International Journal of Internet, Broadcasting and Communication Vol.12 No.3 220-225 (2020)
http://dx.doi.org/10.7236/IJIBC.2020.12.3.220

IJIBC 20-3-28

Non-Intrusive Speech Intelligibility Estimation Using Autoencoder Features with Background Noise Information

Yue Ri Jeong*, Seung Ho Choi**

*Undergraduate Student, Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Seoul, Korea
**Professor, Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Seoul, Korea
*1004yueri@naver.com, **shchoi@snut.ac.kr

Abstract

This paper investigates the non-intrusive speech intelligibility estimation method in noise environments when the bottleneck feature of autoencoder is used as an input to a neural network. The bottleneck feature-based method has the problem of severe performance degradation when the noise environment is changed. In order to overcome this problem, we propose a novel non-intrusive speech intelligibility estimation method that adds the noise environment information along with bottleneck feature to the input of long short-term memory (LSTM) neural network whose output is a short-time objective intelligence (STOI) score that is a standard tool for measuring intrusive speech intelligibility with reference speech signals. From the experiments in various noise environments, the proposed method showed improved performance when the noise environment is same. In particular, the performance was significant improved compared to that of the conventional methods in different environments. Therefore, we can conclude that the method proposed in this paper can be successfully used for estimating non-intrusive speech intelligibility in various noise environments.

Keywords: *Non-intrusive, Speech intelligibility estimation, noise environment, Autoencoder, Bottleneck feature, Long short-term memory (LSTM), STOI*

1. INTRODUCTION

The speech intelligibility estimation methods 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, which calculates the correlation between the reference signal and distorted signal in the frequency domain. The non-intrusive method is necessary in speech communication when the reference signal cannot be obtained. For the non-intrusive measure of speech quality and intelligibility, the data-driven method was developed [3]. Recently, deep learning-based non-intrusive speech intelligibility estimation methods were developed [4, 5]. The non-intrusive estimation method based on deep neural network, which incorporating

Manuscript Received: July. 21, 2020 / Revised: July. 27, 2020 / Accepted: August. 1, 2020
Corresponding Author: shchoi@snut.ac.kr
Tel: +82-2-970-6461, Fax: +82-2-979-7903
Professor, Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Seoul, Korea

STOI values, was studied [6, 7]. In [6] and [7], the input of long short-term memory (LSTM) [8, 9] are the MFCC vector and the bottleneck feature vector of autoencoder [10], respectively, and the output is the frame-wise STOI value for the training of the LSTM, in various noise environments. However, the performance of the non-intrusive speech intelligibility estimation using the bottleneck feature may be significantly deteriorated according to changes in the environment such as background noise. The estimation method in [7] did not use background noise information explicitly for the training and test of LSTM model. In this paper, we propose a novel method that uses background noise information as an input to LSTM along with bottleneck feature to overcome the disadvantages of the bottleneck feature.

2. NON-INTRUSIVE SPEECH INTELLIGIBILITY ESTIMATION BASED ON BOTTLENECK FEATURE WITH BACKGROUND NOISE INFORMATION

The feature used for intelligibility estimation is a frequency domain feature such as MFCC or a deep learning feature such as a bottleneck feature [10] of an autoencoder. In this research work, as shown in Figure 1, the autoencoder is trained with the short-term spectral magnitudes of clean speech frame and noisy frame as input and output in order to extract bottleneck features. The procedure for training the speech intelligibility estimation model in this study is shown in Figure 2. We used LSTM [8, 9] neural network for the intelligibility estimation, which has been successfully applied for the task of speech and acoustic modeling. The LSTM model was trained using feature vector and STOI [2] value as input and output, respectively. The number of input nodes of the LSTM is equal to the order of the feature vectors, and the number of output nodes is one. 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.

