ICASSP 2020

An Attention Enhanced Multi-task Model for Objective Speech Assessment in Real-world Environments

Xuan Dong and Donald S. Williamson
Department of Computer Science, Indiana University, USA

May 2020

Speech Quality and Intelligibility

- Two important attributes of speech
- Important to many applications and products
- Subjective listening studies
    - the most accurate way
    - expensive and time consuming

Computational Measures

- Intrusive metrics
  - require access to the original speech (a limitation)
  - e.g., source-to-distortion ratio (SDR) [1], hearing-aid speech quality index (HASQI) [2], perceptual evaluation of speech quality (PESQ) [3], extended short-time objective intelligibility (ESTOI) [4]
- Non-intrusive measures
  - rely on signal properties and assumptions
  - e.g., IP.563 [5], ANIQUE [6], speech to reverberation modulation energy ratio (SRMR) [7]

Motivation

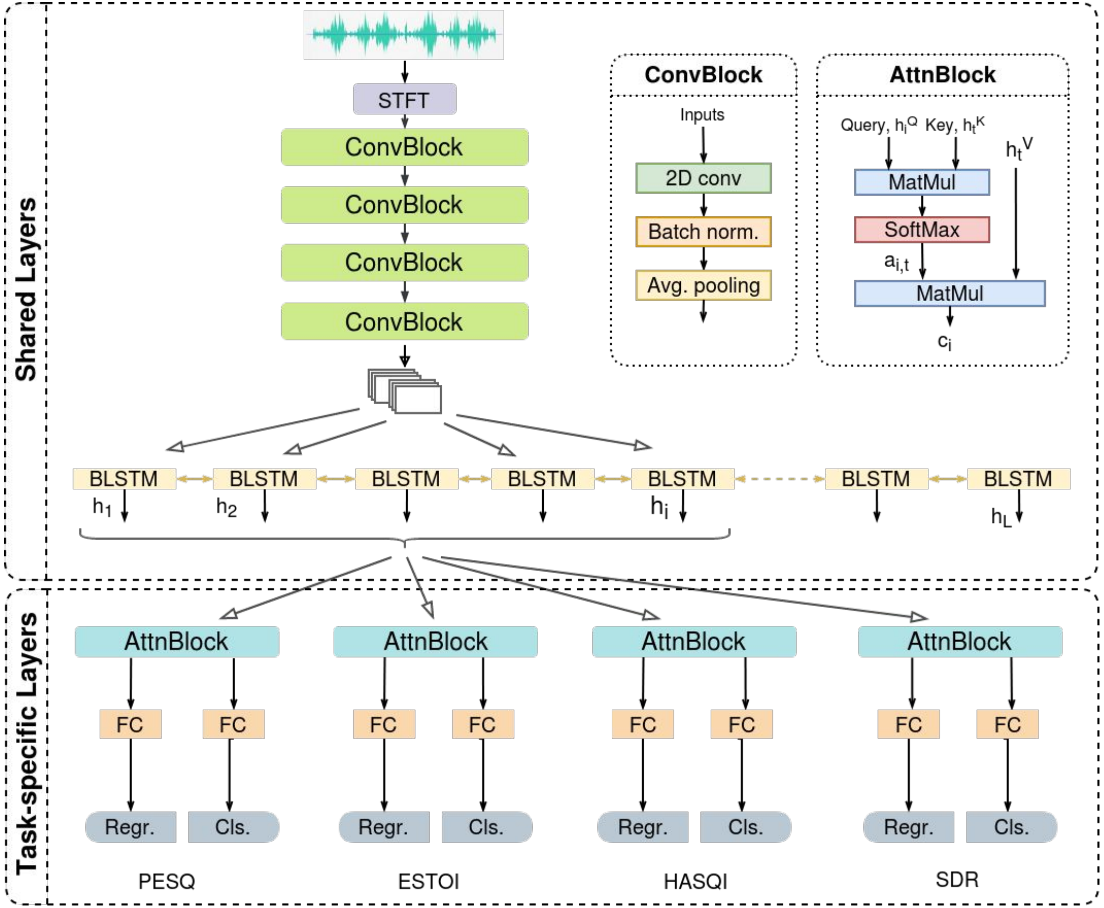
- Data-driven approaches
  - AutoMOS [8], CNN-based [9], DNN-based [10], Quality-Net [11], NISQA [12]
- Limitations
  - not correlated well with human evaluation
  - not reliable in extreme test conditions
  - not generalize well in unseen environment
  - singular quality or intelligibility assessment

Proposed Approach

- The attention enhanced multi-task speech assessment (AMSA) model
  - input: a clip of speech
  - output: estimates of PESQ, ESTOI, HASQI, and SDR metrics
  - multi-task learning: leverages different aspects of speech assessment

AMSA Model

- Shared layers
    - 4 convolutional layers
    - 1 bidirectional LSTM (BLSTM) layer
- Task-specific layers
    - 1 attention layer
    - classification-aided module [13]



Classification-aided Module

- Motivation: reduce estimation outliers
- Raw objective score: $\text{score}_{k,s}$
- Categorical label is calculated as

$$\text{class}_{k,s} = \min(\max\left(1, \text{ceil}\left(\frac{score_{k,s} - L_{k,thres}}{(H_{k,thres} - L_{k,thres})/N_k}\right)\right), N_k).$$

