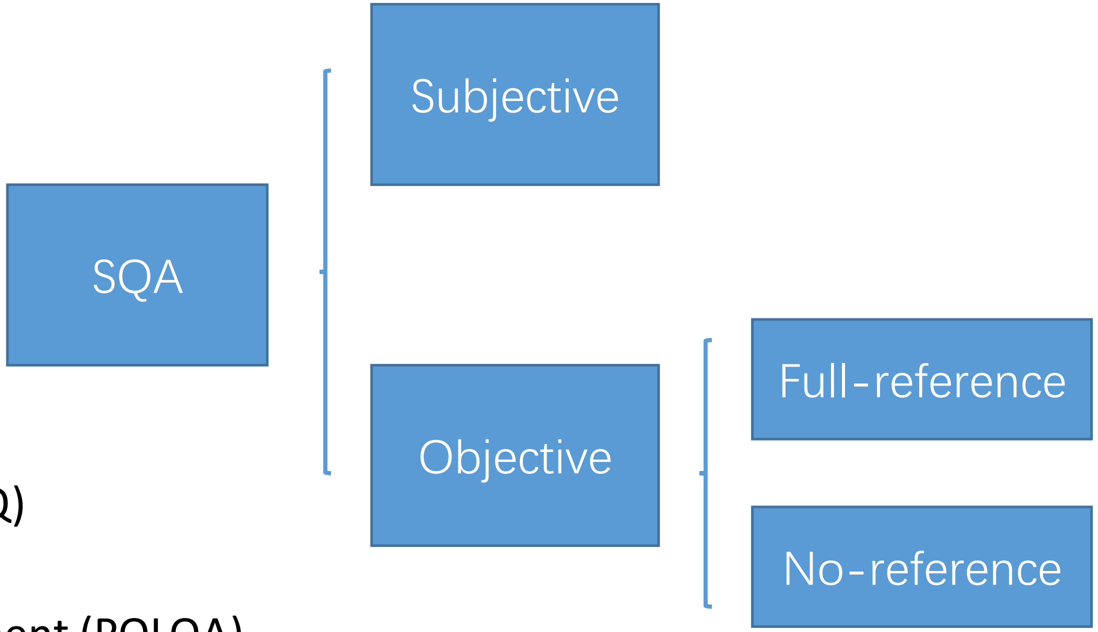
Why we say PESQ is a bad metric ?

Chen Chen

2022/09/30

SQA: Speech quality assessment

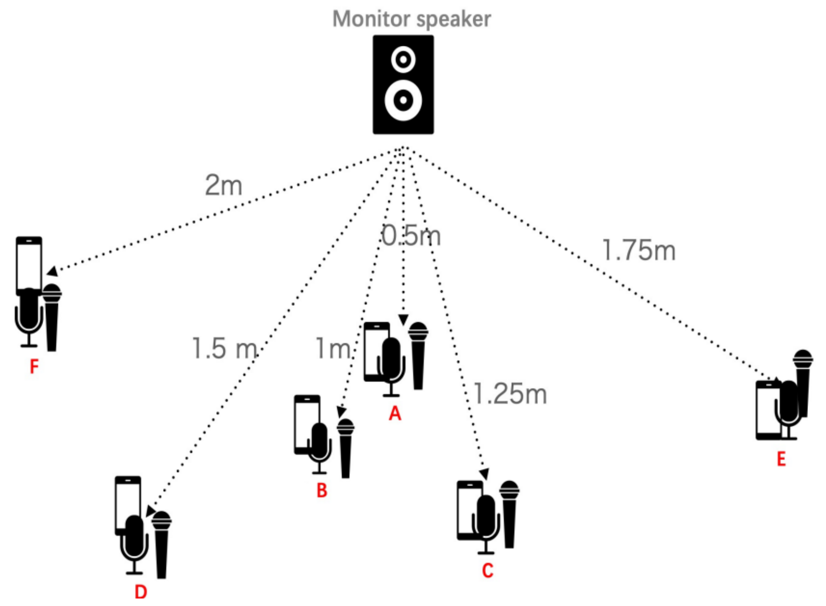
- SQA metrics
  - Subjective metrics
    - Mean Opinion Score (MOS)
  - Objective metrics
    - Full-reference / similarity-based metrics
      - Perceptual evaluation of speech quality (PESQ)
      - Mel-cepstral distance (MCD)
      - Perceptual objective listening quality assessment (POLQA)
      - Virtual Speech Quality Objective Listener (ViSQOL)
    - No-reference / non-intrusive metrics
      - MOSNET
      - Deep Noise Suppression Mean Opinion Score (DNSMOS)
      - Non-Intrusive Speech Quality Assessment (NISQA)



