AN END-TO-END NON-INTRUSIVE MODEL FOR SUBJECTIVE AND OBJECTIVE REAL-WORLD SPEECH ASSESSMENT USING A MULTI-TASK FRAMEWORK

Zhuohuang Zhang, Piyush Vyas, Xuan Dong, Donald Williamson
{zhuozhan, piyush, xuandong}@iu.edu, williads@indiana.edu
Indiana University Bloomington

IEEE ICASSP 2021

INDIANA UNIVERSITY
BLOOMINGTON

- **Introduction**
- Real-world speech data
- Proposed Model
- Experimental Comparisons
- Conclusions

Introduction

- Speech assessment is important for evaluating and improving the performance of many applications, including speech separation, VoIP evaluation, etc.

- Subjective ratings provide the most reliable and accurate form of assessment, but it is costly and time-consuming.
- Objective metrics are often used as they are easy to compute and allow quick assessment of large-scale datasets.

Introduction

Two types of objective metrics:

1. Intrusive:

   Requires a clean reference during assessment, such as PESQ (Rix et al., 2001), eSTOI (Jensen and Taal, 2016).
   However, the clean reference is not always available in real-world environments.

2. Non-intrusive:

   Does not require a clean reference for evaluation, such as ITU-T P.563 (Malfait 2006), ANIQUE (Kim 2005).
   Do not always correlate well with subjective ratings (Andersen et al., 2017; Mittag and Moller, 2019).

Introduction

Recent developments on speech assessment metrics mainly based on data-driven (i.e., deep learning) approaches.

Motivation:

- Models such as Quality-Net (Fu et al., 2018) and NISQA (Mittag and Moller, 2019) have been proposed to estimate some objective scores (e.g., PESQ) of the input speech signal. (not direct estimate of subjective ratings)

- Some other approaches (Patton et al., 2016; Avila et al., 2019) have been proposed to predict mean opinion scores (MOS) collected from human listeners using simulated dataset. (does not reflect real-world environments)

Introduction

Contributions of our work:

1. A non-intrusive speech assessment system that estimates both subjective and objective ratings for real-world recorded speech.

2. An end-to-end model is developed by encoding the time-domain speech with a convolution layer rather than the conventional short-time Fourier transform (STFT).

3. Enabling direct comparisons between real-world and laboratory experiments

- Introduction
- **Real-world Speech Data**
- Proposed Model
- Experimental Comparisons
- Conclusions

Real-world Speech Data

Two real-world collected datasets:

COSINE (Stupakov et al., 2018): The speech-to-noise ratios (SNRs) for COSINE are approximately between -10.1 to 11.4 dB.

VOiCES (Richey et al., 2018)The approximated speech-to-reverberation ratios (SRRs) of these signals range from -4.9 to 4.3 dB.

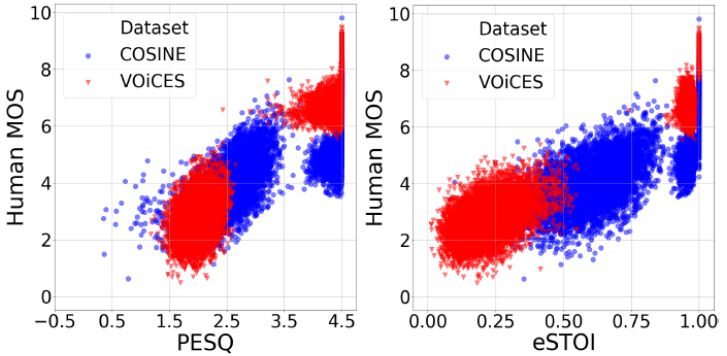


Fig. 1. Correlations between subjective and objective ratings on two real-world corpora (COSINE - blue, VOiCES - red).

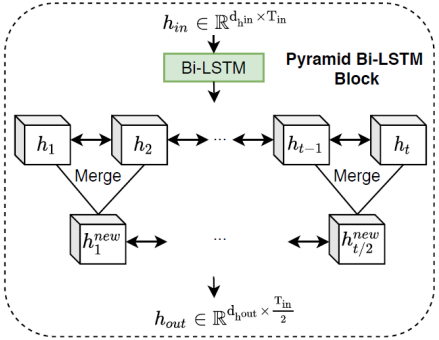
Online listening test for subjective rating (Xuan and Williamson, 2020):

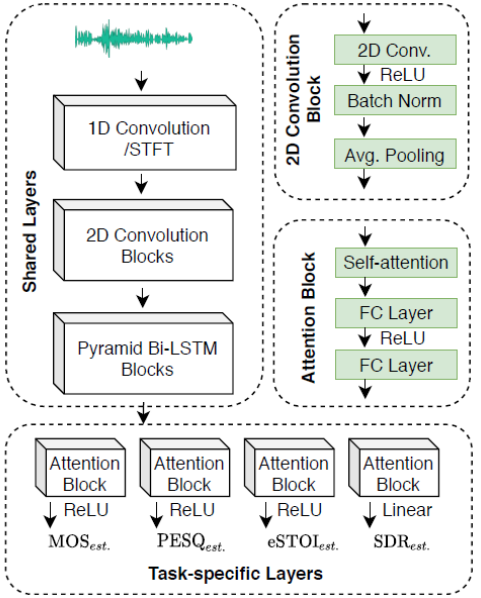
- Amazon Mechanical Turk
- 700 Human intelligence tasks (HITs), each completed by 5 workers (3500 workers in total).
- Following ITU-R BS. 1534. (MUSHRA)
- Ratings on speech quality between 0 to 100.
- 180K responses collected for 36K speech signals (18K signals per dataset).

