TTS Evaluation

HLT 王卉

TTS evaluation

- Speech quality assessment
  - text-to-speech
  - voice conversion
  - RTC(real time communication)
- evaluation
  - Objective assessment
  - Subjective assessment
  - Behavioral assessment

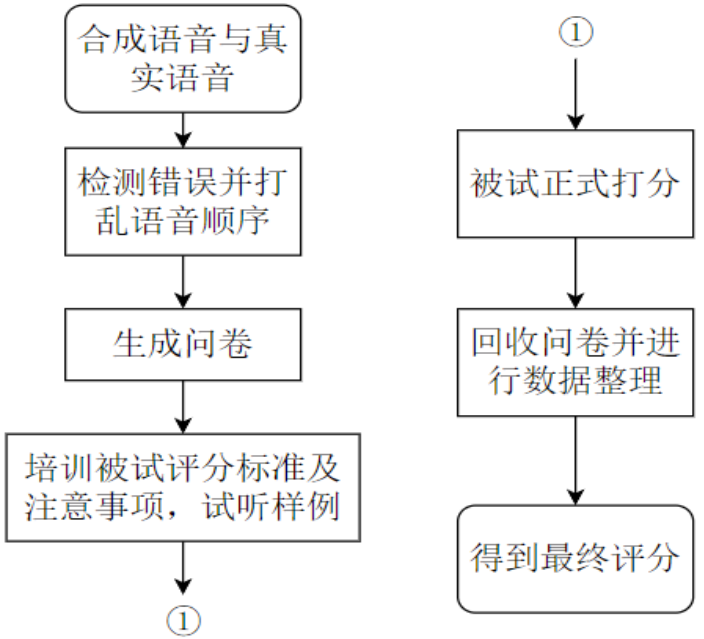
Objective assessment

- Metrics
  - melcepstral distortion (MCD)
  - Short-Time Objective Intelligibility, STOI
  - PESQ family of ITU standards:ITU-T P.861 (MNB)、ITU-T P.862 (PESQ)
  - ITU-T P.863 (POLQA)
- Advantages and disadvantages
  - reduces the need for expensive, time-consuming, and noisy subjective evaluations
  - these metrics do not align well with human perception

[1] https://github.com/AI-Unicamp/TTS-Objective-Metrics

Subjective assessment

- Mean Opinion Score， MOS
  - ITU-T P.800:Absolute Category Rating， ACR
  - naturalness MOS, similarity MOS
- Comparative Mean Opinion Score
  - 7 points (from -3 to 3)
  - sensitive to the difference
- ABX test



等级	MOS 分值	用户满意度
优	5	流畅自然, 可完全放松, 不需要注意力
良	4	失真不明显, 需要注意, 但不需明显集中
中	3	明显失真, 需要中度程度的注意力
差	2	严重的失真, 需要集中注意力
劣	1	语音质量极差, 即使努力去听, 也很难听懂

表 1.1: 主观意见得分的评估标准

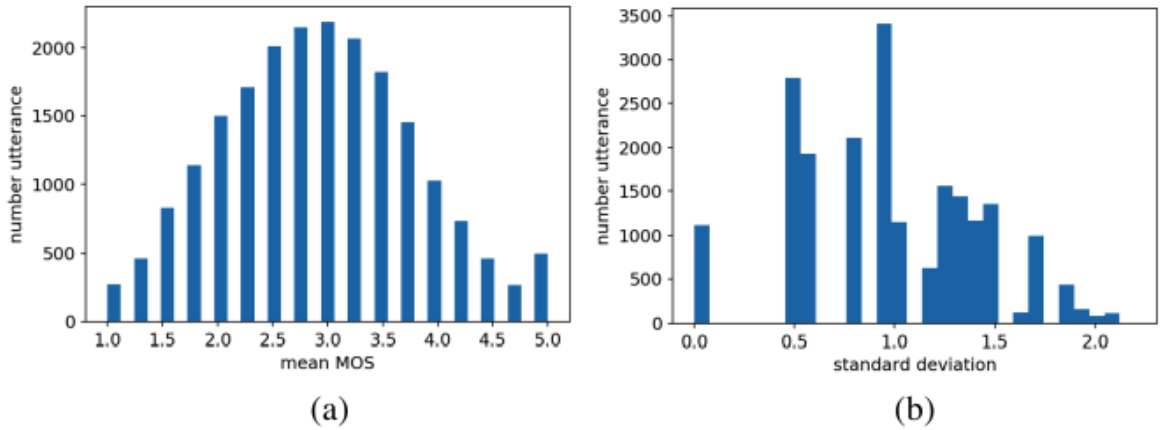
音频级别	平均意见得分	评价标准
优	5.0	很好，听得清楚；延迟小，交流流畅
良	4.0	稍差，听得清楚；延迟小，交流欠流畅，有点杂音
中	3.0	还可以，听不太清；有一定延迟，可以交流
差	2.0	勉强，听不太清；延迟较大，交流需要重复多遍
劣	1.0	极差，听不懂；延迟大，交流不通畅

MOS Prediction

- motivation
  - expensive, time-consuming
  - Application scenarios are limited
- metrics
  - Utterance vs System
  - mean squared error (MSE)
  - Linear Correlation Coefficient (LCC)
  - Spearman Rank Correlation Coefficient (SRCC)
  - Kendall Tau Rank Correlation (KTAU)