Datasets

- DDS
  - A new device-degraded speech dataset for speech enhancement
  - built on top of two existing datasets: DAPS and VCTK

Setting	Count	Description
Speech materials	2	DAPS, VCTK clean sets
Environments	9	conference rooms (2), offices (2), studios (3), living room (1), waiting room (1)
Devices	3	iPad Air (MEMS), Uber Mic (condenser), MPM-1000 (condenser)
Device positions	6	A(50 cm, 0°), B(100 cm, 15°) C(125 cm, 30°), D(150 cm, 45°) E(175 cm, 60°), F(200 cm, 75°)



Datasets

- DDS

Type	DAPS [31]											
	PESQ ↑	VISQOL ↑	DPAM ↓	CDPAM ↓	L1 ↓	L2 ↓	M.STFT ↓	NISQA ↑	SQAPP ↑	DNSMOS ↑	NORESQA ↓	MOS ↑
Clean	-	-	-	-	-	-	-	4.68	3.45	3.85	9.56	4.48
Confroom1	1.55	2.37	2.80	0.30	2.65	30.70	0.19	2.89	3.08	3.46	12.02	2.90
Confroom2	1.33	2.20	2.79	0.34	2.76	31.37	0.20	2.38	2.773	3.17	13.02	2.39
Office1	1.80	2.42	2.73	0.29	2.44	27.86	0.19	3.01	3.10	3.52	11.01	2.99
Office2	1.57	2.38	2.77	0.32	2.52	28.96	0.19	2.71	3.04	3.42	11.52	2.63
Studio1	1.59	2.35	2.78	0.32	2.62	29.40	0.19	2.70	2.94	2.95	12.07	2.63
Studio2	2.03	2.60	2.76	0.27	2.40	27.25	0.18	3.48	3.20	3.65	10.91	3.20
Studio3	2.03	2.54	2.83	0.29	2.53	30.50	0.17	3.58	3.18	3.61	10.86	3.28
Waitingroom1	2.11	2.55	2.75	0.27	2.39	27.80	0.17	3.46	3.23	3.55	10.81	3.42
Livingroom1	1.56	2.36	2.87	0.30	2.67	32.44	0.19	2.82	3.14	3.55	11.28	2.98

Type	VCTK [32]											
	PESQ ↑	VISQOL ↑	DPAM ↓	CDPAM ↓	L1 ↓	L2 ↓	M.STFT ↓	NISQA ↑	SQAPP ↑	DNSMOS ↑	NORESQA ↓	MOS ↑
Clean		-	-	-	-	-	-	4.14	3.37	3.62	9.97	4.18
Confroom1	1.70	2.15	2.80	0.32	2.04	22.84	0.17	2.81	3.03	3.50	13.24	2.77
Confroom2	1.48	2.04	2.75	0.38	2.12	23.31	0.19	2.37	2.75	2.93	14.35	2.29
Office1	1.94	2.14	2.69	0.34	1.88	20.90	0.17	2.81	3.00	3.37	11.82	2.79
Office2	1.70	2.12	2.75	0.37	1.95	21.74	0.18	2.64	2.97	3.32	12.49	2.60
Studio1	1.72	2.04	2.74	0.37	2.06	22.24	0.18	2.65	2.84	3.16	13.51	2.53
Studio2	2.06	2.27	2.67	0.34	1.83	20.05	0.17	3.19	3.07	3.50	11.39	2.96
Studio3	2.03	2.23	2.78	0.33	1.96	22.86	0.16	3.29	3.03	3.41	11.16	3.03
Waitingroom1	2.17	2.23	2.69	0.31	1.83	20.05	0.16	3.21	3.07	3.52	11.21	3.15
Livingroom1	1.71	2.14	2.88	0.35	2.05	24.05	0.17	2.75	3.01	3.48	12.16	2.78

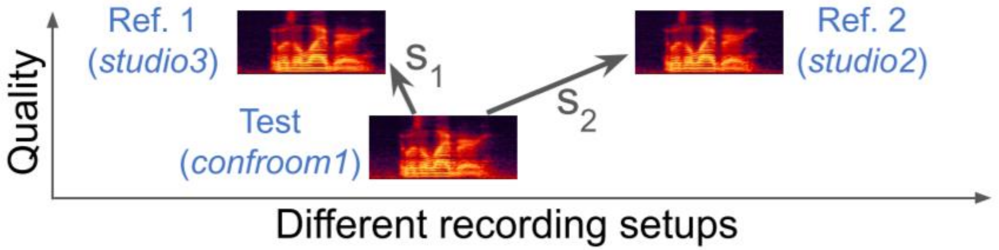
Catalog

1. Audio Similarity is Unreliable as a Proxy for Audio Quality (Interspeech 2022)
* Adobe Research | Princeton University

