**	**0.440**	**0.235**	**0.879**	**0.841**	**0.653**
xlsr	1.924	**0.576**	0.596	0.414	0.274	0.858	0.812	0.621
COM2018								
MN PT	**0.510**	0.398	0.383	0.269	0.404	0.574	0.533	0.386
MN FT-BVCC	0.768	0.420	0.391	0.276	0.458	0.558	0.535	0.387
MN FT+aug	0.797	0.375	0.357	0.251	0.433	0.550	0.522	0.376
w2v_small	1.200	0.476	0.423	0.297	**0.352**	**0.674**	**0.667**	**0.497**
w2v_large	0.951	0.425	0.380	0.268	0.436	0.559	0.535	0.387
xlsr	0.558	**0.501**	**0.480**	**0.341**	1.383	0.369	0.379	0.268

Experiments: OOD

MN PT = MOSNet pretrained

MN FT-BVCC = pretrained MOSNet fine-tuned on BVCC

MN FT + aug = pretrained MOSNet fine-tuned on data-augmented BVCC

Table 5: Out-of-domain utterance-level results

Model	Zero-shot MSE	Zero-shot LCC	Zero-shot SRCC	Zero-shot KTAU	Fine-tune MSE	Fine-tune LCC	Fine-tune SRCC	Fine-tune KTAU
ASV2019								
MN PT	1.912	0.142	0.159	0.112	1.217	0.379	0.386	0.273
MN FT-BVCC	1.641	0.218	0.219	0.154	1.249	0.386	0.401	0.286
MN FT+aug	1.617	0.199	0.218	0.153	1.240	0.368	0.377	0.268
w2v_small	1.498	**0.470**	**0.491**	**0.352**	1.073	0.541	**0.558**	**0.405**
w2v_large	1.589	0.453	0.478	0.344	**1.065**	**0.548**	0.557	0.404
xlsr	**1.371**	0.409	0.423	0.301	1.192	0.518	0.525	0.377
BC2019								
MN PT	**0.823**	0.432	0.402	0.276	0.443	0.738	0.690	0.514
MN FT-BVCC	1.328	0.444	0.470	0.321	0.444	0.743	0.692	0.517
MN FT+aug	2.202	0.407	0.488	0.334	0.406	0.770	0.705	0.526
w2v_small	3.672	0.553	0.559	0.409	0.356	0.878	0.840	0.651
w2v_large	3.023	0.575	**0.618**	**0.440**	**0.235**	**0.879**	**0.841**	**0.653**
xlsr	1.924	**0.576**	0.596	0.414	0.274	0.858	0.812	0.621
COM2018								
MN PT	**0.510**	0.398	0.383	0.269	0.404	0.574	0.533	0.386
MN FT-BVCC	0.768	0.420	0.391	0.276	0.458	0.558	0.535	0.387
MN FT+aug	0.797	0.375	0.357	0.251	0.433	0.550	0.522	0.376
w2v_small	1.200	0.476	0.423	0.297	**0.352**	**0.674**	**0.667**	**0.497**
w2v_large	0.951	0.425	0.380	0.268	0.436	0.559	0.535	0.387
xlsr	0.558	**0.501**	**0.480**	**0.341**	1.383	0.369	0.379	0.268

Experiments: OOD

MN PT = MOSNet pretrained

MN FT-BVCC = pretrained MOSNet fine-tuned on BVCC

MN FT + aug = pretrained MOSNet fine-tuned on data-augmented BVCC

Table 5: Out-of-domain utterance-level results

Model	Zero-shot MSE	Zero-shot LCC	Zero-shot SRCC	Zero-shot KTAU	Fine-tune MSE	Fine-tune LCC	Fine-tune SRCC	Fine-tune KTAU
ASV2019								
MN PT	1.912	0.142	0.159	0.112	1.217	0.379	0.386	0.273
MN FT-BVCC	1.641	0.218	0.219	0.154	1.249	0.386	0.401	0.286
MN FT+aug	1.617	0.199	0.218	0.153	1.240	0.368	0.377	0.268
w2v_small	1.498	**0.470**	**0.491**	**0.352**	1.073	0.541	**0.558**	**0.405**
w2v_large	1.589	0.453	0.478	0.344	**1.065**	**0.548**	0.557	0.404
xlsr	**1.371**	0.409	0.423	0.301	1.192	0.518	0.525	0.377
BC2019								
MN PT	**0.823**	0.432	0.402	0.276	0.443	0.738	0.690	0.514
MN FT-BVCC	1.328	0.444	0.470	0.321	0.444	0.743	0.692	0.517
MN FT+aug	2.202	0.407	0.488	0.334	0.406	0.770	0.705	0.526
w2v_small	3.672	0.553	0.559	0.409	0.356	0.878	0.840	0.651
w2v_large	3.023	0.575	**0.618**	**0.440**	**0.235**	**0.879**	**0.841**	**0.653**
xlsr	1.924	**0.576**	0.596	0.414	0.274	0.858	0.812	0.621
COM2018								
MN PT	**0.510**	0.398	0.383	0.269	0.404	0.574	0.533	0.386
MN FT-BVCC	0.768	0.420	0.391	0.276	0.458	0.558	0.535	0.387
MN FT+aug	0.797	0.375	0.357	0.251	0.433	0.550	0.522	0.376
w2v_small	1.200	0.476	0.423	0.297	**0.352**	**0.674**	**0.667**	**0.497**
w2v_large	0.951	0.425	0.380	0.268	0.436	0.559	0.535	0.387
xlsr	0.558	**0.501**	**0.480**	**0.341**	1.383	0.369	0.379	0.268