Dataset

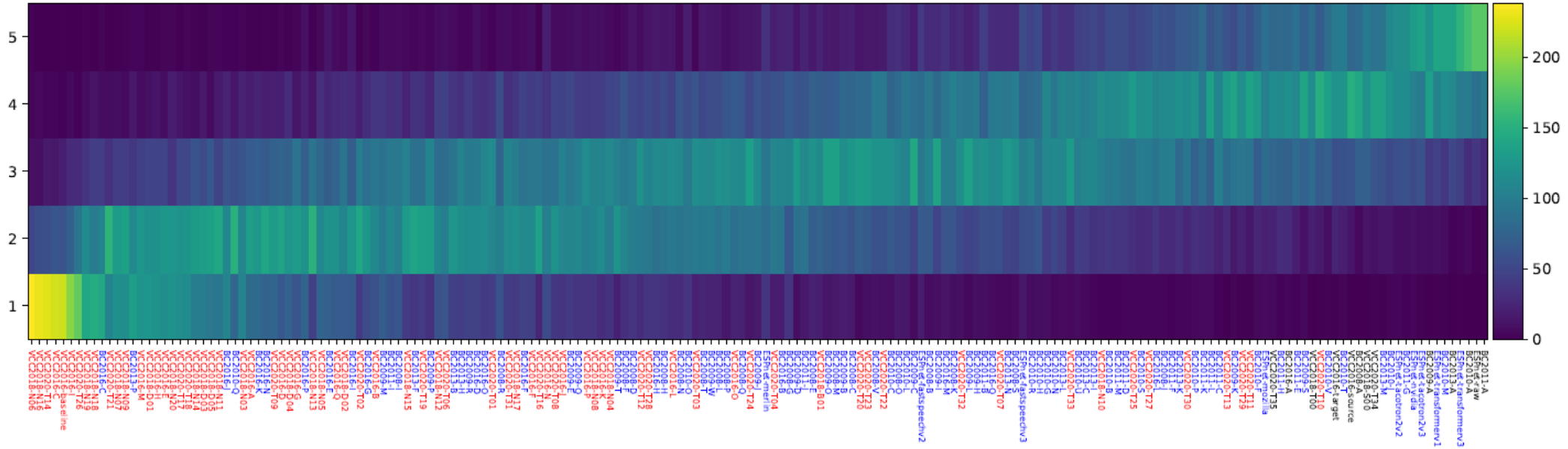
- VCC2018:The Voice Conversion Challenge (VCC) 2018
- 28292×4=113,168: 82,304 naturalness assessments and 30,864 similarity assessments



Lorenzo-Trueba, Jaime et al. “The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods.” Odyssey (2018).

Dataset

- BVCC： Blizzard and Voice Conversion Challenges： 187 systems,7106 samples×8
- train/dev/test:0.7/0.15/0.15



Cooper, E., & Yamagishi, J. (2021). How do Voices from Past Speech Synthesis Challenges Compare Today? ArXiv, abs/2105.02373.

Dataset

- **ASV2019[2]**: English audio samples, from ASVSpoof Challenge2019, scale of 1-10 (different target task)
- **BC2019**：Chinese TTS samples, from 2019 Blizzard Challenge, rated by native speakers of Chinese, train/dev/test:136/136/540 samples,unlabel(540)
- COM2018：Japanese audio samples, 10 systems, using data from the Japanese female speaker “F009” from the XIMERA database[1]

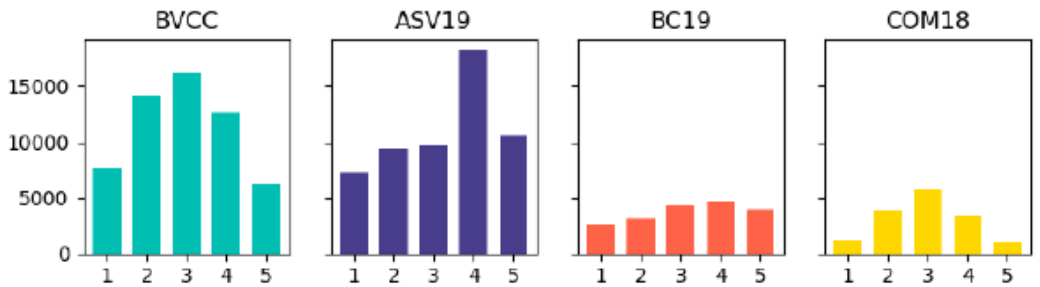


Fig. 1: Distributions of scores for each dataset

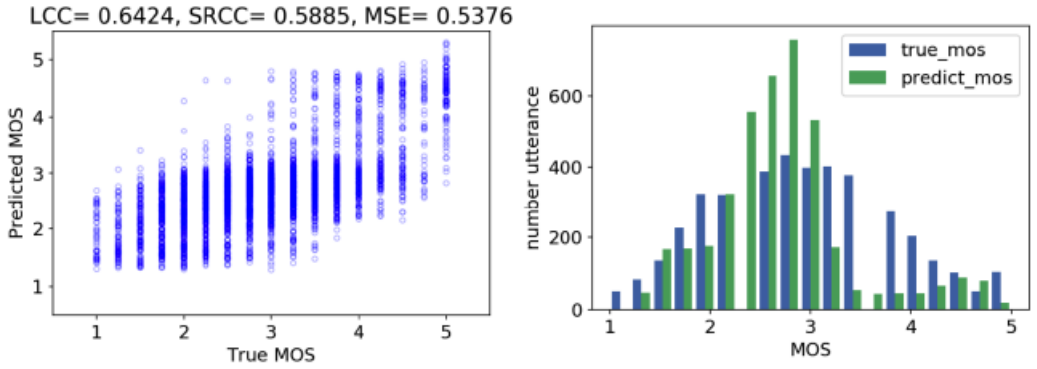
[1] Cooper, E., Huang, W., Toda, T., & Yamagishi, J. (2022). Generalization Ability of MOS Prediction Networks. ICASSP.
[2] https://datashare.ed.ac.uk/handle/10283/3336

Model

- MOSNet

model	BLSTM	CNN	CNN-BLSTM
input layer	input (***N X 257 mag spectrogram***)		
conv. layer		$\left\{\begin{matrix} conv3-(channels)/1 \\ conv3-(channels)/1 \\ conv3-(channels)/3 \end{matrix}\right\} X4$ $channels$ = [16, 32, 64, 128]	
recurrent layer	BLSTM-128		BLSTM-128
FC layer	FC-64, ReLU, dropout	FC-64, ReLU, dropout	FC-128, ReLU, dropout
	FC-1 (***frame-wise scores***)		
output layer	average pool (***utterance score***)		

$$O = \frac{1}{S}\sum_{s=1}^{S}[(\hat{Q}_s - Q_s)^2 + \frac{\alpha}{T_s}\sum_{t=1}^{T_s}(\hat{Q}_s - q_{s,t})^2]$$



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

Lo, C., Fu, S., Huang, W., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H. (2019). MOSNet: Deep Learning based Objective Assessment for Voice Conversion. ArXiv, abs/1904.08352.