2. MOSNet: Deep Learning-based Objective Assessment for Voice Conversion (2019)
* Academia Sinica, Taipei, Taiwan | National Institute of Informatics, Japan

3. InQSS: a speech intelligibility and quality assessment model using a multi-task learning network (Interspeech 2022)
* Academia Sinica, Taiwan

- Motivation
  - PESQ has acknowledged shortcomings, and may not be reliable to detect subtle differences.
  - Quality measures, such as MOS, may not involve a reference that is in parallel with the test signal.
- Inspire
  - different references, although sharing the same quality, may result in different similarities when compared with the same test signal
  - different test signals, although having the same quality rating, may have significantly different similarities to a reference signal
- Experiments
  - Two reference with one test recording
  - One reference with two test recordings
  - Matching datasets acoustically



- Experiments
  - Two reference with one test recording
    - The reference recordings are judged to be equal quality
    - Majority of similarity metrics show large (up to 85%) differences, and fail to provide equal similarity ratings.
    - No-reference metric ratings are very close (up to 3.5% difference), and reflect the MOS ratings well.

Type	MOS ↑ DAPS [31]	MOS ↑ VCTK [32]
Studio2	3.20	2.96
Studio3	3.28	3.03

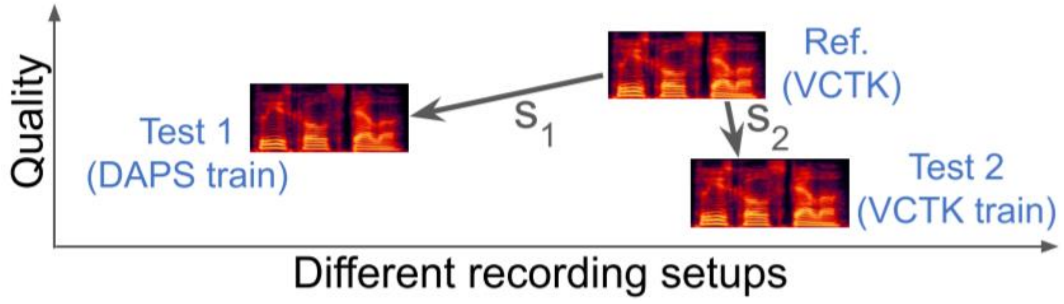
Type	PESQ ↑	VISQOL ↑	DPAM ↓	CDPAM ↓	L1 ↓	L2 ↓	M.STFT ↓	SQAPP ↑	NISQA ↑	DNSMOS ↑	NORESQA ↓
Studio2	1.81	3.32	1.79	0.13	0.75	9.00	0.10	3.71	3.80	3.64	10.79
Studio3	2.75	3.51	2.62	0.07	0.63	7.77	0.09	3.84	3.89	3.55	10.59

Table 2: ***Scenario 1:*** *Performance of similarity and no-reference metrics when reference recordings from* studio2 *and* studio3, *and test recordings from* confroom1 *are selected. We see that similarity metrics show different similarities, even though no-reference metrics (and subjective ratings - Table 1) suggest the two references are of equal quality.*

Audio Similarity is Unreliable as a Proxy for Audio Quality (Interspeech 2022)

* Adobe Research | Princeton University

- Experiments
  - One reference with two test
    - similarity measures do not reflect subjective ratings.
    - no-reference metrics reflects subjective quality well.



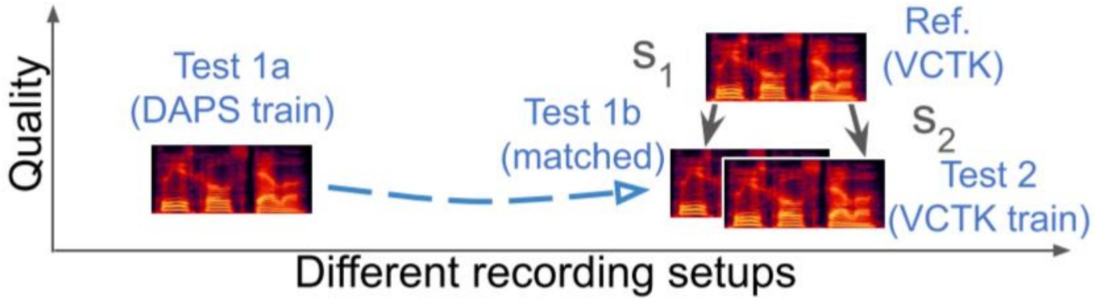
	DAPS Model	VCTK Model	Valset Noisy	Valset Clean
PESQ ↑	2.16	3.13	1.96	-
VISQOL ↑	3.60	4.18	3.81	-
DPAM ↓	2.77	1.35	1.71	-
CDPAM ↓	0.21	0.07	0.30	-
L1 ↓	2.24	0.30	0.92	-
L2 ↓	20.42	2.14	5.90	-
Multi-res STFT ↓	0.14	0.07	0.17	-
SQUAPP ↑	4.06	3.76	2.73	3.85
NISQA ↑	4.90	4.65	3.06	4.60
DNSMOS ↑	3.64	3.55	3.02	3.55
NORESQA ↓	9.57	10.53	12.88	9.44
MOS↑	4.10	4.02	2.48	4.29

