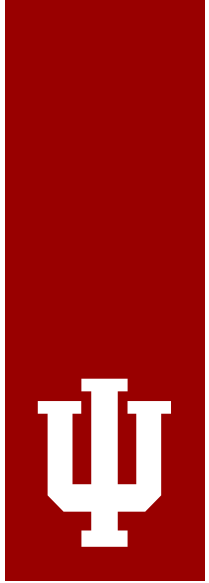


Incorporating Embedding Vectors from a Human Mean-Opinion Score Prediction Model for Monaural Speech Enhancement

Khandokar Md. Nayem and Donald S. Williamson

Department of Computer Science, Indiana University, IN, USA

LUDDY
SCHOOL OF INFORMATICS,
COMPUTING, AND ENGINEERING

INTERSPEECH 2021

Introduction

Monaural speech enhancement is a challenging problem that aims to remove unwanted noise from a target speech signal.

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**.

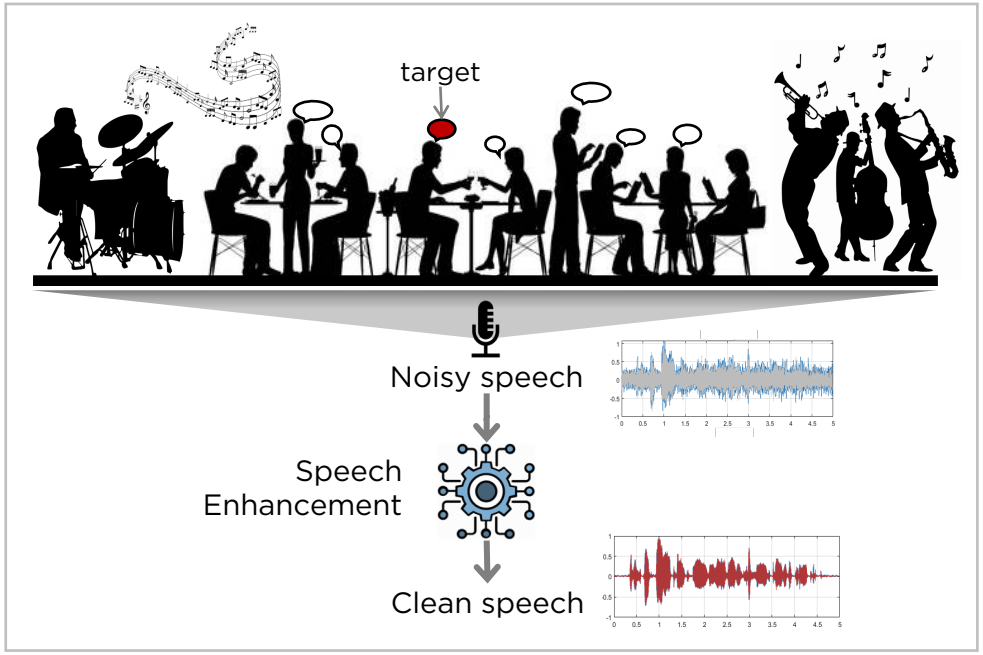


Image source, https://clipground.com/

Speech Objective Functions

Short-Time Spectral-Amplitude Mean square error (STSA-MSE) (Ephraim et al. 1984)

Short-time objective intelligibility (STOI) [1]

Fu et al. 2018 jointly optimize with STOI and MSE, which improves speech intelligibility.

Zhang et al. 2018 apply gradient approximation and Koizumi et al. 2018 use a policy gradient method.

Kolbæk et al. 2019 compares MSE and STOI as loss function.

Perceptual evaluation of speech quality (PESQ) [2]

Fu et al. 2018 report that PESQ does not been increase when optimizing with STOI.

Martin-Donas et al. 2018 propose a PESQ-inspired objective function.

Fu et al. 2018 formulate Quality-Net for non-intrusive PESQ estimation, which is used to enhance speech as a maximization criteria in Fu et al. 2019 and as a model selection parameter in Zezario et al. 2019.

Signal-to-distortion ratio (SDR) [3]

Kawanaka et al. 2010 perform SDR as objective cost function.

[1] C. H. Taal et al., "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE/ACM TASLP, 2011.
[2] A. W. Rix et al., "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in Proc. ICASSP, 2001.
[3] C. Fe´vote et al., "Bss eval toolbox user guide–revision 2.0," 2005.