Model

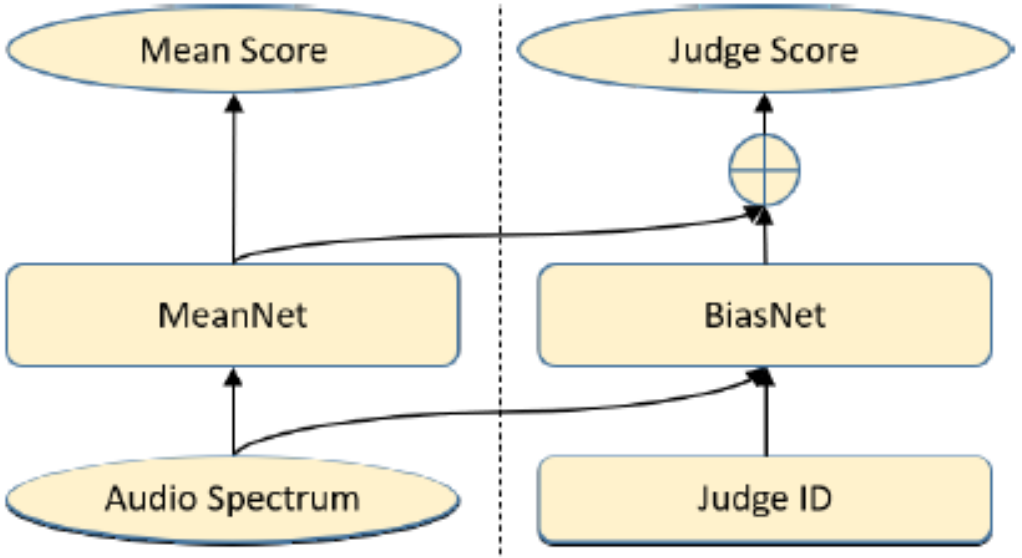
- utilize every judge score in MOS datasets
- use Self-supervised learning model
- design loss function
- data agumentation

MOS Prediction

- **utilize every judge score in MOS datasets**
  - MBnet
  - LDNet

$$loss = \mathcal{L}_C(W_m(a_i), \bar{s}_i) + \lambda \mathcal{L}_C(W_b(a_i, J_k) + W_m(a_i), s_i^k) \quad (2)$$

Model	VCC 2018						VCC 2016		
	utterance-level			*system-level*			*system-level*		
	LCC	SRCC	MSE	LCC	SRCC	MSE	LCC	SRCC	MSE
MOSNet [14]	0.638	0.611	0.465	0.964	0.922	0.047	0.915	0.862	0.308
MBNet	**0.680**	**0.647**	**0.426**	**0.977**	**0.949**	**0.029**	**0.941**	**0.920**	**0.188**



Leng, Y., Tan, X., Zhao, S., Soong, F.K., Li, X., & Qin, T. (2021). MBNET: MOS Prediction for Synthesized Speech with Mean-Bias Network. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 391-395.