Table 3: ***Scenario 2:*** *Performance of similarity metrics, no-reference metrics and MOS ratings (± 0.02) across speech enhancement (SE) models trained on two datasets (DAPS and VCTK), and evaluated on the VCTK evaluation set.*

Audio Similarity is Unreliable as a Proxy for Audio Quality (Interspeech 2022)

* Adobe Research | Princeton University

- Experiments
  - Matching datasets acoustically



Type	PESQ ↑	VISQOL ↑	DPAM ↓	CDPAM ↓	L1 ↓	L2 ↓	M.STFT ↓	SQAPP ↑	NISQA ↑	DNSMOS ↑	NORESQA ↓	MOS ↑
VCTK Model	3.13	4.18	1.35	0.07	0.30	2.14	0.07	3.76	4.65	3.55	10.53	4.02
DAPS Model	2.16	3.60	2.77	0.21	2.24	20.42	0.14	4.06	4.90	3.64	9.57	4.10
EQ Match.	2.32	3.68	2.73	0.22	2.00	25.99	0.13	3.96	4.89	3.61	10.32	4.11
Breath rem.	2.67	4.18	2.67	0.13	1.95	25.97	0.12	3.84	4.84	3.63	10.29	4.19
Energy norm.												
itr0	2.32	4.17	2.68	0.14	1.92	24.85	0.13	3.87	4.77	3.56	10.09	4.14
itr200	2.29	4.17	2.69	0.14	1.92	24.83	0.14	3.94	4.78	3.55	10.07	4.10
itr1000	2.29	4.17	2.69	0.15	1.91	24.77	0.14	3.95	4.78	3.55	10.07	4.10
Orig. Phase	3.16	4.44	1.21	0.05	0.21	0.25	0.13	3.97	4.78	3.60	10.35	4.19

Table 4: ***Scenario 3:*** *Objective measures and MOS ratings (±0.03) across pre-processing stages (Section 2.3) when recordings from DAPS trained SE model are matched to the VCTK trained SE model.*

- Conclusion
  - similarity metrics (like PESQ) are an unreliable proxy for audio quality, and should be used cautiously

MOSNet: Deep Learning-based Objective Assessment for Voice Conversion (Interspeech 2019)

* Academia Sinica, Taipei, Taiwan | National Institute of Informatics, Japan

- # Motivation
  - Objective evaluation metrics for voice conversion (VC) are not always correlated with human perception
  - Subjective evaluation metrics are time-consuming and expensive.
- # Dataset
  - large-scale listening test results of the Voice Conversion Challenge (VCC) 2018

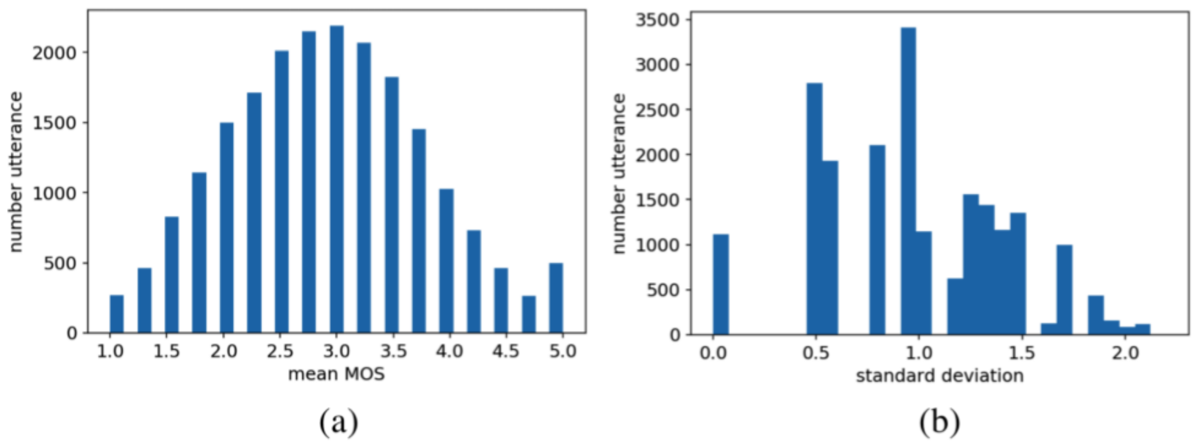


Figure 1: *Histograms of the mean (a) and standard deviation (b) of four MOS ratings for each utterance in the VCC 2018.*

