

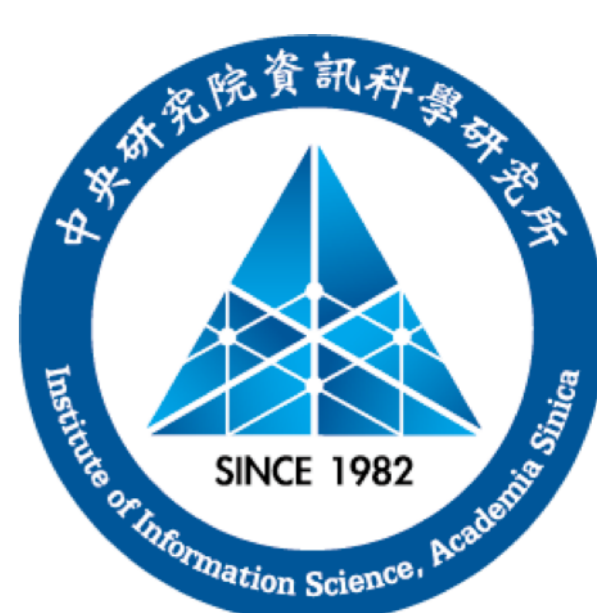
MOSNet: Deep Learning-based Objective Assessment for Voice Conversion

Chen-Chou Lo¹, Szu-Wei Fu², Wen-Chin Huang¹, Xin Wang³, Junichi Yamagishi³, Yu Tsao²,
Hsin-Min Wang³

¹ Institute of Information Science, Academia Sinica, Taiwan

² Research Center for Information Technology Innovation, Academia Sinica, Taiwan

³ National Institute of Informatics, Japan



ABSTRACT

• We propose a deep learning-based assessment model, termed MOSNet, to predict human mean opinion score of converted speech.

▶ The model is trained and tested on large-scale listening test results of the **Voice Conversion Challenge 2018**.

▶ Results show the predicted scores are **highly correlated (0.957)** with human MOS ratings at the **system-level** while **fairly correlated (0.642)** at the **utterance-level**.

• Meanwhile, we've modified MOSNet to predict the similarity scores.

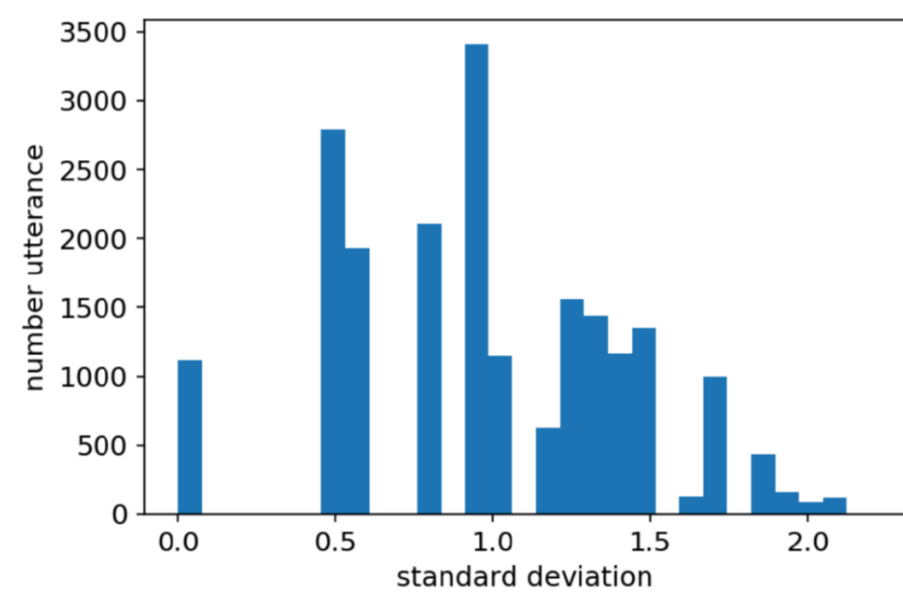
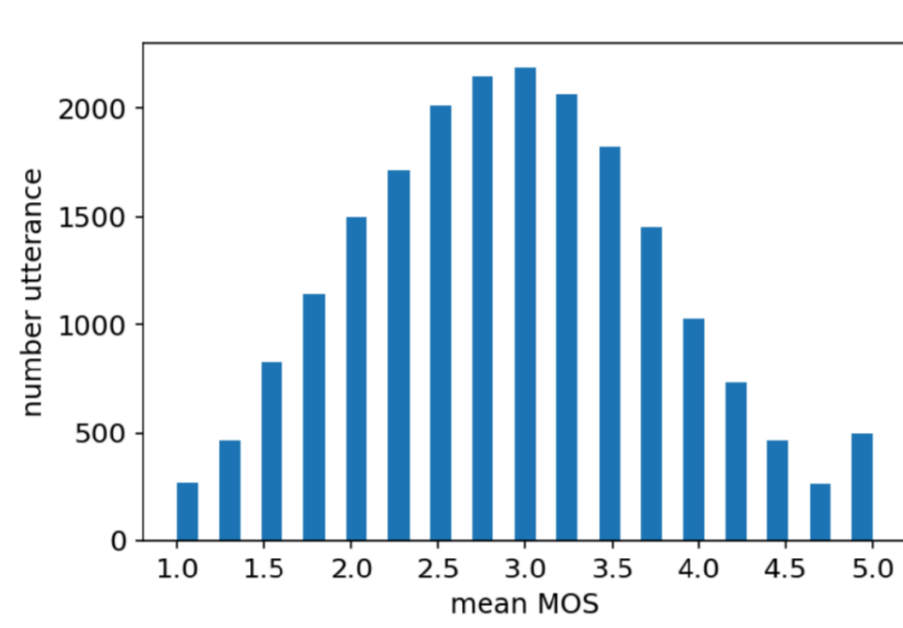
▶ The preliminary results show a **fairly correlation to human ratings**.