Motivation

Rix et al. 2006 & Emiya et al. 2011 show that optimizing with objective measures of success is not always optimal since they do not always strongly correlate with subjective measures.

Subjective evaluation – Mean opinion score (MOS)

Patton et al. 2016, Avila et al. & Lo et al. 2019 et al. proposed human-assessed MOS model separately.

Dong et al. 2020 formulate another MOS estimate model in real-world environments.

Joint learning

Speech estimation with other training targets are successful, e.g. Donahue et al. 2018 (speech recognition), Lee et al. 2019 (phase response), Schulze-Forster et al. 2020 (phoneme class), and Ji et al. 2020 (speaker identification).

In a similar manner, joint learning of speech enhancement task and speech-quality estimation can provide better speech quality acoustically and perceptually.

Proposed approach

We propose a joint learning scheme that estimates a speech quality MOS score and enhances the speech signal.

MOS estimation model produces encoded embedding vectors that extract perceptually useful features that are important for human-based assessment.

Proposed **speech enhancement model** is conditioned on that embedding vector and enhances the noisy speech using a separate encoder- decoder framework.

Our proposed model jointly updates both the MOS-prediction and speech-enhancement models during training, using speech enhancement and MOS prediction loss terms.

Problem Formulation

In the time domain, $m_t = s_t + n_t$

In the time-frequency (T-F) domain, $\mathbf{M} = |\boldsymbol{M}|e^{i\boldsymbol{\theta}^M}$

$\boldsymbol{S} = |\boldsymbol{S}|e^{i\boldsymbol{\theta}^S}$

$|\boldsymbol{S}|$ → clean speech magnitude

$\boldsymbol{\theta}^S$ → clean speech phase

$\boldsymbol{N} = |\boldsymbol{N}|e^{i\boldsymbol{\theta}^N}$

$|\boldsymbol{N}|$ → noise magnitude

$\boldsymbol{\theta}^N$ → noise phase

m_t → noisy speech
s_t → clean speech
n_t → noise
t → time index

$\boldsymbol{M}$ → T-F noisy speech
$|\boldsymbol{M}|$ → magnitude response
$\boldsymbol{\theta}^M$ → phase response
f → frequency index

Estimate clean speech, $|\hat{\boldsymbol{S}}| = \mathcal{F}(|\boldsymbol{M}|)$ with noisy phase $\boldsymbol{\theta}^M$.

Speech Quality Assessment Model

We adopt a data-driven MOS prediction model proposed by Dong et al. 2020.

MOS prediction model consists of an attention-based encoder-decoder structure that uses stacked pyramid bi-directional long-short term memory (pBLSTM) networks in the encoder. We denote this model as Pyramid-MOS (PMOS).

The output of a pBLSTM node is an embedding vector, h_t^l, that is as defined below:

$$h_t^l = pBLSTM\left(h_{t-1}^l, \left[h_{\Upsilon\times t-\Upsilon+1}^{l-1}, h_{\Upsilon\times t}^{l-1}\right]\right)$$

where Υ is the reduction factor between successive pBLSTM layers and l is the layer number.