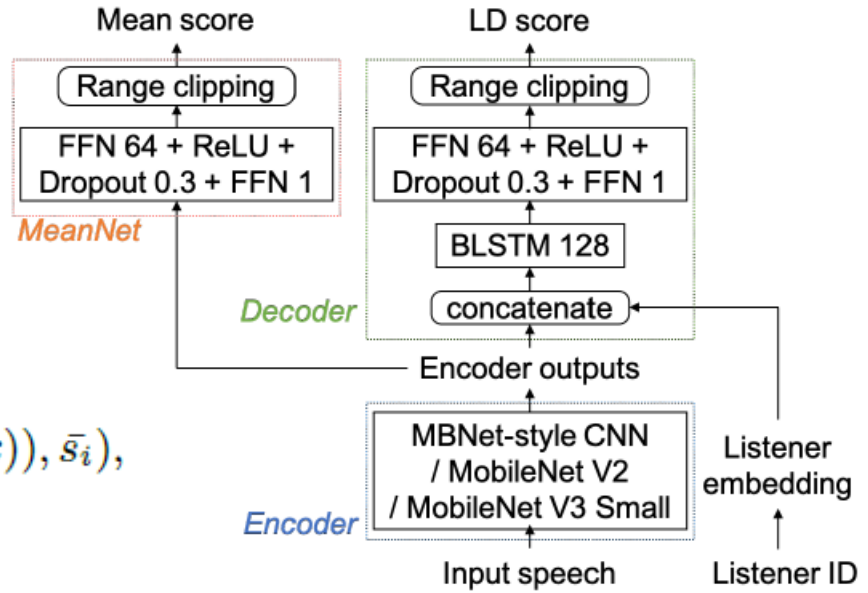
MOS Prediction

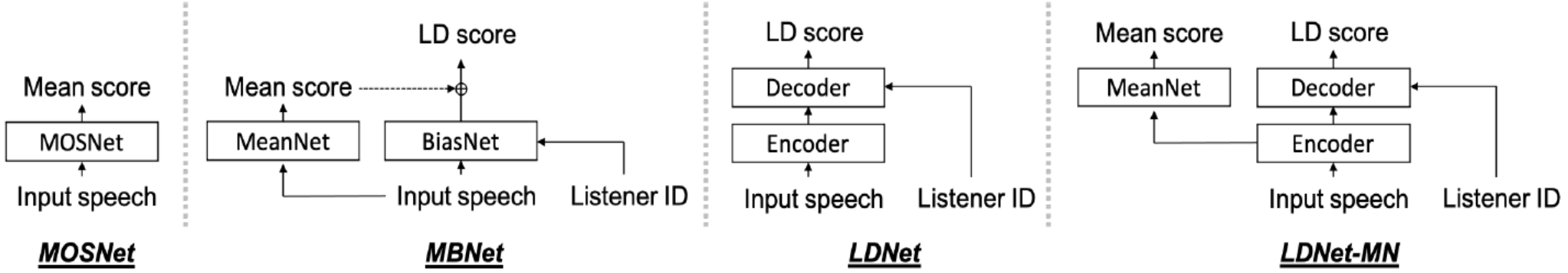
- **utilize every judge score in MOS datasets**
  - MBnet
  - LDNet
    - all listeners
    - mean listener

$$\mathcal{L}_{\text{LDNet-MN}} = \alpha\mathcal{L}_{\text{MTL}} + \lambda\mathcal{L}_{\text{LD}}.$$

$$\mathcal{L}_{\text{MTL}} = \text{MSE}(\text{MeanNet}(\text{Encoder}(x)), \bar{s}_i),$$

$$\mathcal{L}_{\text{LDNet}} = \mathcal{L}_{\text{LD}} = \text{MSE}(f(x_i, l_i^j), s_i^j).$$





MOS Prediction

- LDNet
  - all listeners
  - mean listener

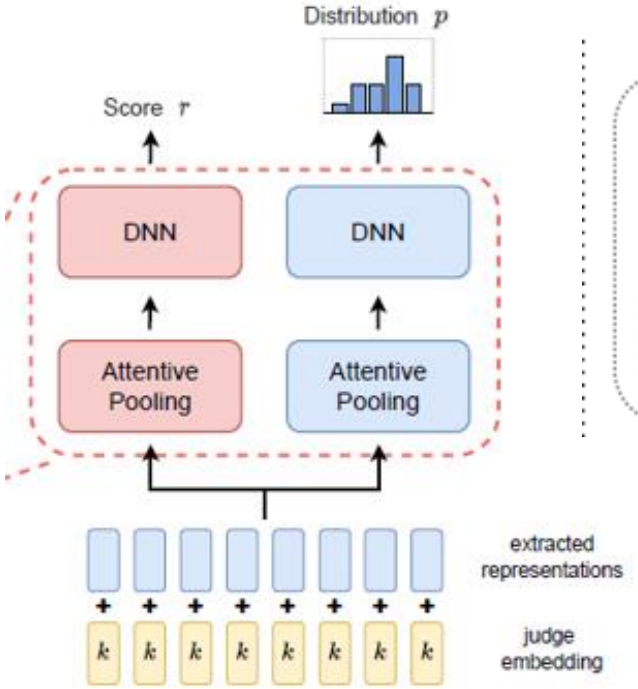
Table 1: Results on the VCC2018 test set and the BVCC test set. "MN", "All" and "ML" stand for mean net, all listeners and mean listener inference, respectively. For MSE, the smaller the better; for LCC and SRCC, the larger the better.

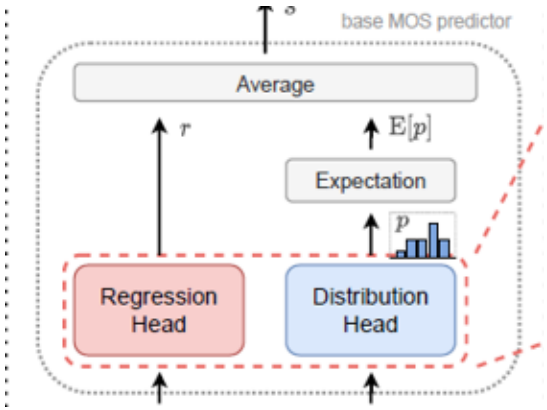
Model	Config (Enc./Dec./MeanNet)	Model size	Mode	VCC2018 Utterance level MSE	VCC2018 Utterance level LCC	VCC2018 Utterance level SRCC	VCC2018 System level MSE	VCC2018 System level LCC	VCC2018 System level SRCC	BVCC Utterance level MSE	BVCC Utterance level LCC	BVCC Utterance level SRCC	BVCC System level MSE	BVCC System level LCC	BVCC System level SRCC
MOSNet	Numbers from [6]	-	MN	0.538	0.642	0.589	0.084	0.957	0.888	-	-	-	-	-	-
MOSNet	Numbers from [7]	-	MN	0.465	0.638	0.611	0.047	0.964	0.922	-	-	-	-	-	-
MOSNet	Model from [6]	-	MN	-	-	-	-	-	-	0.816	0.294	0.263	0.563	0.261	0.266
MBNet†	Numbers from [7]	-	MN	**0.426**	**0.680**	**0.647**	0.029	0.977	0.949	-	-	-	-	-	-
(a) MBNet	Self implementation	1.38M	MN	0.955	0.658	0.630	0.549	0.978	0.957	0.669	0.757	0.765	0.522	0.854	0.860
			All	0.615	0.656	0.627	0.154	0.980	0.966	0.492	0.758	0.765	0.271	0.856	0.860
(b) LDNet	MBNet-style/RNN/-	1.18M	All	0.465	0.650	0.617	0.040	0.973	0.955	0.397	0.740	0.734	0.189	0.856	0.855
(c) LDNet	MobileV2/RNN/-	1.73M	All	0.461	0.646	0.603	0.037	0.984	0.958	0.328	0.793	0.791	0.179	0.878	0.876
(d) LDNet	MobileV3/RNN/-	1.48M	All	0.432	0.676	0.641	0.020	**0.989**	0.976	0.324	0.794	0.790	0.174	0.876	0.871
(e) LDNet	MobileV3/FFN/-	0.96M	All	0.457	0.661	0.621	**0.013**	0.988	0.976	0.333	0.788	0.784	0.173	0.876	0.870
(f) LDNet-MN	MobileV3/RNN/FFN	1.49M	All	0.437	0.671	0.635	0.023	0.987	0.971	0.324	0.794	0.791	0.187	0.869	0.868
(g) LDNet-ML	MobileV3/FNN/-	0.96M	All	0.463	0.653	0.617	0.024	0.983	0.975	**0.316**	**0.795**	**0.794**	**0.157**	0.881	0.881
			ML	0.479	0.648	0.613	0.021	0.983	**0.979**	0.333	**0.795**	**0.794**	0.169	**0.885**	**0.886**