MOSNet: Deep Learning-based Objective Assessment for Voice Conversion (Interspeech 2019)

* Academia Sinica, Taipei, Taiwan | National Institute of Informatics, Japan

• Method

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

- Utterance loss
- Frame-wise loss

model	**BLSTM**	**CNN**	**CNN-BLSTM**
input layer	input (*N X 257 mag spectrogram*)		
conv. layer		$\left\{\begin{array}{l} conv3-(channels)/1 \\ conv3-(channels)/1 \\ conv3-(channels)/3 \end{array}\right\} X4$ $channels$ = [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*)		

- Experiments

Table 3: *Utterance-level and system-level prediction results for different models, where the subscript denotes the batch size.*

	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

LCC: Linear Correlation Coefficient
SRCC: Spearman Rank Correlation Coefficient

MOSNet: Deep Learning-based Objective Assessment for Voice Conversion (Interspeech 2019)
* Academia Sinica, Taipei, Taiwan | National Institute of Informatics, Japan

• Experiments

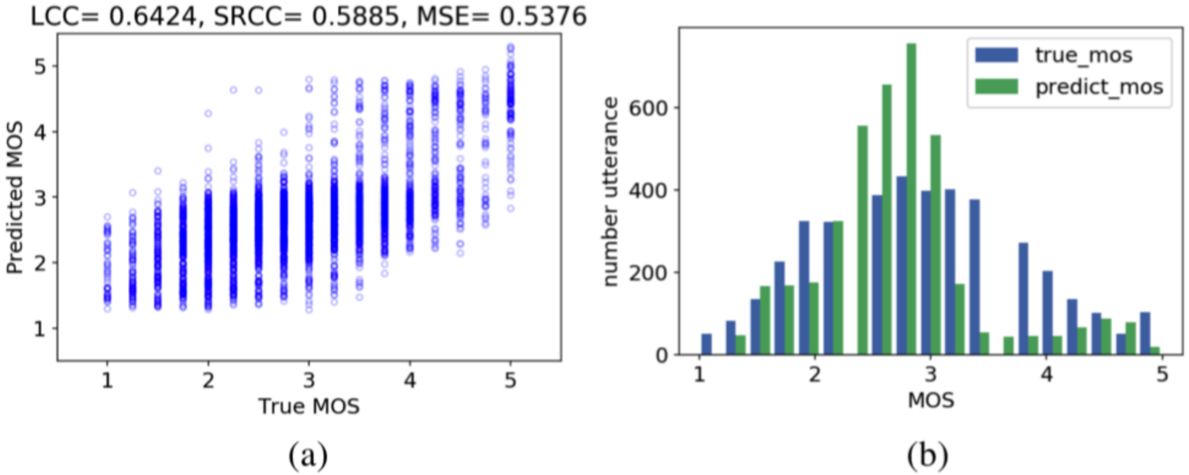


Figure 2: *Scatter plot (a) and histogram (b) of the utterance-level predictions of CNN-BLSTM$_{64}$.*