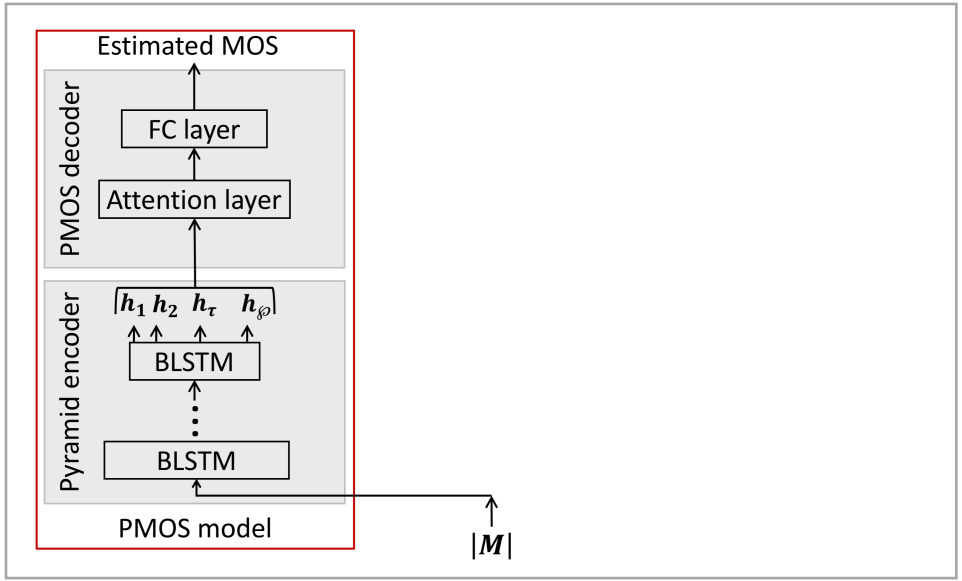


Fig. Proposed network for speech quality assessment (PMOS model).

Speech Quality Assessment Model

The encoder output is generated by concatenating the hidden states of the last pBLSTM layer into vector $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_\tau, \cdots, \boldsymbol{h}_\wp\}$, where $\wp$ is the total number of final embedding vectors with index τ.

The decoder of the PMOS model is implemented as an attention layer followed by a fully-connected (FC) layer and outputs estimated MOS of input speech.

Self-attention mechanism uses pyramid encoder output at the i-th and k-th time steps to compute attention weights, $\alpha_{i,k}^{PMOS}$ and context vector c_i^{PMOS},

$$\alpha_{i,k}^{PMOS} = \frac{exp(\boldsymbol{h}_i^\dagger \boldsymbol{Q} \boldsymbol{h}_k)}{\sum_{i=1}^{\wp} exp(\boldsymbol{h}_i^\dagger \boldsymbol{Q} \boldsymbol{h}_k)}$$

$$c_i^{PMOS} = \sum_{k=1}^{\wp} \alpha_{i,k}^{PMOS} \cdot \boldsymbol{h}_k$$

where $\boldsymbol{Q}$ is the PMOS attention weight matrix.

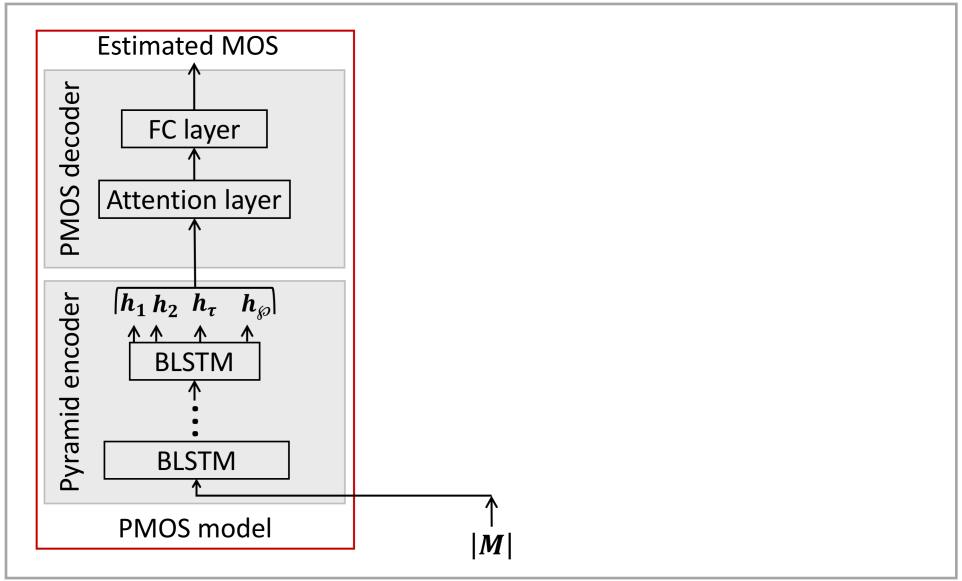


Fig. Proposed network for speech quality assessment (PMOS model).

Attention-based SE Model

Proposed SE model is also an encoder-decoder structure. SE encoder takes a single time-frame of a noisy-speech mixture, $|\boldsymbol{M}_t|$, as input and multiple BLSTM layers, are stacked together to create a hidden representation of the frame, $\boldsymbol{g}_t$.

An attention mechanism is applied using the mixture encoding from the SE model, $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_T\}$, and the PMOS encoding, $\boldsymbol{H}$, from the MOS prediction model.