†: The results on this row are not directly comparable with the rows below since it is unclear what data split the authors used. We suggest readers to compare results from (a) to (g).

MOS Prediction

- **utilize every judge score in MOS datasets**
  - MBnet, LDNet
  - DDOS





entropy to train this head. When a non-zero judge id is provided, we let the target of p be a one-hot vector with the s_i-th dimension equal to 1 and the other dimensions equal to 0 since $\forall i, s_i \in \{1, 2, 3, 4, 5\}$. When the 0 judge id is input, the target of p is the opinion score distribution of the utterance

$$\frac{1}{K}\left(\sum_{t=1}^{K} 1_{\{s_t=1\}}, \sum_{t=1}^{K} 1_{\{s_t=2\}}, \cdots, \sum_{t=1}^{K} 1_{\{s_t=5\}}\right),$$

DDOS (ours)	0.212	**0.880**	**0.880**	**0.707**	0.110	**0.933**	**0.932**	**0.782**
- DNN in head	0.208	0.878	0.879	0.706	0.116	0.927	0.927	0.774
- Distribution Head	0.205	0.878	0.876	0.702	0.105	0.930	0.928	0.774
- Regression Head	0.221	0.877	0.878	0.704	0.115	0.932	0.931	0.779

Tseng, W., Kao, W., & Lee, H. (2022). DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores. ArXiv, abs/2204.03219.

Model

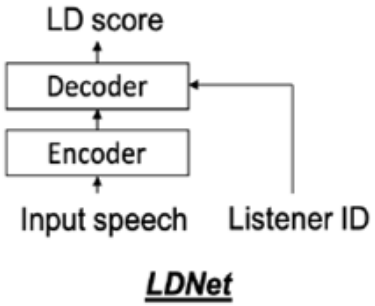
- utilize every judge score in MOS datasets
- use Self-supervised learning model
  - which model is better
  - ssl vs mosnet...
  - how to improve
- design loss function
- data agumentation

MOS Prediction

- **use Self-supervised learning model**
  - ssl-mos

Name	Training data	# params	Out dim.
wav2vec2			
w2v_small	Librispeech [27]	95m	768
libri960_big	Librispeech	317m	1024
w2v_vox_new	Libri-Light [28]	317m	1024
w2v_large	Libri-Light, CommonVoice [29], Switchboard [30], Fisher [31]	317m	1024
xlsr	MLS [32], CommonVoice, BABEL [33]	317m	1024
HuBERT			
hubert_base_ls960	Librispeech	95m	768
hubert_large_ll60k	Libri-Light	316m	1024

Model	Utterance-level	System-level	
	VCC 2018	VCC 2016	VCC 2018
wav2vec 2.0	0.734	0.966	0.99
TERA	0.727	0.943	0.987
APC	0.678	0.891	0.964
CPC	0.699	0.890	0.980
MFCC	0.183	0.196	0.326
Mel-spec.	0.215	0.487	0.618



Base model	Test set: Utterance level MSE	LCC	SRCC	KTAU	System level 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
From scratch	0.777	0.304	0.261	0.178	0.504	0.239	0.181	0.117

Cooper, E., Huang, W., Toda, T., & Yamagishi, J. (2022). Generalization Ability of MOS Prediction Networks. ICASSP.

MOS Prediction

- **use ssl (Self-supervised learning) model**
  - ssl-mos: out of domain

Name	Training data	# params	Out dim.
	wav2vec2		
w2v_small	Librispeech [27]	95m	768
libri960_big	Librispeech	317m	1024
w2v_vox_new	Libri-Light [28]	317m	1024
w2v_large	Libri-Light, CommonVoice [29], Switchboard [30], Fisher [31]	317m	1024
xlsr	MLS [32], CommonVoice, BABEL [33]	317m	1024
	HuBERT		
hubert_base_ls960	Librispeech	95m	768
hubert_large_ll60k	Libri-Light	316m	1024

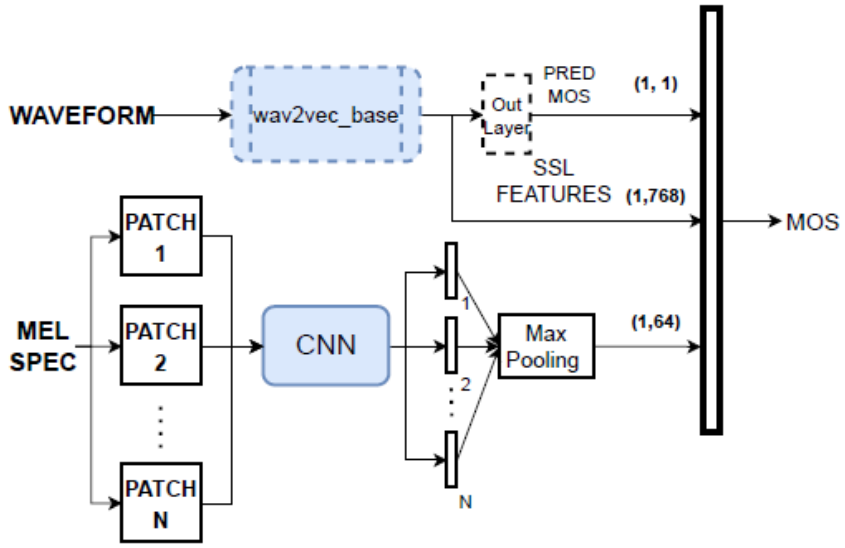
Table 5: Out-of-domain utterance-level results

Model	Zero-shot MSE	LCC	SRCC	KTAU	Fine-tune MSE	LCC	SRCC	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

Cooper, E., Huang, W., Toda, T., & Yamagishi, J. (2022). Generalization Ability of MOS Prediction Networks. ICASSP.

