Incorporating Embedding Vectors from a Human Mean-Opinion Score Prediction Model for Monaural Speech Enhancement

Khandokar Md. Nayem and Donald S. Williamson

Department of Computer Science, Indiana University, IN, USA

LUDDY
SCHOOL OF INFORMATICS,
COMPUTING, AND ENGINEERING

INTERSPEECH 2021

Motivation

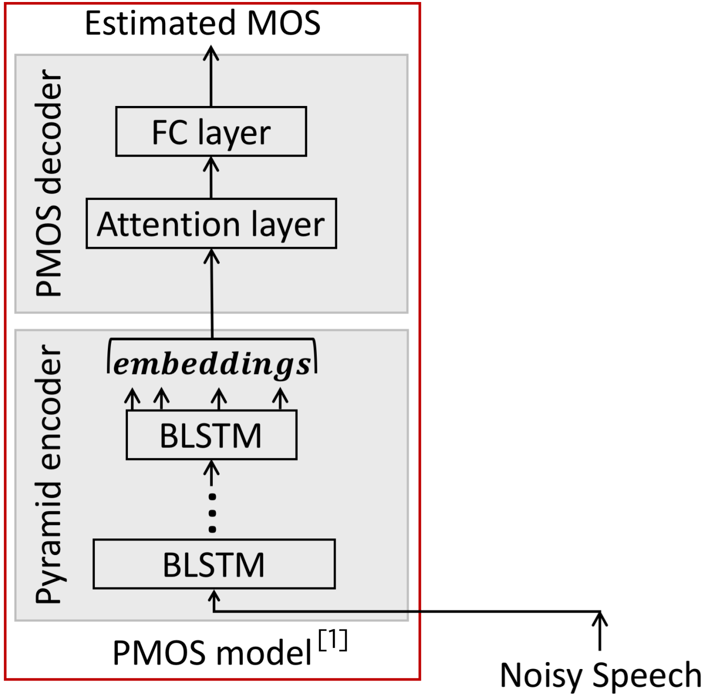
Speech Enhancement (SE) systems target maximization of speech quality and intelligibility measured by various proposed **objective functions**.

Current speech quality objective functions are often not strongly correlated with **human subjective evaluations**.

Automatic speech assessment that measures the subjective score of enhanced speech can help SE systems to estimate better perceptual quality speech.