- Introduction
- Real-world Speech Data
- **Proposed Model**
- Experimental Comparisons
- Conclusions

Proposed Model

Our proposed model consists of two types of layers:





1. Shared-encoder layers:
   - 1D convolution layer to extract features
   - 2D convolution and pBi-LSTM to capture the shared information

2. Task-specific decoding layers:
   - Using self-attention mechanism, the decoder part focuses on different aspects of the learned latent feature and makes task specific estimates of the different metrics (i.e., MOS, PESQ, eSTOI and SDR).

Weighted loss (empirically set to 10, 1, 12 and 0.1):

$$\mathcal{L}_{\text{Model}} = \sum_{k=1}^{K} \alpha_k \mathcal{L}_{\text{MSE}}^{k}$$

- Introduction
- Real-world Speech Data
- Proposed Model
- **Experimental Comparisons**
- Conclusions

Experimental Comparisons

Models evaluated:

1. AMSA: a multi-task model for objective score estimation (Dong and Williamson, 2020).
2. Quality-net (Fu et al., 2018).
3. NISQA (Mittag and Moller, 2019).
4. A deep neural network (DNN) based model (Avila et al., 2019).
5. pLSTM+att (Dong and Williamson, 2020).

Experimental Comparisons

Speech Data:

- All signals are resampled at 16 kHz, audios are truncated to 4s.
- Two datasets are split into training (80%), validation (10%) and testing (10%) subsets.

Network Setup:

- 20-sample kernel with 10-sample stride is used for 1D-convolution (257 output channels).
- 3X3 kernel used for 2D-convolution layers, with output channels set to 16, 32, 64 and 128.
- 3 blocks of pBi-LSTM (128, 64 and 32 units in each direction) are used.
- Trained 100 epochs using Adam optimizer.

Experimental Comparisons

Table 1. Average performance between comparison and our proposed models. '-' indicates that the model is not capable of estimating such scores. The best results are in **bold**.

	MOS			PESQ			eSTOI			SDR		
Systems	MAE	PCC	SRCC	MAE	PCC	SRCC	MAE	PCC	SRCC	MAE	PCC	SRCC
AMSA [23]	-	-	-	0.30	0.94	0.79	0.11	0.90	0.78	5.20	0.94	0.83
DNN [18]	0.49	0.94	0.88	0.19	0.96	0.83	0.05	0.96	0.86	3.50	0.98	0.88
Quality-Net [14]	0.48	0.93	0.87	0.15	0.97	0.81	0.06	0.95	0.80	2.72	0.96	0.88
NISQA [13]	0.50	**0.96**	**0.90**	0.18	**0.98**	0.88	0.06	0.96	**0.88**	2.20	0.98	**0.93**
pBi-LSTM+Att [22]	0.44	0.94	0.88	0.17	0.95	0.78	0.05	0.95	0.74	3.58	0.94	0.83
Prop. System (STFT)	0.42	0.95	0.88	0.17	0.95	0.80	**0.04**	0.94	0.85	2.69	0.97	0.89
Prop. System (1D-Conv)	**0.40**	**0.96**	**0.90**	**0.12**	**0.98**	**0.89**	**0.04**	**0.97**	**0.88**	**1.87**	**0.99**	**0.93**

1. The proposed system achieves slightly better correlation with human ratings on real-world datasets.
2. The proposed system also achieves slightly better performance in other targets.
3. A learnable 1D-convolution layer leads to improvements for our system.

Experimental Comparisons

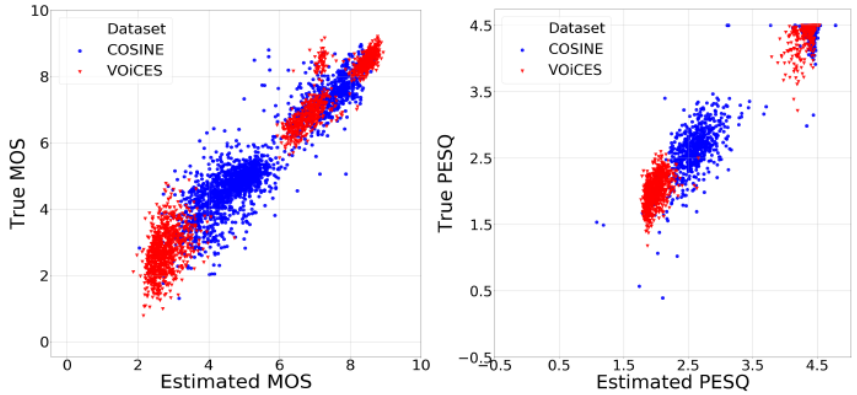


Fig. 4. Correlation between estimated scores and groundtruth on COSINE and VOiCES datasets (left: MOS, right: PESQ).

Our model has most of its estimations fall on the diagonals which reflects the high correlation Results.

Conclusions

- We proposed a novel multi-task data-driven non-intrusive speech assessment model that can analyze the speech quality from subjective and many objective perspectives.
- The proposed model achieves similar while slightly higher correlations, lower estimation errors when compared to the other speech assessment systems on the evaluated dataset.
- Better encoding capability with a 1D convolution layer than conventional STFT.
- Future work will evaluate the generalization ability and move on to subjective intelligibility estimation.

Thank you for your attention!

Audio, Speech and Information Retrieval (ASPIRE) Group
https://aspire.sice.indiana.edu

INDIANA UNIVERSITY
BLOOMINGTON