DATA & EVALUATION

• Listening test results of VCC 2018

▶ 113,168 human evaluations rated by 267 listeners: **82,304 for naturalness** and **30,864 for similarity**.

▶ 82,304 naturalness evaluations for 20,576 submitted utterances, each rated by 4 listeners, the average score as ground-truth.



• Data distribution and predictability

▶ We use **bootstrap estimation with 1,000 replications** to estimate the intrinsic predictability of the dataset.

▶ For each replication, **half of the listeners were randomly sampled** to measure the mean MOS, which was **compared to the mean MOS of the whole set**.

▶ This analysis shows that the MOS at the utterance-level can be predicted only up to a certain extent, but **the MOS at the system-level is predictable**.

Level	LCC	SRCC	MSE
Utterance	0.805	0.806	0.396
System	0.994	0.978	0.005

• Loss Function

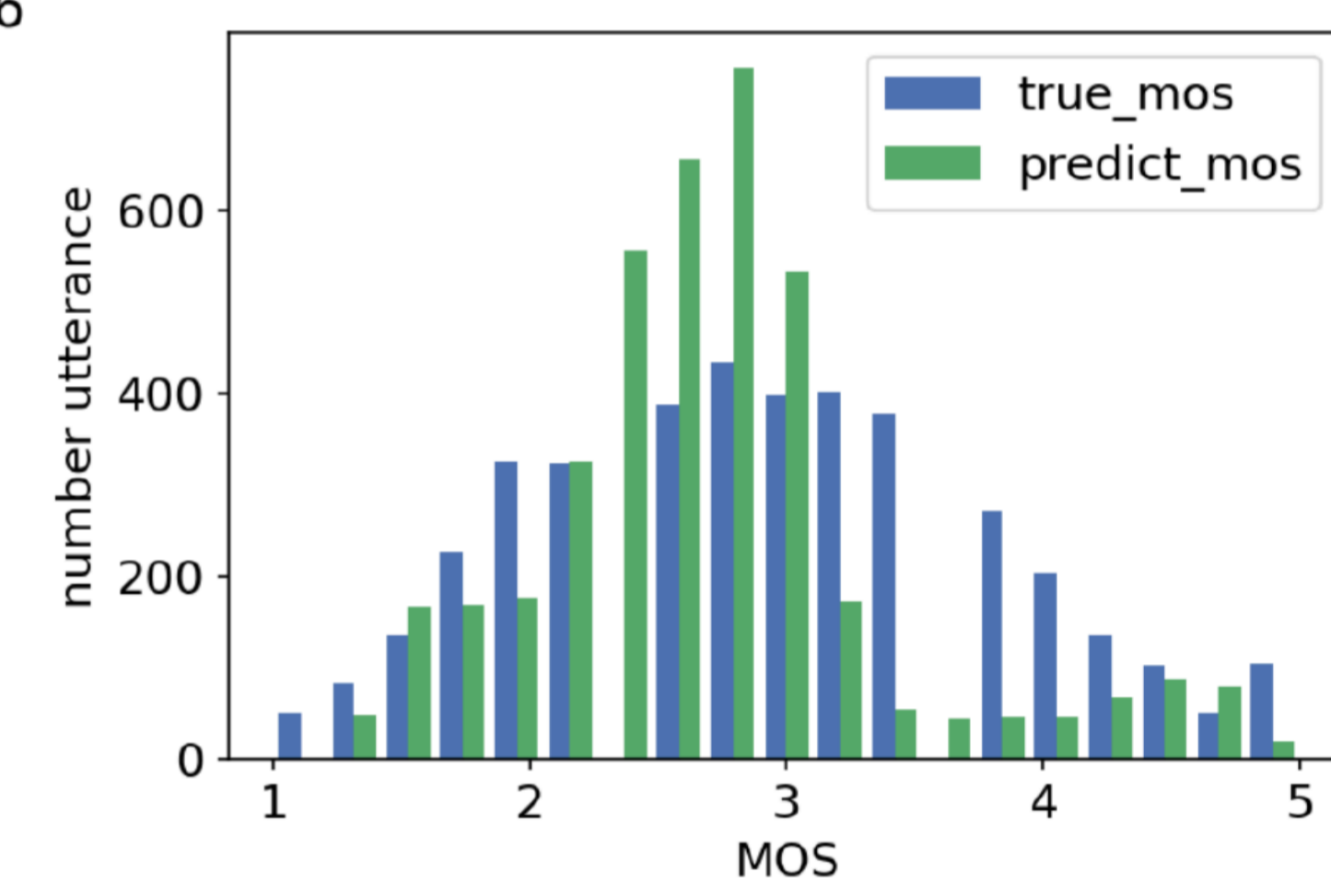
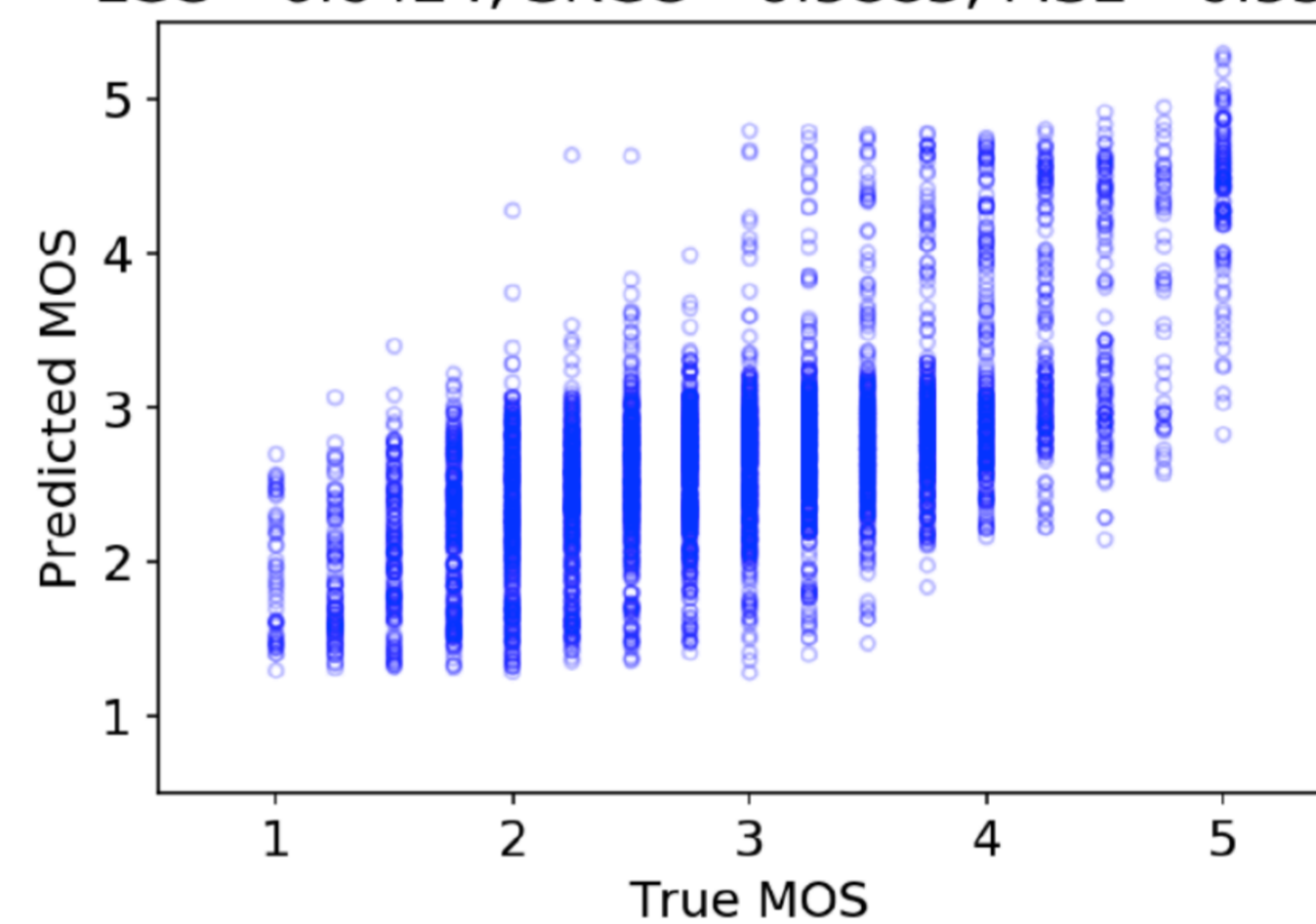
$$O = \frac{1}{S} \sum_{s=1}^S [(\hat{Q}_s - Q_s)^2 + \frac{\alpha}{T_s} \sum_{t=1}^{T_s} (\hat{Q}_s - q_{s,t})^2]$$