- Objective function

$$\mathcal{L}_{total} = \sum_{k=1}^{K} \beta_k(\mathcal{L}_{k,regr} + \lambda_k * \mathcal{L}_{k,cls})$$

Experiment Setup

- Speech materials: TIMIT speech corpus
- Test conditions: simulated noisy, reverberant, and noisy-reverberant environments
- Performance is measured with root mean square error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (PCC)

Experimental Results I

	PESQ			ESTOI			HASQI			SDR		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
AutoMOS [8]	0.35	0.30	0.84	0.14	0.10	0.83	0.12	0.12	0.83	2.71	2.56	0.87
CNN [9]	0.29	0.27	0.86	0.07	0.06	0.93	0.08	0.06	0.90	2.13	1.97	0.91
DNN [10]	0.19	0.18	0.90	0.11	0.08	0.86	0.06	0.07	0.88	1.90	1.84	0.91
Quality-Net [11]	0.16	0.17	0.91	0.05	0.04	0.96	0.04	0.04	**0.91**	1.52	1.48	0.92
NISQA [12]	0.19	0.17	0.90	0.06	0.06	0.94	0.05	0.04	**0.91**	1.24	1.27	0.92
AMSA	**0.11**	**0.10**	**0.94**	**0.02**	**0.03**	**0.97**	**0.02**	**0.02**	**0.91**	**0.62**	**0.65**	**0.95**

Test on Real-world Corpora

- COnversational Speech In Noisy Environments (COSINE) corpus [14]
  - multi-party conversations with background noise and interfering speakers
- Voices Obscured in Complex Environmental Settings (VOiCES) corpus [15]
  - background noise played in conjunction with foreground speech in two furnished rooms

Experimental Results II

	PESQ			ESTOI			HASQI			SDR		
	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC
Quality-Net [11]	0.56	0.63	0.69	0.17	0.19	0.56	0.10	0.12	**0.71**	4.37	5.69	0.67
NISQA [12]	0.34	0.38	0.77	0.14	0.18	0.63	0.06	0.08	**0.75**	4.13	4.55	0.71
AMSA	**0.25**	**0.29**	**0.84**	**0.06**	**0.05**	**0.81**	**0.05**	**0.05**	**0.79**	**2.63**	**2.30**	**0.81**

Summary

- Propose an attention enhanced multi-task model for speech assessment
- Apply a single model to predict a number of objective speech quality and intelligibility metrics simultaneously
- Significantly reduce the estimation error and improves the generalization ability in real-world acoustic environments

References

[1] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” IEEE TASLP, vol. 14, 2006

[2] J. Kates and K. Arehart, “The hearing-aid speech quality index (HASQI) version 2,” J. Audio Eng. Soc., vol. 62, 2014.

[3] ITU-T P.862, “Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs,” 2001.

[4] J. Jensen and C. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” IEEE TASLP, vol. 24, no. 11, pp. 2009–2022, 2016.

[5] L. Malfait, J. Berger, and M. Kastner, “P.563-The ITU-T standard for single-ended speech quality assessment,” IEEE TASLP, vol. 14, pp. 1924–1934, 2006.

[6] D. Kim, “ANIQUE: An auditory model for single-ended speech quality estimation,” IEEE TSAP, vol. 13, no. 5, 2005.

[7]T. Falk, C. Zheng, and W. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” IEEE TASLP, vol. 18, no. 7, pp. 1766–1774, 2010.

[8] B. Patton, Y. Agiomyrgiannakis, M. Terry, et al., “AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech,” Workshop NIPS, 2016.

[9] A. Andersen, J. Haan, Z. Tan, and J. Jensen, “Nonintrusive speech intelligibility prediction using convolutional neural net- works,” IEEE TASLP, vol. 26, pp. 1925–1939, 2018.

References

[10] A. Avila, H. Gamper, C. Reddy, R. Cutler, et al., “Non- intrusive speech quality assessment using neural networks,” in Proc. ICASSP. IEEE, 2019, pp. 631–635.

[11] S. Fu, Y. Tsao, H. Hwang, et al., “Quality-Net: An end-to-end non-intrusive speech quality assessment model based on blstm,” Interspeech, 2018.

[12] G. Mittag and S. Möller, “Non-intrusive speech quality assessment for super-wideband speech communication networks,” in Proc. ICASSP. IEEE, 2019, pp. 7125–7129.

[13] X. Dong and D. S Williamson, “A classification-aided framework for non-intrusive speech quality assessment,” in Proc. WASPAA, 2019.

[14] A. Stupakov, E. Hanusa, J. Bilmes, et al., “COSINE-a corpus of multi-party conversational speech in noisy environments,” in Proc. ICASSP. IEEE, 2009, pp. 4153–4156.

[15] C. Richey, M. Barrios, Z. Armstrong, et al., “Voices obscured in complex environmental settings (VOICES) corpus,” arXiv preprint arXiv:1804.05053, 2018.

Thank You

For more information please refer to our paper.

Welcome to our ASPIRE Research Group!

https://aspire.sice.indiana.edu/index.html