The intelligibility estimation performance using bottleneck feature may be significantly deteriorated when the noise environment is changed. Therefore, we devised a new method to use the background noise information together with the bottleneck feature as the LSTM input as shown in Figure 3. For the background noise information, as shown in the following equation (1), we used the average of the spectral magnitude for each frequency bin using the N frames of noise interval.

$$|\bar{X}(k)| = \frac{1}{N}\sum_{i=0}^{N-1}|X(i,k)|, \text{ k} = 0, \cdots, \text{M-1} \tag{1}$$

where i and k are frame index and frequency bin index, respectively.

In this research work, ReLU (Rectified Linear Units) [11] and ADAM (ADAmptive Moment Estimation) [12] are used as the activation function and the statistical optimization algorithm for learning the neural network parameters.

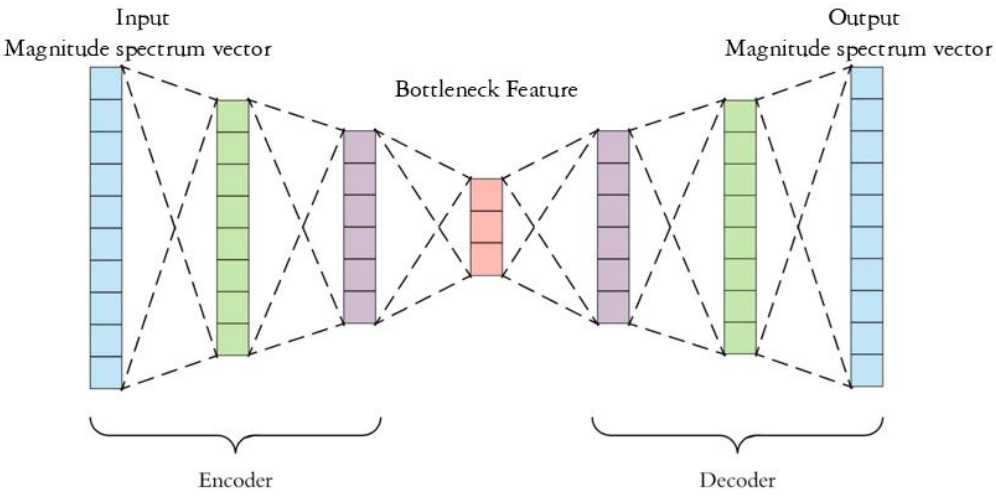

Figure 1. Autoencoder structure and bottleneck feature

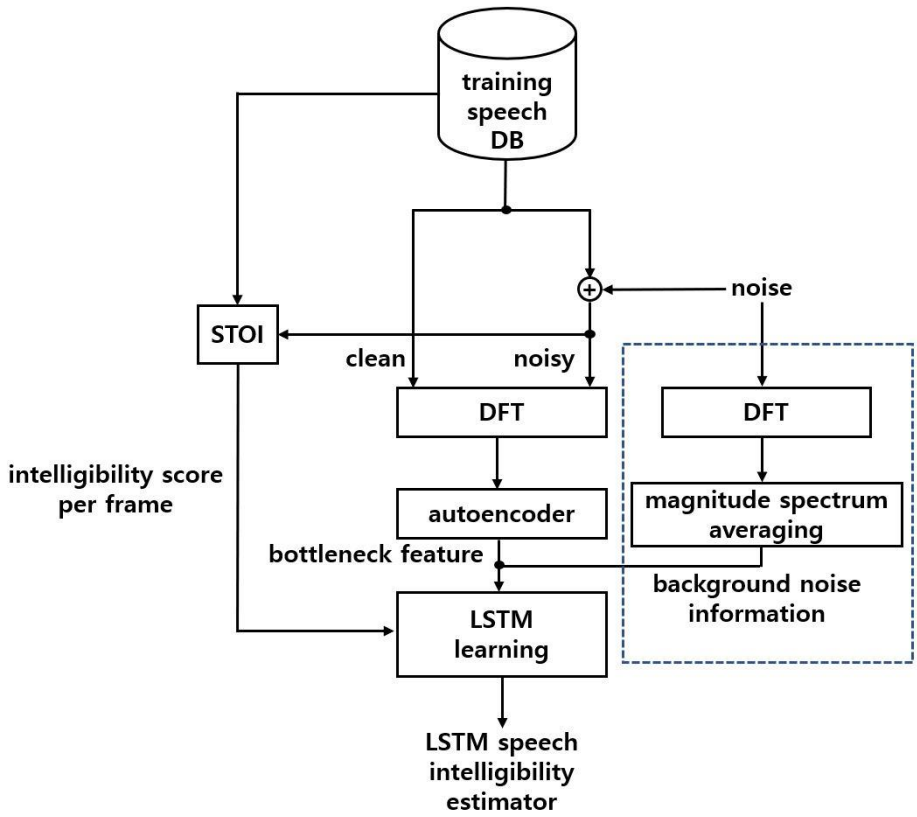

Figure 2. Training procedure for speech intelligibility estimator

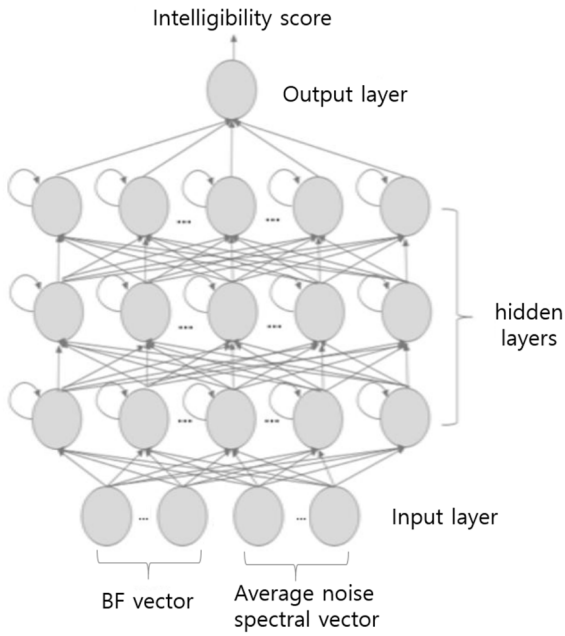

Figure 3. LSTM model structure of the proposed method

3. EXPERIMENTS AND RESULTS

In order to evaluate the performance of the conventional and proposed methods, we used the TIMIT speech database [13], where each sentence was about 4 secs long and was resampled at a rate of 8 kHz. We used 700 and 500 sentences for the training and test phases. The input and output for the autoencoder are 129-dimensional short-time spectrum magnitude vectors extracted from 20 msec frames with the frame shift of 10 msec in clean and various noisy environments. We extracted the 40-demensional bottleneck features from the trained autoencoder. The noise data were seven types: Cell phone, Ring tone, Step noise, Dog, TV, Car horn, and Siren. The number of hidden layers and hidden units for each hidden layer of the autoencoder network was 5 and 128, respectively. For the training of the autoencoder, the number of epochs was 50 and the learning was 0.001. The number of hidden layers and hidden units of LSTM network was 3 and 128, respectively, and the number of epochs was 10 and the learning was 0.001 for the training.

First, we examined the performance degradation of the conventional LSTM-based method [6] using bottleneck features without using the background noise information, according to the changes in the environment such as background noise. Table 1 shows the normalized correlation coefficient (NCC) and the root-mean-square error (RMSE) between the estimated intelligibility scores and STOI values. In the table, 'Matched condition' or 'Mismatched condition' means that the noise environment in training and test is same or different, respectively. As a result of the experiment, in 'Matched condition', bottleneck feature showed superior performance compared to MFCC. However, MFCC is superior in 'Mismatched condition'. That is, the performance deterioration is severe because the bottleneck feature is sensitive to environmental changes.

