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

3995

Zhuohuang Zhang[1,2]; Piyush Vyas[1]; Xuan Dong[1]; Donald S. Williamson[1]
[1]Department of Computer Science, Indiana University; [2]Department of Speech, Language and Hearing Sciences, Indiana University
{zhuozhan, piyush, xuandong}@iu.edu; williads@indiana.edu

Introduction

Speech assessment is crucial for many applications. In this paper, we propose a novel multi-task non-intrusive approach that is capable of simultaneously estimating both **subjective** and **objective** scores of real-world speech.

The encoder first uses convolution and pyramid Bi-LSTM (pBi-LSTM) layers to extract features locally and temporally at different resolutions, directly from the time-domain signal. Then an attention mechanism is applied in a decoder for multi-task learning.

The proposed system assesses speech from multiple perspectives, including subjective and objective speech quality (i.e., human MOS and PESQ), objective intelligibility and signal distortions (i.e., eSTOI and SDR).

The **contributions** are:

1. The first non-intrusive speech assessment system that estimates both subjective and objective ratings for real-world recorded speech.
2. An end-to-end model that encodes the time-domain speech with a convolution layer rather than using the conventional short-time Fourier transform (STFT) which may not be optimal.
3. Although not shown here, this approach enables direct comparisons between real-world and laboratory experiments.

Real-world Speech Data

We use the **VOiCES** (Richey et al., 2018) and the **COSINE** (Stupakov et al., 2018) corpora as the source for the real-world speech materials.

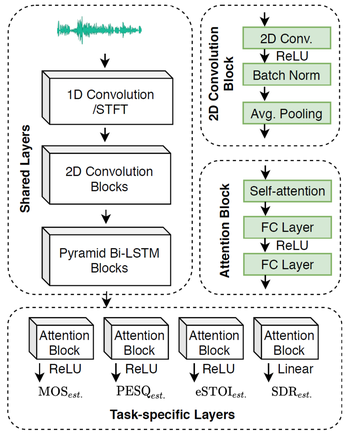
VOiCES:

- Recorded in acoustically challenging and reverberant environments using twelve mics strategically placed around different rooms.
- Recordings from two of the mics are used as reverberant stimuli and the foreground speech is used as the reference signal.
- The approximated speech-to-reverberation ratios (SRRs) of these signals range from -4.9 to 4.3 dB.

COSINE:

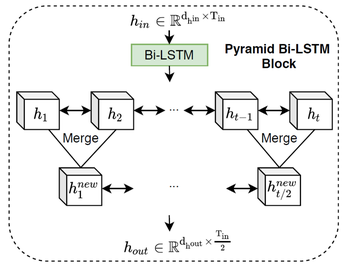
- Recorded in a multi-party conversational setting both indoor and outdoor to represent various background noises with seven mics.
- The close-talking mic captures high quality speech and is used as the clean reference. Recordings from the shoulder mic and the 4-mic array are used as noisy signals.
- The speech-to-noise ratios (SNRs) for COSINE are approximately between -10.1 to 11.4 dB.

Approach



Network Structure

$$h^{t}_{\text{pBi-LSTM}} = h^{2t-1}_{\text{Bi-LSTM}} + h^{2t}_{\text{Bi-LSTM}}$$



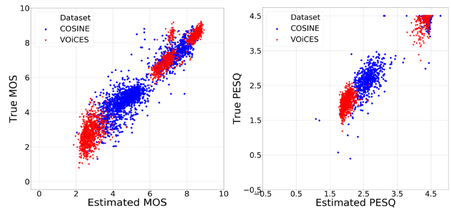
- Audios are truncated to 4 s with 16 kHz sampling rate.
- 20-sample kernel and 10-sample stride are used for 1D convolution with 257 output channels.
- Four 2D convolution blocks are used with 3X3 kernel. The output channels are set to 16, 32, 64 and 128. Three blocks of the pBi-LSTM (with 128, 64 and 32 units in each direction) are adopted.
- The training targets are human MOS (0-10), PESQ (-0.5 to 4.5), eSTOI (0 to 1) and SDR (-25 to 36 dB).
- Weights for each loss term are determined empirically, where $\alpha_1 = 10$, $\alpha_2 = 1$, $\alpha_3 = 12$, $\alpha_4 = 0.1$ for MOS, PESQ, eSTOI and SDR, respectively.

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

Experimental Results

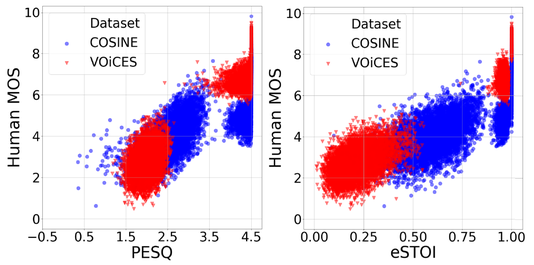
Systems	MOS MAE	MOS PCC	MOS SRCC	PESQ MAE	PESQ PCC	PESQ SRCC	eSTOI MAE	eSTOI PCC	eSTOI SRCC	SDR MAE	SDR PCC	SDR SRCC
AMSA [23]	-	-	-	0.30	0.94	0.79	0.11	0.90	0.78	5.20	0.94	0.83
DNN [18]	0.49	0.94	0.88	0.19	0.96	0.83	0.05	0.96	0.86	3.50	0.98	0.88
Quality-Net [14]	0.48	0.93	0.87	0.15	0.97	0.81	0.06	0.95	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**

- The proposed system (with 1D-Conv) achieves better correlation with human ratings on real-world datasets.
- The proposed system achieves better performance in almost all targets than other methods.
- A learnable 1D-convolution layer leads to slight improvements for all targets.



Online Data Collection

- Online listening tests (Dong et al., 2020) on Amazon Mechanical Turk (700 human intelligence tasks).
- In total 3,500 workers participated (1,455 females and 2,045 males), aged from 18 to 65 years old.
- All participants are native English speakers and self-reported to have normal hearing.
- Each task contains 15 trials of evaluations that follow ITU-R BS.1534. Users provide quality ratings (between 0 to 100) for all stimuli.
- A total of 180K responses are collected for 36K signals (18K signals per dataset) with a total duration of approximately 45 hours.



- There are some correlations between the subjective and objective scores, jointly learning may be beneficial.

Conclusions

- Our proposed approach (i.e. with 1D-Conv layer) achieves the best performance according to all evaluation metrics.
- The learnable 1D convolution layer leads to slight improvements for all targets in nearly all criteria.
- A novel multi-task data-driven non-intrusive speech assessment model that is capable of analyzing the speech quality from subjective and different objective perspectives.

Acknowledgments

This work is supported by an NSF grant (IIS-1755844) and is also supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

INDIANA UNIVERSITY