EXPERIMENTS & RESULTS

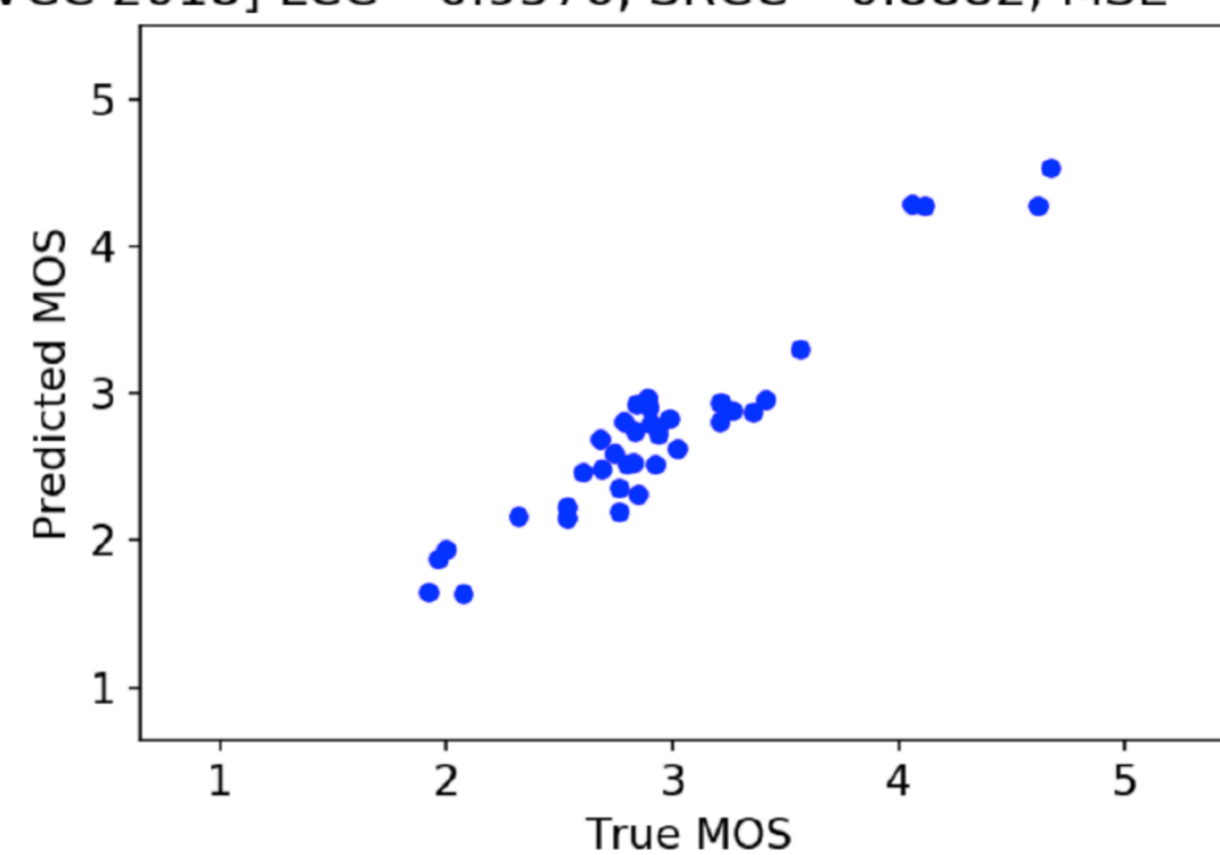
• Utterance-level and system-level prediction results for different models