We compute a score for each embedding vector $\boldsymbol{h}_\tau^l$ using a learnable weight matrix, $\boldsymbol{W}$.

$$score_{t,\tau} = \boldsymbol{g}_t^T \boldsymbol{W} \boldsymbol{h}_\tau$$

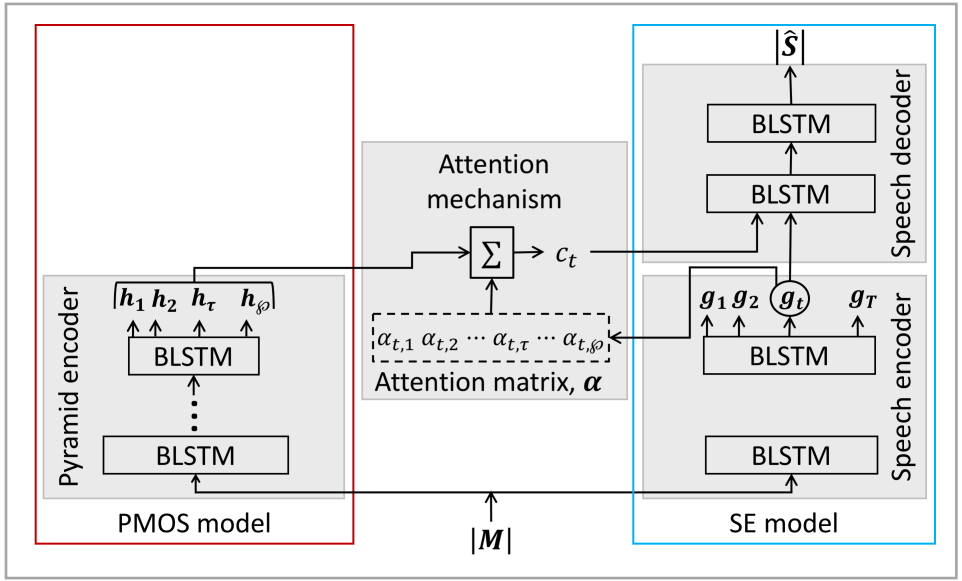


Fig. Proposed network for speech enhancement (SE model).

Attention-based SE Model

Attention weights for SE model $\alpha_{t,\tau}$ forms a context vector $\boldsymbol{c}_t$ for each mixture frame. Prior computing $\boldsymbol{c}_t$, $\boldsymbol{h}_\tau^L$ passes through a linear layer l using below equations:

$$\alpha_{t,\tau} = \frac{exp\left(score_{t,\tau}\right)}{\sum_{t=1}^{T} exp\left(exp\left(score_{t,\tau}\right)\right)}$$

$$\boldsymbol{c}_t = \sum_{\tau=1}^{\wp} \alpha_{t,\epsilon} \cdot l(\boldsymbol{h}_\tau)$$

Context vector and SE-model embedding vector are concatenated (e.g., $[\boldsymbol{c}_t, \boldsymbol{g}_t]$) and passed to the decoder module.

Decoder consists of a linear layer with a $\tanh(\cdot)$ activation function, two BLSTM layers, and a linear layer with ReLU. It outputs the estimated enhanced speech $|\hat{\boldsymbol{S}}|$.