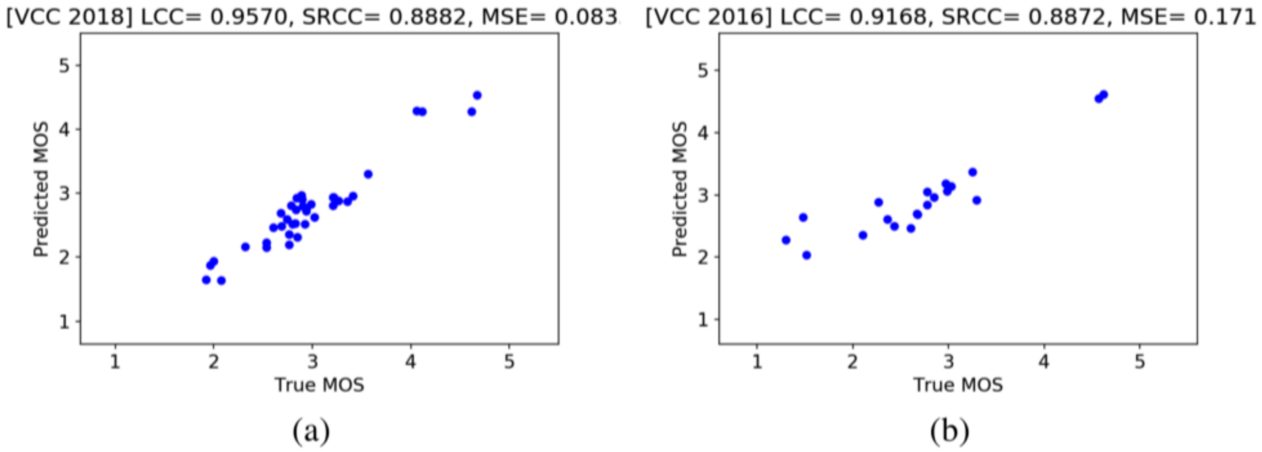


Figure 3: *Scatter plot of system-level predictions on the testing set in (a) the VCC 2018 and (b) the VCC 2016.*

InQSS: a speech intelligibility and quality assessment model using a multi-task learning network (Interspeech 2022)
* Academia Sinica, Taiwan

- Motivation
  - Similarity-based metrics
    - reference clean speech signals are often unavailable in real-world applications
    - the results might not correlate well with the listening test results
  - Non-Intrusive metrics
    - only a few studies have investigated multi-task models for intelligibility and quality assessment due to the limitations of available data
- Main work
  - Released TMHINT-QI1, a Chinese speech dataset with subjective quality and intelligibility scores
  - Propose the InQSS, a multi-task learning framework using training-from-scratch and pretrained self-supervised learning (SSL) models

- Dataset: TMHINT-QI
    - for SE
    - 16kHz, 16bit
    - 10 Chinese characters, ~3s
    - Two parts
        - First
            - 6 spk (3F3M) * 200 utts = 1200 clean utts
            - 5 noise types, 8 SNR levels
            - 3 nn-based SE methods (FCN, DDAE, Trans)
        - Second
            - 2 spk (1F1M) * 115 utts = 230 clean utts
            - 4 noise types, 4 SNR levels
            - 5 SE models
    - Total: 24,408 samples with 14,919 unique utterances

InQSS: a speech intelligibility and quality assessment model using a multi-task learning network (Interspeech 2022)

* Academia Sinica, Taiwan

- Dataset: TMHINT-QI

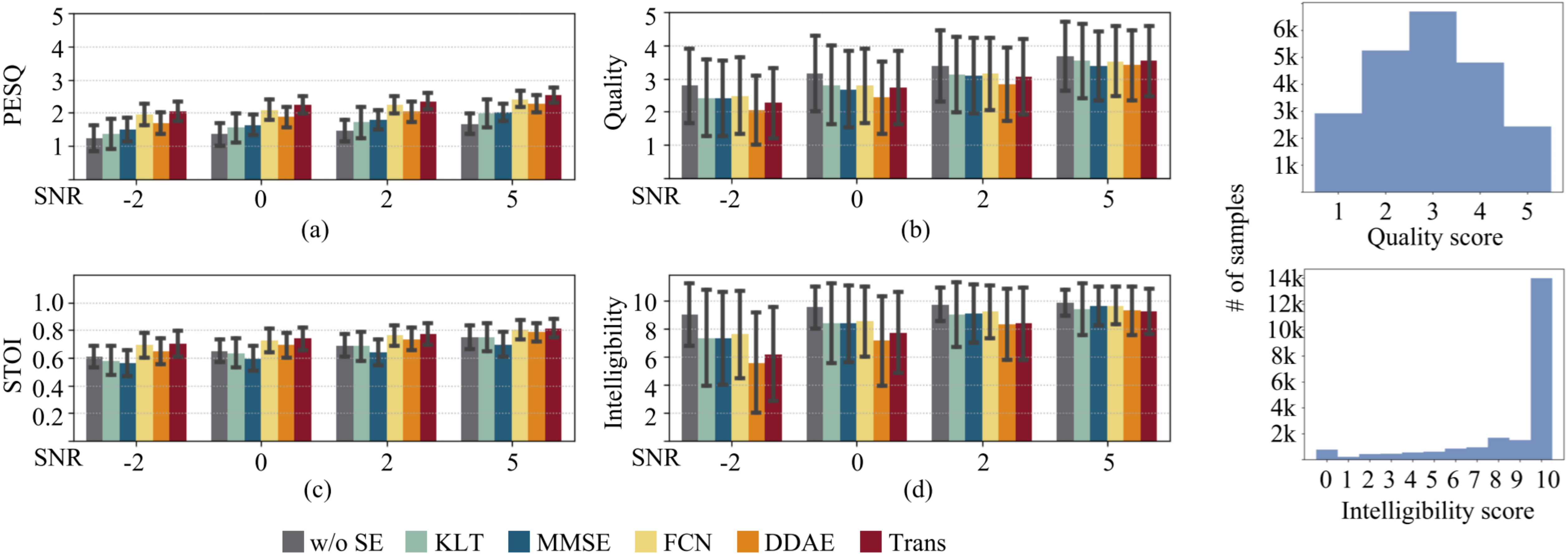


Figure 1: *Comparison between objective assessment metrics and the listening test (left), and histograms of the quality scores and the intelligibility scores (right).*