MOS Prediction

- **use ssl (Self-supervised learning) model**
  - 2022 VoiceMOS



	Utterance Level				System Level			
	MSE	SRCC	LCC	KTAU	MSE	SRCC	LCC	KTAU
CMP	0.325	0.790	0.790	0.601	0.169	0.872	0.874	0.696
CMP-AE	0.321	0.792	0.800	0.609	0.165	0.883	0.890	0.707
CMP-AE+wM	0.206	0.875	**0.879**	**0.702**	0.125	0.920	**0.927**	0.766
CMP-AE+wM+wF	**0.205**	**0.877**	0.878	0.701	**0.116**	**0.925**	**0.927**	**0.776**
NISQA Scratch	0.327	0.797	0.801	0.614	0.171	0.883	0.880	0.709
NISQA Pre-Trained	0.299	0.825	0.832	0.644	0.157	0.878	0.884	0.705
wav2vec2_base	0.263	0.872	0.875	0.697	0.130	0.915	0.922	0.758
wav2vec2_base NC	0.566	0.600	0.609	0.428	0.412	0.700	0.709	0.506
wav2vec2_base PSTN	1.222	0.793	0.757	0.596	1.017	0.819	0.803	0.621
wav2vec2_large	0.277	0.869	0.869	0.690	0.148	0.924	0.921	0.769
MOSA-Net	0.309	0.809	0.811	0.621	0.162	0.899	0.900	0.729
LDNet	0.334	0.787	0.785	0.599	0.178	0.873	0.873	0.691

Ragano, A., Benetos, E., Chinen, M., Martinez, H.B., Reddy, C.K., Skoglund, J., & Hines, A. (2022). A Comparison of Deep Learning MOS Predictors for Speech Synthesis Quality.

MOS Prediction

- **use ssl (Self-supervised learning) model**
  - 2022 VoiceMOS
    - UTMOS: stacking method