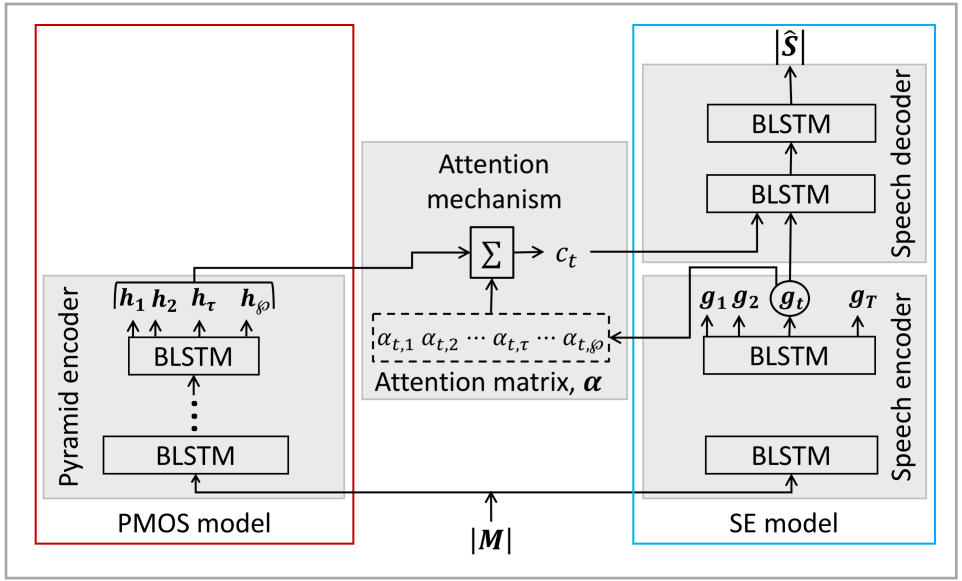


Fig. Proposed network for speech enhancement (SE model).

Joint-learning SE Model

Our joint-learning objective function uses a weighted average of a signal-approximation loss $\mathcal{L}_{sa}$ (from the SE model), the MSE of the magnitude spectrum $\mathcal{L}_{mse}$ (from the SE model) and the MSE of the MOS estimation $\mathcal{L}_{mos}$ (from the PMOS model).

With hyper-parameters λ_1 and λ_2, the overall loss function:

$$\mathcal{L} = \lambda_1[\lambda_2\mathcal{L}_{mse} + (1-\lambda_2)\mathcal{L}_{sa}] + (1-\lambda_1)\mathcal{L}_{mos}$$

We first train the PMOS model using $\mathcal{L}_{mos}$ (e.g. $\lambda_1 = 0$), then we train the SE model using $\lambda_1 = 1$, while running the PMOS model in inference mode.

Finally, we train both models jointly using $\mathcal{L}$ to further reduce any correctional differences between the true MOS and estimated MOS in the PMOS model, and to increase perceptual quality of the enhanced speech.

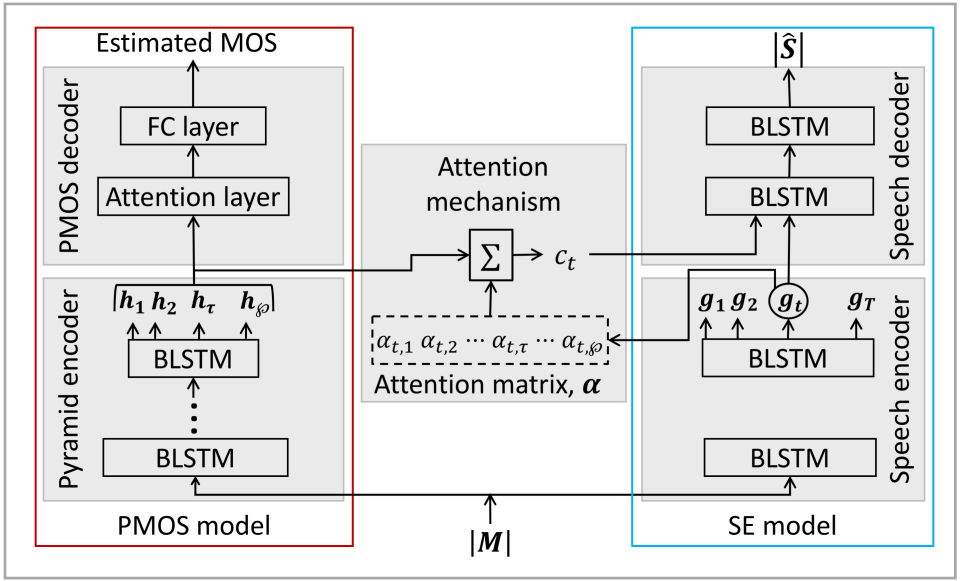


Fig. Proposed network for speech enhancement & speech quality assessment (SE+PMOS model).

Dataset

Train and evaluate using COnversational Speech In Noisy Environments (COSINE)[4] and Voices Obscured in Complex Environmental Settings (VOiCES)[5] speech corpora.

> COSINE contains 150 hours audio, which captured using 7-channel wearable microphones, with multiparty conversations in a variety of noisy environments (e.g., street, cafeteria, bus, wind noise, etc). The approximated SNRs range from -10.1 to 11.4 dB.
>
> VOiCES records audio with 12 microphones placed in two rooms with different background noises to capture reverberant-noisy speech. The approximated speech-to-reverberation ratios (SRRs) ranges from -4.9 to 4.3 dB.