Experiments: OOD

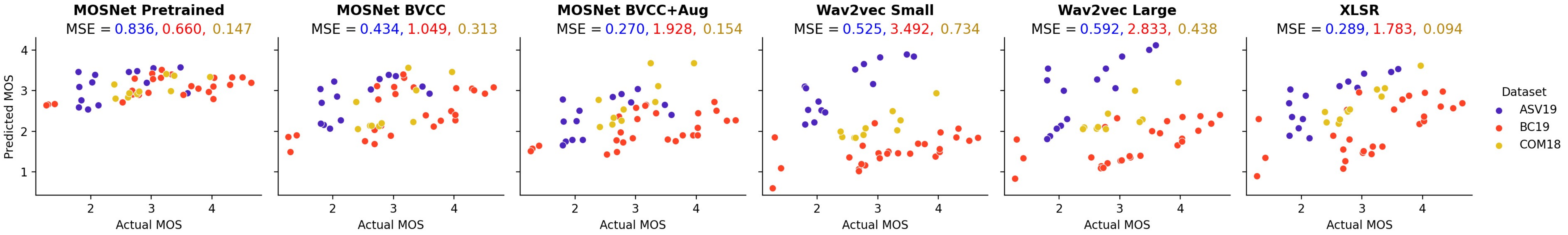


Fig. 2: Scatter plot of system-level zero-shot prediction results for each system.

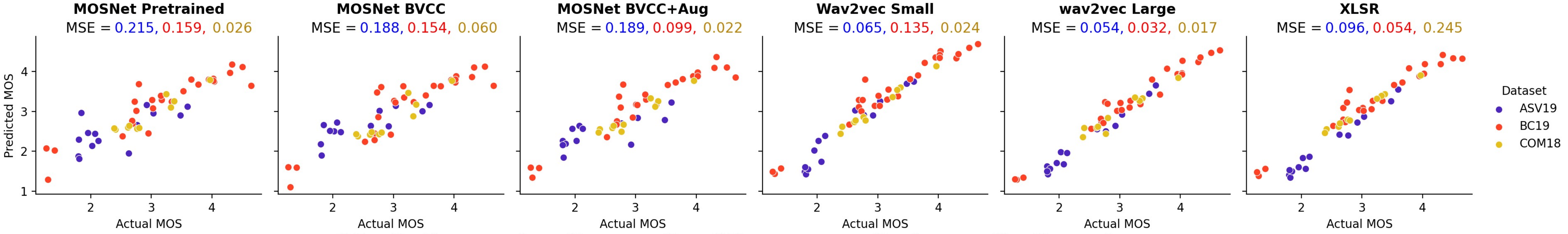


Fig. 3: Scatter plot of system-level fine-tune prediction results for each system.

Analysis of difficulty of unseen categories

Table 6: Analysis of unseen categories. Mean and standard deviations of squared errors for the unseen categories are shown. Unseen categories whose mean squared error is significantly higher than their seen counterparts are shown in bold.

Data	MN PT	MN FT	MN FT-aug	w2v_sm	w2v_lg	xlsr
Unseen speakers						
ASV19	**1.33±1.65**	1.28±1.52	1.23±1.48	1.02±1.72	1.04±1.77	1.18±2.04
Unseen systems						
ASV19	**1.36±1.45**	**1.43±1.51**	**1.43±1.54**	**1.23±1.58**	**1.26±1.82**	**1.43±2.15**
BC19	**0.77±1.11**	**0.67±1.04**	**0.76±1.10**	**0.87±0.98**	**0.41±0.61**	**0.56±0.78**
COM18	0.42±0.61	0.50±0.71	0.47±0.68	0.33±0.48	**0.52±0.74**	0.35±0.51
Unseen listeners						
ASV19	0.76±1.13	0.70±1.19	**0.71±1.25**	**0.58±1.46**	0.55±1.55	0.57±1.62
Unseen texts						
BC19	0.30±0.31	0.26±0.36	0.35±0.52	0.26±0.43	0.13±0.16	0.23±0.40
COM19	0.43±0.69	0.51±0.82	0.48±0.76	**0.47±0.71**	0.49±0.75	**0.51±0.78**

Conclusions

- **A large and diverse training dataset** can help to improve MOS prediction
  - In the case where pretrained models are **finetuned.**
  - Although our dataset had broad coverage of different synthesis methods, it was not enough data to train MOSNets from scratch.
- **Data augmentation** can help to improve automatic MOS prediction
  - But only for smaller models like MOSNet.
  - Data augmentation did **not** improve SSL-based models.
- **Self-supervised learning** (SSL) based speech models can be successfully finetuned for the MOS prediction task
  - And they can also generalize well to new listening tests with further **finetuning** using only a small amount of OOD data.
- **Unseen systems** are the most challenging category for MOS predictors to predict.

The VoiceMOS Challenge 2022

- Accepted as a special session at Interspeech 2022!

VoiceMOS Challenge