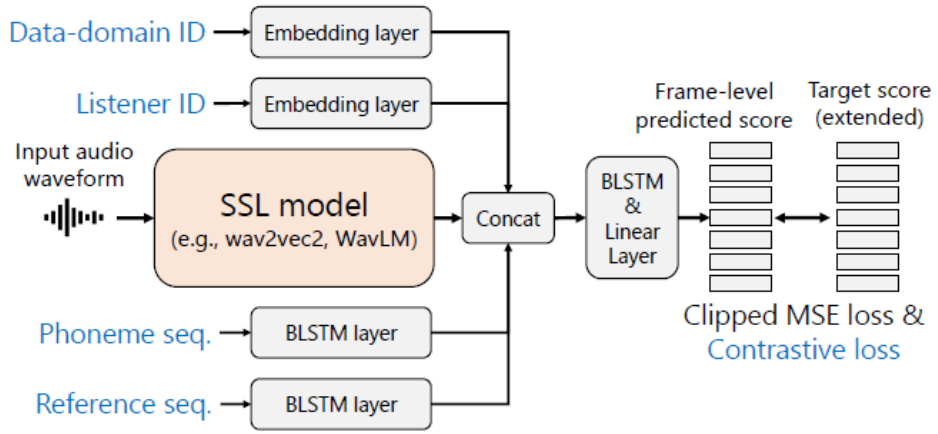


Figure 1: *Architecture of the proposed strong learner.*

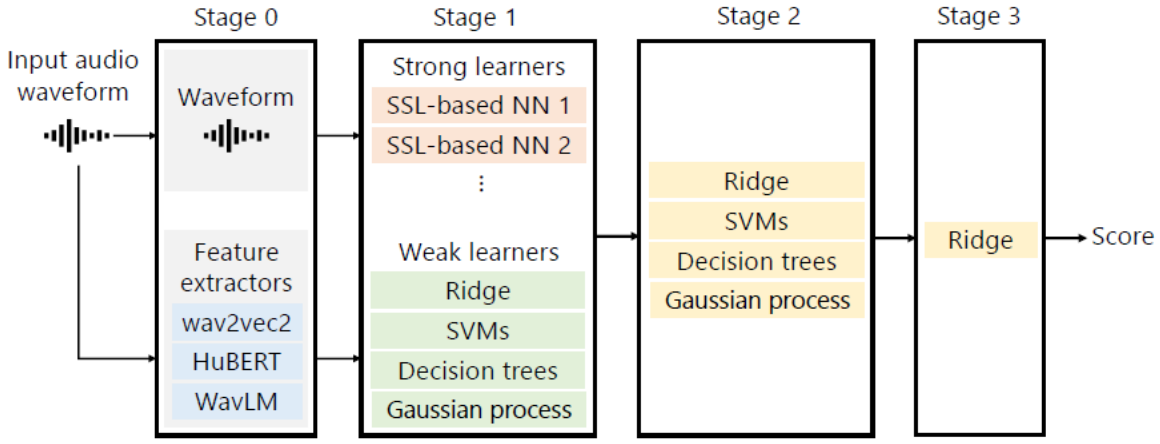
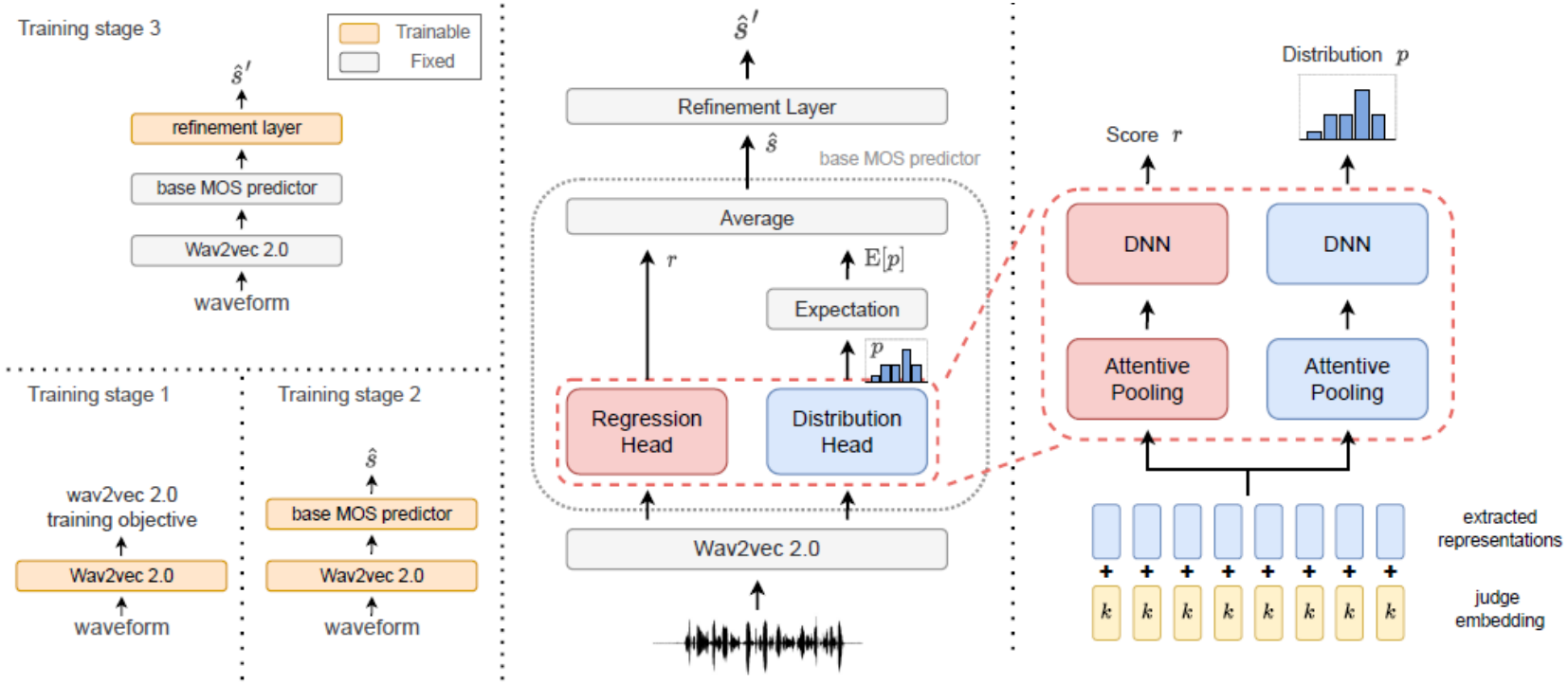


Figure 2: *Flow of stacking with strong and weak learners.*

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., & Saruwatari, H. (2022). UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. ArXiv, abs/2204.02152.

MOS Prediction

- **use ssl (Self-supervised learning) model**
  - 2022 VoiceMOS: DDOS
    - domain-adaptive pre-training (DAPT), base mos predictor, refinement layer



MOS Prediction

- **use ssl (Self-supervised learning) model**
  - 2022 VoiceMOS:
    - DDOS

Table 3: *Zero-shot ablation study.*

System-level	MSE↓	LCC↑	SRCC↑	KTAU↑
DDOS	1.119	0.766	0.797	0.637
- DNN in head	0.896	0.793	0.822	0.668
- Distribution Head	0.921	0.755	0.779	0.614
- Regression Head	4.682	0.773	0.804	0.645
- DAPT	2.813	0.696	0.704	0.559
- data augmentation	0.948	0.813	0.813	0.660

Table 1: *The performances of our framework and baselines on BVCC test set.*

	Utterance-level				System-level			
	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑
LDNet [3]	0.334	0.787	0.785	0.599	0.178	0.873	0.873	0.691
MOSA-Net [11]	0.309	0.809	0.811	0.621	0.162	0.899	0.900	0.729
SSL-MOS [4]	0.277	0.869	0.869	0.690	0.148	0.924	0.921	0.769
DDOS (ours)	0.212	**0.880**	**0.880**	**0.707**	0.110	**0.933**	**0.932**	**0.782**
- DNN in head	0.208	0.878	0.879	0.706	0.116	0.927	0.927	0.774
- Distribution Head	0.205	0.878	0.876	0.702	0.105	0.930	0.928	0.774
- Regression Head	0.221	0.877	0.878	0.704	0.115	0.932	0.931	0.779
- refinement layer	0.230	0.880	0.880	0.707	0.106	0.933	0.932	0.782
- DAPT	**0.201**	0.877	0.875	0.701	**0.102**	0.928	0.926	0.770
- data augmentation	0.213	0.880	0.879	0.705	0.118	0.927	0.925	0.774

MOS Prediction

- **loss function**
  - utterance-level score mse, mean score mse + bias score mse
  - frame-level score mse(MOSnet), segment-level score mse[1]
  - utterance-level score contrastive loss(UTMOS)
  - clipped MSE loss(MBNet, LDNet,UTMOS...)