MOS data was captured from the listening study that is outlined by Dong et al. 2020, which contains MOS quality ratings for 18,000 COSINE signals and 18,000 VOiCES signals.

Noisy or reverberant stimuli of each dataset are divided into training (70%), validation (10%), and testing (20%) sets, and trained separately.

[4] A. Stupakov et al., “Cosine-a corpus of multi-party conversational speech in noisy environments,” in Proc. ICASSP. IEEE, 2009.
[5] C. Richey et al., “Voices obscured in complex environmental settings (voices) corpus,” arXiv preprint arXiv:1804.05053, 2018.

Experimental Setup

PMOS model

PMOS encoder uses $L = 3$ pBLSTM layers (with 128, 64 and 32 nodes in each direction, respectively) on top of a BLSTM layer that has 256 nodes.

The reduction factor $\Upsilon = 2$ is adopted here. Therefore, the final latent representation h_τ is reduced in the time resolution by a factor of $\Upsilon^3 = 8$.

The context vector is passed to a FC layer with 32 units.

SE model

SE model uses a BLSTM based encoder-decoder architecture, where the encoder consists of 2 BLSTM layers.

Each BLSTM layer contains 200 nodes and the linear layer has 321 nodes.

Input feature vector is the magnitude of the mixture spectrogram computed using a hamming window with 50% overlap after normalization

Results

Table 1: Performance comparison with MOS prediction models.

	MAE	RMSE	PCC	SRCC
NISQA[6]	0.62	0.7	0.71	0.79
PMOS [7]	0.51	0.57	0.88	0.88
SE+PMOS (proposed)	**0.45**	**0.52**	**0.9**	**0.91**

MAE = Mean absolute error,

RMSE = Root mean squared error,

PCC = Pearson's correlation coefficient,

SRCC = Spearman's rank correlation coefficient

[6] G. Mittag et. al, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in Proc. ICASSP, 2019.
[7] X. Dong et. al, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in Proc. Interspeech, 2020.

Results

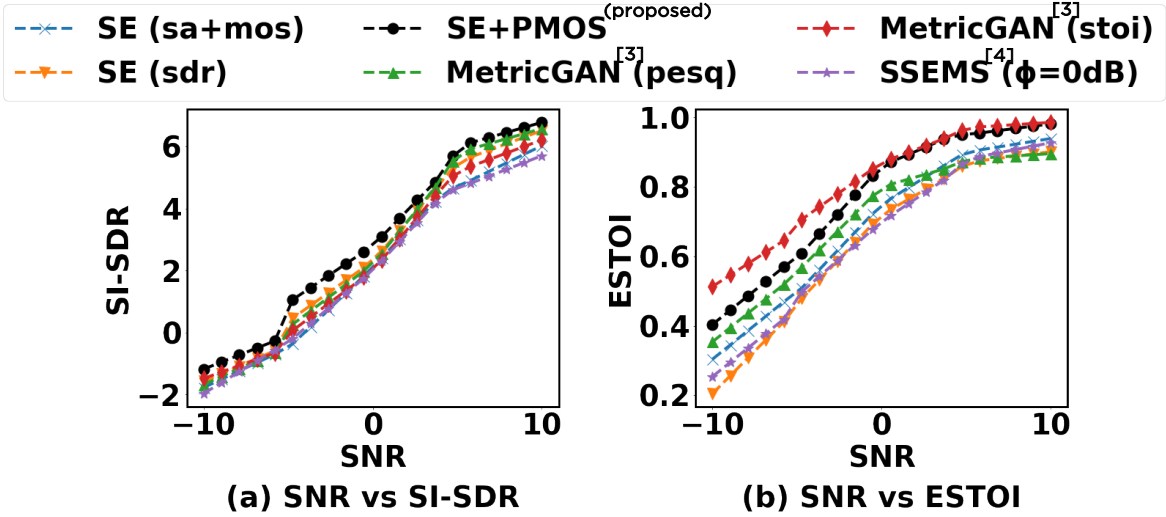
Table 2: Average results of the SE models in different performance metrics.

		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[8]	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[9]	qnet (ϕ=0dB)	2.85	2.9	0.83	3	2.4	1.8	0.7	2.8

[8] S.-W. Fu et. al, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in Proc. ICML, 2019.
[9] R. E. Zezario et. al, "Specialized speech enhancement model selection based on learned non-intrusive quality assessment metric." in Proc. Interspeech, 2019.

Results

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

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