A Non-Intrusive Speech Intelligibility Estimation Method Based on Deep Learning Using Autoencoder Features

Yoonhee KIM†, Deokgyu YUN†, Hannah LEE†, Nonmembers, and Seung Ho CHOI†a, Member

SUMMARY This paper presents a deep learning-based non-intrusive speech intelligibility estimation method using bottleneck features of autoencoder. The conventional standard non-intrusive speech intelligibility estimation method, P.563, lacks intelligibility estimation performance in various noise environments. We propose a more accurate speech intelligibility estimation method based on long-short term memory (LSTM) neural network whose input and output are an autoencoder bottleneck features and a short-time objective intelligibility (STOI) score, respectively, where STOI is a standard tool for measuring intrusive speech intelligibility with reference speech signals. We showed that the proposed method has a superior performance by comparing with the conventional standard P.563 and mel-frequency cepstral coefficient (MFCC) feature-based intelligibility estimation methods for speech signals in various noise environments.

key words: autoencoder, bottleneck feature, STOI, deep learning, long short-term memory (LSTM)

1. Introduction

The methods for estimating objective speech intelligibility are divided into intrusive or non-intrusive methods according to the presence or absence of a reference speech signal. The P.563[1] is a representative non-intrusive method, and STOI (Short-Time Objective Intelligibility measure) [2] is an intrusive method for estimating speech intelligibility. The non-intrusive method is less accurate than the intrusive method because there is no reference signal.

However, since there is no reference signal for the receiver such as communication, a non-intrusive intelligibility estimation method is required. In [3], the speech intelligibility is predicted using convolutional neural network which is trained with measured intelligibility scores that humans listen and evaluate. The work in [4] presented the method of speech intelligibility prediction by using automatic speech recognition (ASR) system based deep neural networks. Recently, the non-intrusive speech intelligibility estimation method based on a recurrent neural network (RNN) with a mel-frequency cepstrum coefficient (MFCC) vector was proposed [5]. In this paper, we propose a novel method of estimating intelligibility score with higher performance than conventional methods in noise environments based on deep learning using autoencoder bottleneck feature and STOI values. We confirmed the superiority of the performance of the proposed method using normalized correlation coefficients (NCC) and the root mean square error (RMSE) with respect to STOI.

2. LSTM-Based Intelligibility Estimation Using Autoencoder Bottleneck Features

In this research work, the autoencoder is trained with the short-term spectral magnitudes as input and output in order to extract bottleneck features [6], [7]. Long-short term memory (LSTM) [8], [9] neural network is used for intelligibility estimation, which was designed to model temporal sequences and have been successfully applied for the task of speech and acoustic modeling. The LSTM neural network is trained with the bottleneck feature vectors and STOI scores as input and output, respectively, in various noisy environments as shown in Fig. 1.

In the test, the intelligibility score per frame is estimated based on LSTM speech intelligibility estimator by using the bottleneck features of the noisy input speech. Then, an intelligibility score of the test utterance is obtained by averaging the frame-wise intelligibility scores.

In this paper, ReLU (Rectified Linear Units) [10] and ADAM (ADAmptive Moment Estimation) [11] are used as the activation function and the statistical optimization algorithm for learning the neural network parameters.

3. Experiments and Results

In order to evaluate the performance of the proposed
method, we used the TIMIT speech database [12], where each sentence was about 4 secs long and was resampled at a rate of 8 kHz. The input and output data for the autoencoder is 129-dimensional short-time spectrum magnitude vector in various noisy environments. We extracted the 39-dimensional bottleneck features from the trained autoencoder, which were used as input data of LSTM neural network. The LSTM neural network was trained under seven environments that consist of clean and noisy conditions. The tests were conducted in 21 environments, including seven training environments and 14 noisy environments that were not used in the training process. The noise data were seven types: Cell phone, Ring tone, Step noise, Dog, TV, Car horn, and Siren.

Table 1 shows normalized correlation coefficients (NCC) and root mean square error (RMSE) between the intelligibility scores obtained by each method and STOI values. This table is divided into trained noise environments (known environments) and test environments (unknown environments). Experimental results show that the proposed non-intrusive speech intelligibility estimation method based on LSTM using autoencoder feature is superior to the conventional methods, which are P.563 and non-intrusive speech intelligibility estimation using MFCC.

### Acknowledgments

This work was supported by the research fund of Signal Intelligence Research Center supervised by Defense Acquisition Program Administration and Agency for Defense Development of Korea.

### References

1. L. Malfait, J. Berger, and M. Kastner, “P.563—The ITU-T standard for single-ended speech quality assessment,” IEEE Trans. Audio, Speech, Language Process., vol.14, no.6, pp.1924–1934, 2006.
2. C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Trans. Audio, Speech, Language Process., vol.19, no.7, pp.2125–2136, 2011.
3. A.H. Andersen, J.M. de Haan, Z. Tan and J. Jensen, “Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks,” IEEE/ACM Trans. Audio, Speech, Language Process., vol.26, no.10, pp.1925–1939, Oct. 2018.
4. C. Spille, S.D. Ewert, B. Kollmeier, and B.T. Meyer, “Predicting speech intelligibility with deep neural networks,” Computer Speech & Language, vol.48, pp.51–66, 2018.
5. D. Yun, H. Lee, and S.H. Choi, “A deep learning-based approach to non-intrusive objective speech intelligibility estimation,” IEICE Trans. Inf. & Syst., vol.E101-D, no.4, pp.1207–1208, 2018.
6. T.N. Sainath, B. Kingsbury, and B. Ramab, “Auto-encoder bottleneck features using deep belief networks,” Proc. ICASSP, pp.4153–4156, 2012.
7. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech Enhancement Based on Deep Denoising Autoencoder,” Proc. INTERSPEECH, pp.436–440, 2013.
8. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol.9, no.8, pp.1735–1780, Nov. 1997.
9. H. Sak, A.W. Senior, and F. Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling models,” Proc. INTERSPEECH, pp.338–342, 2014.
10. V. Nair and G.E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” Proc. 27th international conference on machine learning (ICML-10), pp.807–814, 2010.
11. D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
12. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S. Pallett, “DARPA TIMT acoustic phonetic continuous speech corpus CDROM,” NIST, 1993.