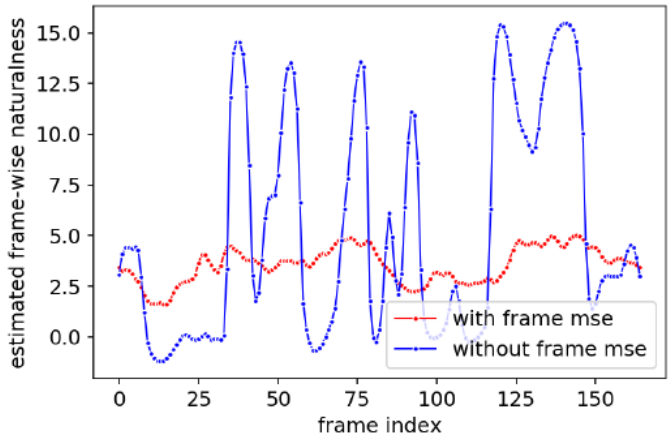


Figure 4: *Frame-wise MOS predictions of CNN-BLSTM$_{64}$.*

Table 4: *The effect of frame-level MSE on the utterance-level predictions of CNN-BLSTM$_{64}$.*

	LCC	SRCC	MSE
with frame MSE	**0.642**	**0.589**	**0.538**
without frame MSE	0.560	0.528	2.525

[1]Tseng, W., Huang, C., Kao, W., Lin, Y.Y., & Lee, H. (2021). Utilizing Self-supervised Representations for MOS Prediction. Interspeech.

Model

- utilize every judge score in MOS datasets
- use Self-supervised learning model
- design loss function
- data agumentation

MOS Prediction

- **loss function**
  - utterance-level score mse, mean score mse + bias score mse
  - frame-level score mse(MOSnet), segment-level score mse
  - utterance-level score contrastive loss(UTMOS)
  - clipped MSE loss(MBNet, LDNet,UTMOS...)

$$(d_{x_1,x_2} = s_1 - s_2)$$

$$(\widehat{d}_{x_1,x_2} = \widehat{s}_1 - \widehat{s}_2)$$

$$\mathcal{L}^{\text{con}}_{x_1,x_2} = \max(0, |d_{x_1,x_2} - \widehat{d}_{x_1,x_2}| - \alpha)$$

	Utterance-level				System-level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
UTMOS strong	0.276	0.883	0.881	0.708	0.148	0.930	0.925	0.774
w/o contrastive loss	0.241	0.881	0.879	0.706	0.114	0.932	0.930	0.781
w/o listener ID	0.307	0.880	0.878	0.704	0.160	0.935	0.933	0.784
w/o phoneme encoder	0.249	0.881	0.882	0.709	0.119	0.935	**0.936**	**0.790**
w/o data augmentation	0.226	**0.885**	**0.882**	**0.710**	**0.103**	**0.936**	0.933	0.784
w/o MSE loss	**0.219**	0.882	0.880	0.707	0.114	0.932	0.929	0.778
SSL-MOS	0.380	0.869	0.871	0.695	0.223	0.920	0.918	0.758

MOS Prediction

- **loss function**
  - utterance-level score mse, mean score mse + bias score mse
  - frame-level score mse(MOSnet), segment-level score mse
  - utterance-level score contrastive loss(UTMOS)
  - clipped MSE loss(MBNet, LDNet,UTMOS...)

$$\mathcal{L}_C(y, \hat{y}) = \mathbb{1}(|y - \hat{y}| > \tau)(y - \hat{y})^2$$

the indicator function whose output is 1 when the condition is true otherwise 0

Model	SRCC		
	Utterance 18'	System 18'	System 16'
MBNet	0.647	**0.949**	**0.920**
– BiasNet	0.633	0.934	0.916
– MeanNet	0.343	0.887	0.873
– CMSE	**0.652**	0.946	0.915
– Reppad	0.644	0.929	0.912

Model

- utilize every judge score in MOS datasets
- use Self-supervised learning model
- design loss function
- data agumentation

MOS Prediction

- **data agumentation**
  - speaking-rate-changing（UTMOS，DDOS, MOSNet）
  - pitch-shifting（UTMOS，DDOS）
  - adding silence（DDOS, MOSNet）

	Utterance-level				System-level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
UTMOS strong	0.276	0.883	0.881	0.708	0.148	0.930	0.925	0.774
w/o contrastive loss	0.241	0.881	0.879	0.706	0.114	0.932	0.930	0.781
w/o listener ID	0.307	0.880	0.878	0.704	0.160	0.935	0.933	0.784
w/o phoneme encoder	0.249	0.881	0.882	0.709	0.119	0.935	**0.936**	**0.790**
w/o data augmentation	0.226	**0.885**	**0.882**	**0.710**	**0.103**	**0.936**	0.933	0.784

Table 3: MOSNet BVCC results

Model	Utterance level				System level			
	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
Pretrained [2]	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**

DDOS (ours)	0.212	**0.880**	**0.880**	**0.707**	0.110	**0.933**	**0.932**	**0.782**
- DNN in head	0.208	0.878	0.879	0.706	0.116	0.927	0.927	0.774
- Distribution Head	0.205	0.878	0.876	0.702	0.105	0.930	0.928	0.774
- Regression Head	0.221	0.877	0.878	0.704	0.115	0.932	0.931	0.779
- refinement layer	0.230	0.880	0.880	0.707	0.106	0.933	0.932	0.782
- DAPT	**0.201**	0.877	0.875	0.701	**0.102**	0.928	0.926	0.770
- data augmentation	0.213	0.880	0.879	0.705	0.118	0.927	0.925	0.774

conclusion

- motivation
- dataset
  - in-domain: VCC2018, BVCC
  - out-of-domain: ASV2019, BC2019, COM2018
- improvement
  - **utilize every judge score in MOS datasets**
  - **utilize Self-supervised learning model**
  - design loss function
  - data agumentation
- challenges
  - Generalization Ability

Thanks and Q&A