Table 1. Estimation performance for matched and mismatched conditions

	Matched condition		Mismatched condition	
	MFCC	Bottleneck feature	MFCC	Bottleneck feature
RMSE	0.056	0.028	0.094	0.145
NCC	0.919	0.979	0.832	0.619

We conducted the experiments in order to evaluate the performance of the proposed method that tries to overcome the disadvantages of bottleneck feature. In the test, the average noise spectral vector is the spectral average values of N frames in which noise only exists before the start of the speech interval, and N is set to 5. As can be seen in Table 2, the proposed method has improved performance compared to the previous results in 'Matched condition' and 'Mismatched condition', especially when compared to the result in 'Mismatched condition'. This means that the bottleneck feature can be successfully used under various noise environments by adding noise information to the input of LSTM.

Table 2. Estimation performance of the proposed method

	Bottleneck feature with noise information
RMSE	0.012
NCC	0.996

4. CONCLUSION

This paper concerned the non-intrusive speech intelligibility estimation method using a bottleneck feature as an input to a neural network. First, we showed that the bottleneck feature-based method has the problem of severe performance degradation when the noisy environment is changed. In order to overcome this, a new non-intrusive speech intelligibility estimation method was proposed, which adds the noise environment information to the input of LSTM. As a result of the experiment, the proposed method showed improved performance in both 'Matched condition' and 'Mismatched condition'. In particular, the improvement was significant compared to the 'Mismatched condition'. Therefore, we can conclude that the method proposed in this paper can be successfully used for estimating non-intrusive speech intelligibility.

ACKNOWLEDGEMENT

This study was supported by the Research Program funded by the SeoulTech (Seoul National University of Science and Technology).

REFERENCES

[1] Ludovic Malfait, Jens Berger, and Martin Kastner, “P.563 —The ITU-T standard for single-ended speech quality assessment,” *IEEE Transactions on Audio, Speech, and Language Processing* 14.6, pp.1924-1934, 2006.

DOI**:** 10.1109/TASL.2006.883177

[2] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125-2136, 2011.
DOI: https://www.doi.org/10.1109/TASL.2011.2114881

[3] Dushyant Sharma, Yu Wang, Patrick A. Naylor, Mike Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Communication*, vol. 80, June 2016, pp. 84-94, June 2016.
DOI: https://doi.org/10.1016/j.specom.2016.03.005

[4] A. H. Andersen, J. M. de Haan, Z. tan and J. Jensen, "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1925-1939, Oct. 2018.
DOI: 10.1109/TASLP.2018.2847459

[5] Anderson R. Avila, Hannes Gamper, Chandan Reddy, Ross Cutler, Ivan Tashev, and Johannes Gehrke, "Non-intrusive Speech Quality Assessment Using Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 18777982, May 2019.
DOI: 10.1109/ICASSP.2019.8683175

[6] D. K. Yun, H. N. Lee, and S. H. Choi, "A Deep Learning-Based Approach to Non-Intrusive Speech Intelligibility Estimation," *IEICE Trans. Information and Systems*, pp. 1207-1208, Apr. 2018.
DOI: 10.1587/transinf.2017EDL8225

[7] Y. H. Kim, D. K. Yun, H. N. Lee, and S. H. Choi, "A Non-Intrusive Speech Intelligibility Estimation Method Based on Deep Learning Using Autoencoder Features" *IEICE Trans. Information and Systems,* Vol.E103-D No.3, March. 2020.
DOI: 10.1587/transinf.2019EDL8150

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
DOI: 10.1162/neco.1997.9.8.1735

[9] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling models," *Proc. INTERSPEECH*, pp. 338-342, 2014.

[10] Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramab, "Auto-encoder bottleneck features using deep belief networks," *Proc. ICASSP*, pp. 4153-4156, 2012.
DOI: 10.1109/ICASSP.2012.6288833

[11] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," *Proc. of the 27th international conference on machine learning (ICML-10)*, pp. 807-814. 2010.
DOI: https://dl.acm.org/citation.cfm?id=3104425

[12] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
DOI: https://arxiv.org/abs/1412.6980

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," NIST, 1993.