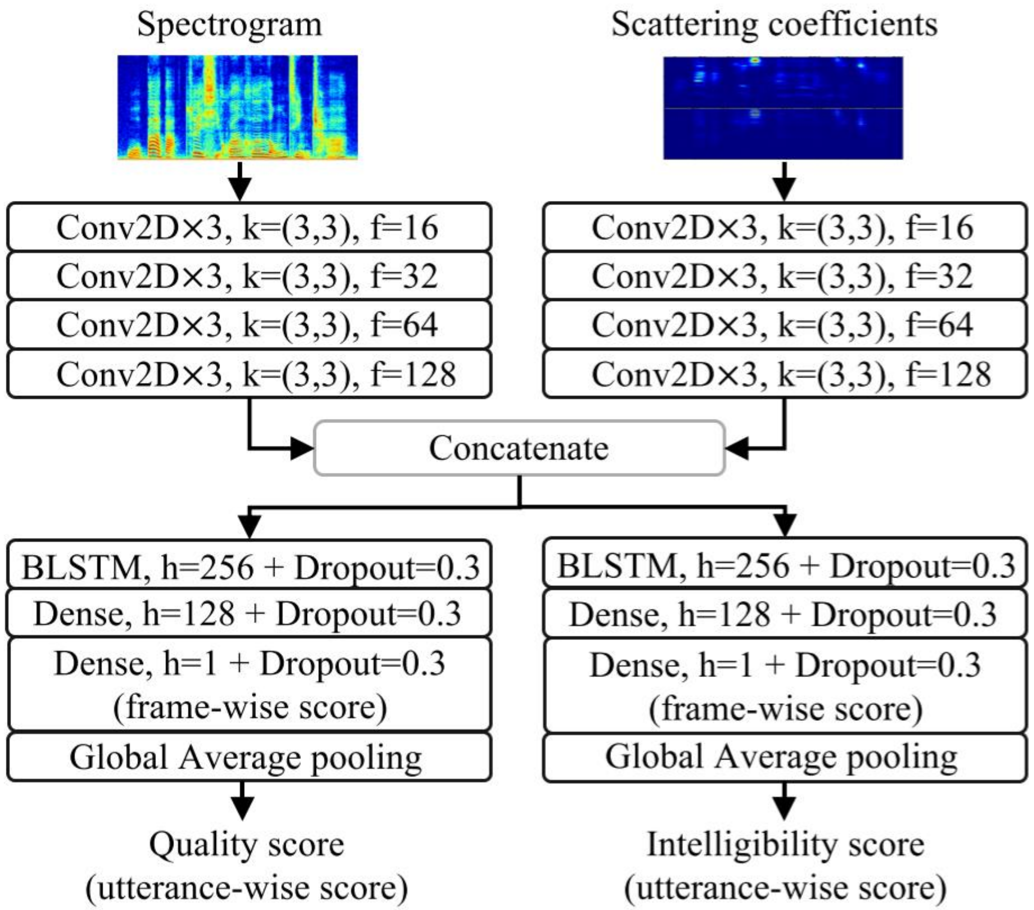
InQSS: a speech intelligibility and quality assessment model using a multi-task learning network (Interspeech 2022)