Speech Quality Assessment Model

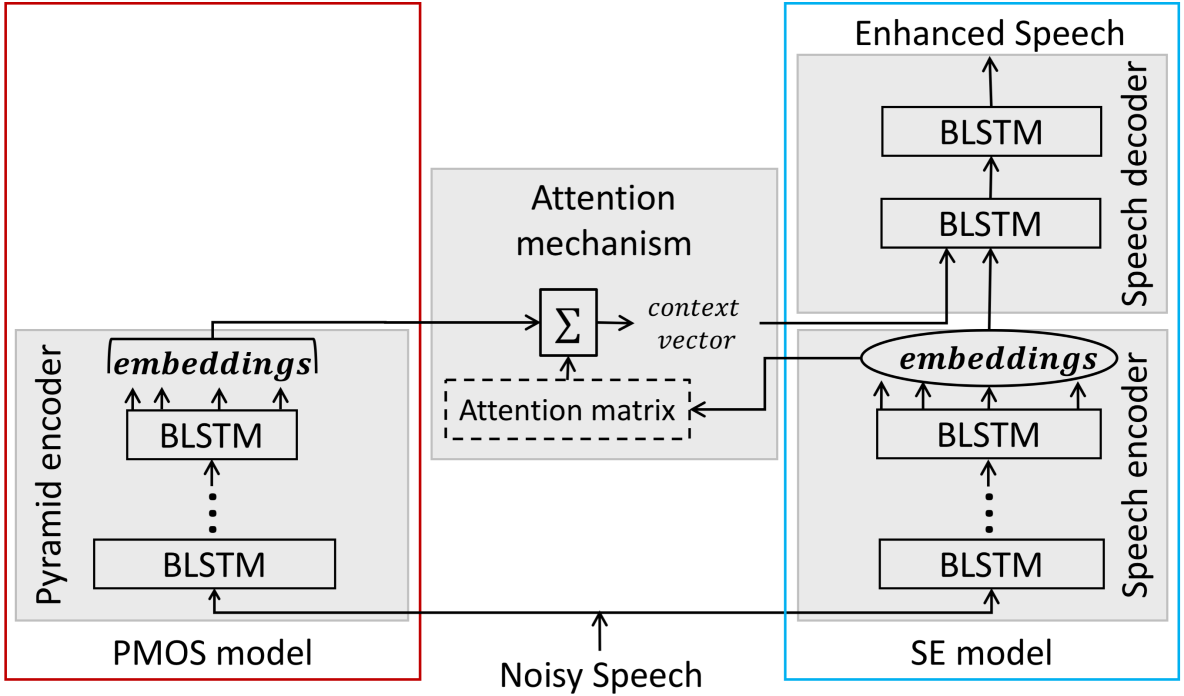
PMOS model



[1] X. Dong et. al, “A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals,” in Proc. Interspeech, 2020.