Model _{batchsize}	utterance-level			system-level		
	LCC	SRCC	MSE	LCC	SRCC	MSE
BLSTM ₇ [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 ₇	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 ₇	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

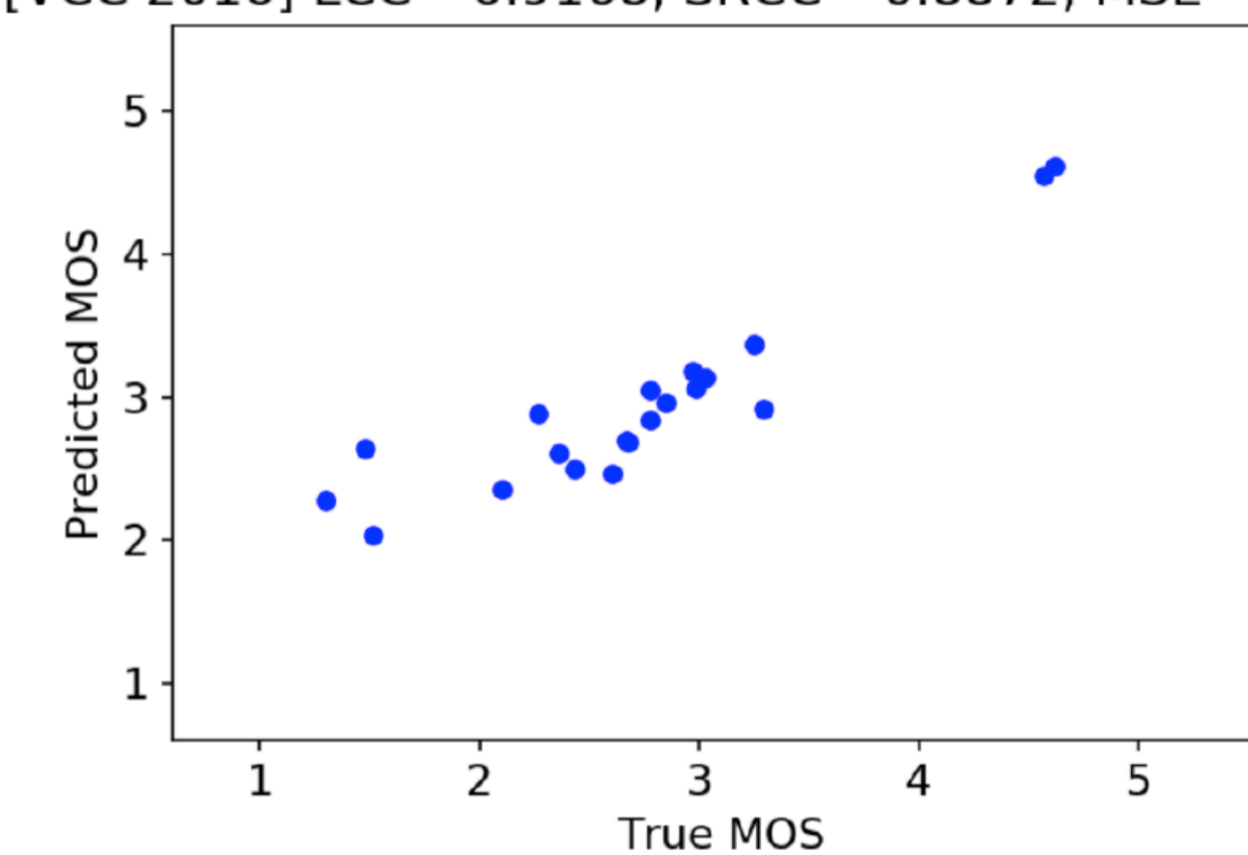
LCC= 0.6424, SRCC= 0.5885, MSE= 0.5376



[VCC 2018] LCC= 0.9570, SRCC= 0.8882, MSE= 0.083

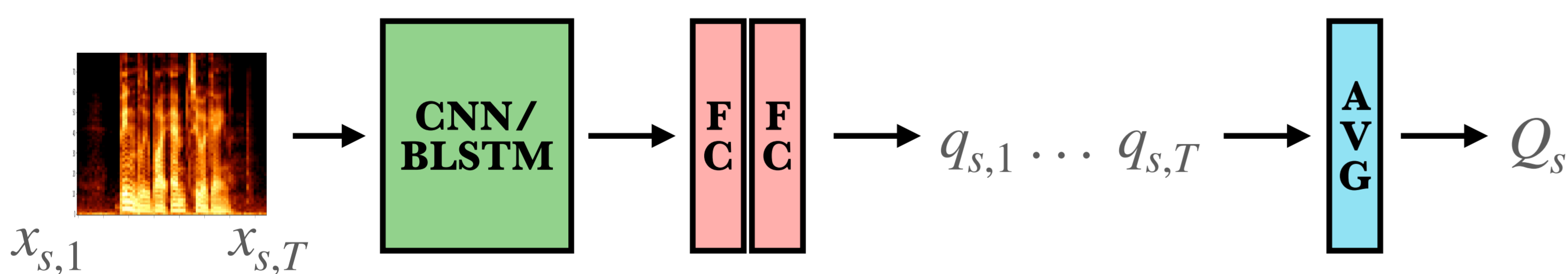


[VCC 2016] LCC= 0.9168, SRCC= 0.8872, MSE= 0.171



MODEL

• Model



model	BLSTM	CNN	CNN-BLSTM
input layer	input ($N \times 257$ mag spectrogram)		
conv. layer	$\left\{ \begin{array}{l} conv3 - (channels)/1 \\ conv3 - (channels)/1 \\ conv3 - (channels)/3 \end{array} \right\} \times 4$ $channels = [16, 32, 64, 128]$		
recurrent layer	BLSTM-128		BLSTM-128
FC layer	FC-64, ReLU, dropout	FC-64, ReLU, dropout	FC-128, ReLU, dropout
	FC-1 (<i>frame-wise scores</i>)		
output layer	average pool (<i>utterance score</i>)		

• Similarity prediction results of the modified MOSNet model

model	Level	ACC	LCC	SRCC	MSE
CNN (scalar)	utterance	0.696	0.453	0.455	0.197
	system	0.701	0.394	0.395	0.195
CNN (2 classes)	utterance	0.670	0.329	0.329	0.336
	system	0.674	0.292	0.292	0.326

CONCLUSION

- We presented a **deep learning-based quality assessment model, MOSNet**, for the Voice Conversion task.
- Experimental results show that MOSNet yields predictions with a **high correlation to human ratings at the system-level** while a fair correlation at the utterance-level.
- Modified MOSNet can fairly predict the similarity scores of the converted speech relative to the target speech.
- To the best of our knowledge, MOSNet is the **first speech objective assessment model for VC**.