* Academia Sinica, Taiwan

- # Method

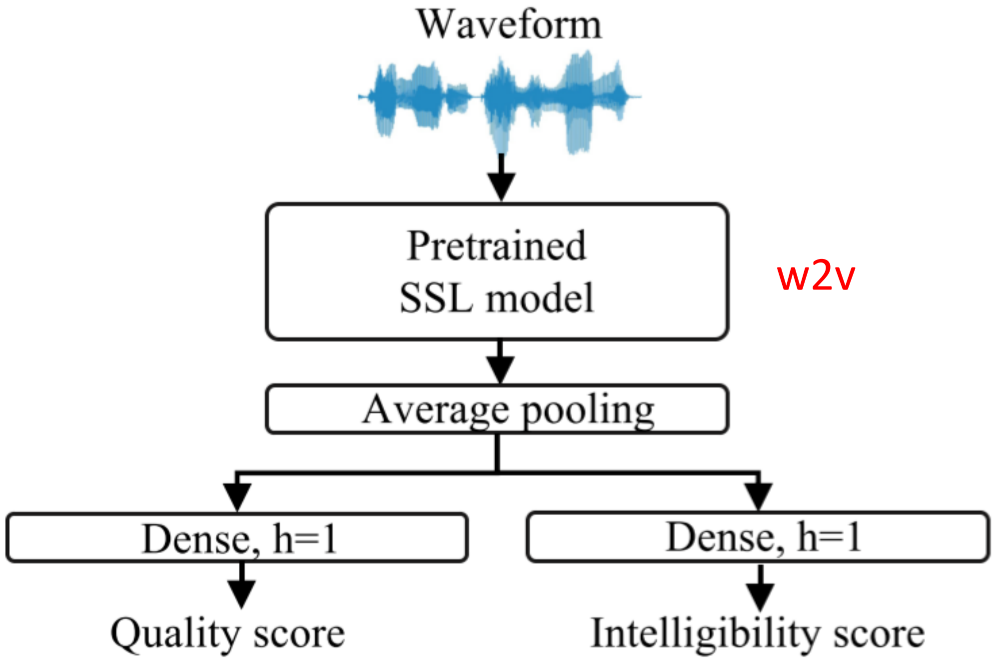
小波散射变换

wavelet scattering transform



train with L2 loss

InQSS-MOSNet



w2v

train with L1 loss

InQSS-SSL

- Experiments on intelligibility
  - scattering coefficients are more useful for intelligibility prediction than spectrograms
  - multi-task is better than only one task
  - combine mos with ssl is best

Model / Method	Input	MSE	PCC	SRCC	
In-MOSNet-S_{I}	spec	2.562	0.695	0.610	1
In-MOSNet-S_{II}	scat	2.425	0.708	0.633	2
In-MOSNet-S_{III}	spec+scat	2.393	0.714	0.642	3
InQSS-MOSNet	spec+scat	2.117	0.755	0.682	4
In-SSL	wav	2.571	0.749	0.645	5
InQSS-SSL	wav	2.552	0.754	0.664	6
In-MOSSSL	wav spec+scat	2.015	0.777	0.668	7
InQSS-MOSSSL	wav spec+scat	**2.017**	**0.791**	**0.700**	8
STOI [1]	-	5.573	0.482	0.461	9
Google-ASR	-	7.305	0.710	0.679	10

InQSS: a speech intelligibility and quality assessment model using a multi-task learning network (Interspeech 2022)

* Academia Sinica, Taiwan

• Experiments on quality

- multi-task is better than only one task
- the SSL model in [35] was trained on a much larger speech quality dataset than the TMHINT-QI, and therefore has a better generalizability than our model
- MSE evaluation results are inconsistent with the PCC and SRCC results on out-of-domain datasets

Model	Dataset	MSE	PCC	SRCC	
Q-MOSNet*	TMHINT-QI	0.439	0.753	0.698	1
InQSS-MOSNet*	TMHINT-QI	0.422	0.763	0.715	2
Q-SSL*	TMHINT-QI	0.388	0.794	0.750	3
InQSS-SSL*	TMHINT-QI	0.365	0.800	0.754	4
InQSS-MOSSSL*	TMHINT-QI	**0.353**	**0.804**	**0.759**	5
DNSMOS [9]	TMHINT-QI	0.915	0.496	0.311	6
NISQA [40]	TMHINT-QI	3.140	0.529	0.348	7
SSL [35]	TMHINT-QI	4.417	0.574	0.405	8
Q-MOSSSL	DAPS	1.261	0.617	0.599	9
InQSS-MOSSSL	DAPS	1.100	0.639	0.639	10
DNSMOS [9]	DAPS	0.665	0.515	0.510	11
NISQA [40]	DAPS	0.663	0.519	0.389	12
SSL [35]	DAPS	**0.475**	**0.710**	**0.718**	13

- Conclution
  - a multitask learning network can improve the performance of a single task without increasing the model complexity
  - SSL-based models can achieve high performance on multi-task speech assessment and require less time to convergence than the training-from-scratch models
  - a simple ensemble approach, averaging the final predictions of two models, can effectively improve the results

Conclution

- PESQ (and other similarity-depend metrics) is not a good idea.
- Non-Intrusive metrics still didn’t show their reliability.
- Human hearing test is the best.

Reference

- Audio Similarity is Unreliable as a Proxy for Audio Quality
  - https://www.isca-speech.org/archive/interspeech_2022/manocha22_interspeech.html
- MOSNet: Deep Learning based Objective Assessment for Voice Conversion
  - http://arxiv.org/abs/1904.08352
- InQSS: a speech intelligibility and quality assessment model using a multi-task learning network
  - https://www.isca-speech.org/archive/interspeech_2022/chen22i_interspeech.html
- PPT from HLT 王卉