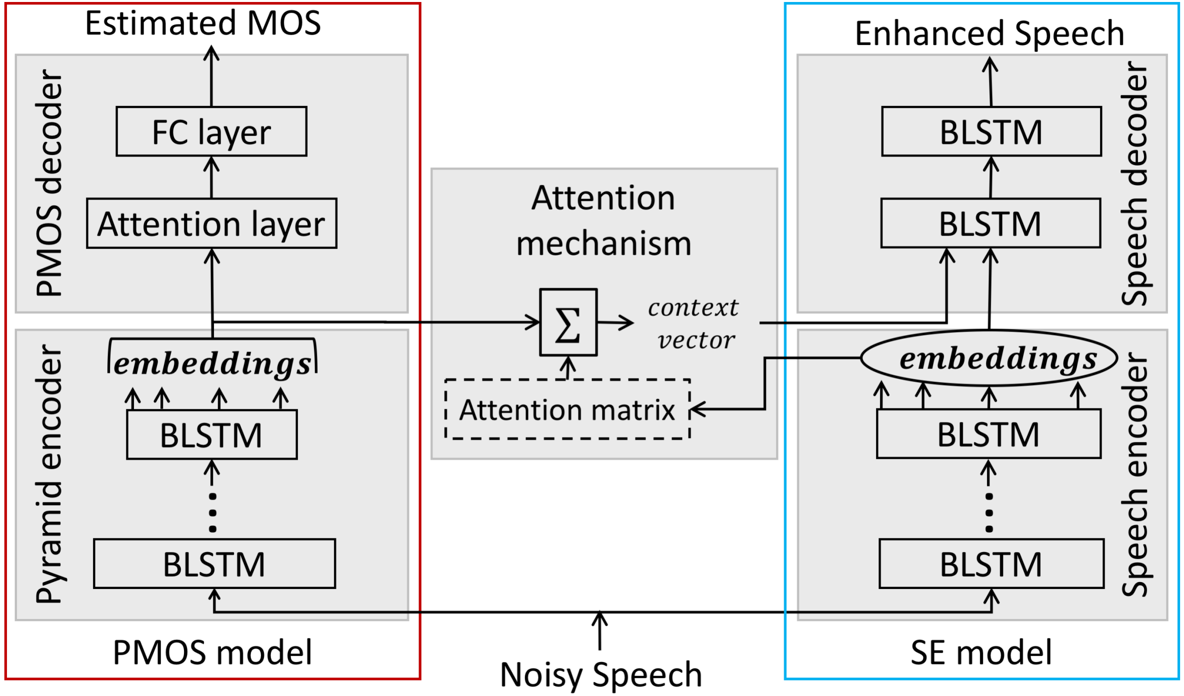
Attention-based SE Model

SE model



Joint-learning SE Model

SE + PMOS model



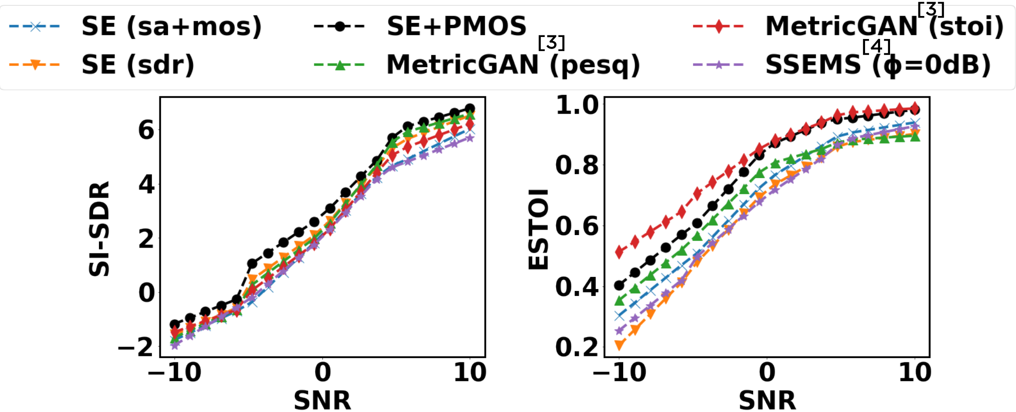
Results

Table 1: Performance comparison with MOS prediction models.

	MAE	RMSE	PCC	SRCC
NISQA[2]	0.62	0.7	0.71	0.79
PMOS [1]	0.51	0.57	0.88	0.88
SE+PMOS (proposed)	0.45	0.52	0.9	0.91

PCC = Pearson's correlation coefficient,
SRCC = Spearman's rank correlation coefficient,
SI-SDR = Scale-invariant signal-to-distortion ratio,
ESTOI = extended short-time objective intelligibility,

Figure 1: Average (a) SI-SDR, (b) ESTOI performance of SE models on test speech in different SNRs.



(a) SNR vs SI-SDR (b) SNR vs ESTOI

[1] X. Dong et. al, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in Proc. Interspeech, 2020.
[2] G. Mittag et. al, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in Proc. ICASSP, 2019.
[3] S.-W. Fu et. al, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in Proc. ICML, 2019.
[4] R. E. Zezario et. al, "Specialized speech enhancement model selection based on learned non-intrusive quality assessment metric." in Proc. Interspeech, 2019.

Results

Table 2: Average results of the SE models in different performance metrics.

		COSINE				VOiCES			
	Loss function	**PESQ**	**SI-SDR**	**ESTOI**	**MOS-LQO**	**PESQ**	**SI-SDR**	**ESTOI**	**MOS-LQO**
Mixture	-	1.46	0.53	0.62	4.04	1.26	-1.3	0.48	2.74
SE	mse	2.68	2.8	0.8	3.2	2.3	1.2	0.69	3.5
	mos	2.8	3.8	0.82	4.2	2.37	1.66	0.74	5.3
	mse, sa	2.72	3.1	0.82	4	2.35	1.6	0.7	3.8
	sa, mos	2.89	4.1	0.85	4.4	2.42	1.72	0.77	5.7
	sdr	2.7	4.5	0.82	4	2.32	2.01	0.72	4.5
SE+PMOS (proposed)	mse	3.1	4	0.85	4.2	2.48	1.8	0.8	6
	mse, sa	3.19	**4.6**	0.93	4.8	2.54	**2.08**	0.86	6.3
	mse, sa, mos	3.19	4.5	0.92	**5.1**	2.53	2.06	0.84	**6.5**
MetricGAN[3]	pesq	**3.28**	4.4	0.9	5	**2.67**	2.01	0.83	6.1
	stoi	3.19	4.3	**0.94**	4.8	2.5	2	**0.87**	5.8
SSEMS[4]	qnet (ϕ=0dB)	2.85	2.9	0.83	3	2.4	1.8	0.7	2.8

Conclusion

Our proposed **speech enhancement** model utilizes a speech quality **MOS assessment metric** in a joint learning manner.

Results show that proposed **SE+PMOS** model outperforms other models in different noisy environments.

We evaluate our model's subjective score using an **MOS- estimation model**.

Our assessment model provides **utterance-level feedback**, which may be sub-optimal since the model's embeddings are calculated at the frame level.

Thank You

Khandokar Md. Nayem

knayem@iu.edu

Donald S. Williamson

williads@indiana.edu

Department of Computer Science
Audio, Speech and Information Retrieval (ASPIRE) lab
https://aspire.sice.indiana.edu/
Indiana University, IN, USA