PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

On non-reference speech intelligibility estimation using DNN noise reduction

Hiroto Takahashi[1]; Kazuhiro Kondo[2]

[1] Graduate School of Science and Engineering, Yamagata University, Japan

[2] Graduate School of Science and Engineering, Yamagata University, Japan

ABSTRACT

Methods for estimating speech intelligibility are classified into two types. One is the full reference method that estimates the subjective evaluation values using both degraded speech that passed through the evaluation system and the original speech before degradation. The other is the non-reference method that estimates intelligibility from only degraded speech. In speech intelligibility estimation, it is assumed that the original speech cannot be obtained. Therefore, from the viewpoint of practicality, a non-reference method is required. In this research, we consider a method to apply the full reference method on degraded speech and original speech estimated from degraded speech. The model used here is an intelligibility estimation model using Deep Neural Network (DNN) which shows higher estimation accuracy than other methods such as logistic regression and random forests. In this paper, we compared the estimation accuracy between estimation using original speech and estimation using estimated speech. In the closed test, the correlation coefficient was 0.9721 and the RMSE was 0.0743, showing the same degree of accuracy as when the original speech was used. However, in the open test, the correlation coefficient and RMSE was 0.8307, 0.1839, respectively, indicating room for improvement.

Keywords: Speech intelligibility, Noise reduction, Deep neural network

1. INTRODUCTION

In speech communication using mobile phones and the Internet, speech quality deteriorates due to various factors. For example, mixing of noise from the surroundings, distortion due to data compression, packet loss generated at the time of data transmission, and the like can be listed. In order to design and manage speech communication systems, it is necessary to evaluate and guarantee the quality of the speech that includes such deteriorations.

Speech intelligibility is one of the measures of speech quality. It is a measure of how accurately words and sentences of speech are transmitted to the other party. Currently, the most reliable speech intelligibility evaluation method is the subjective evaluation method in which a person actually using the system listens to speech and evaluates the speech quality. However, because of its nature, the subjective evaluation method requires an excessive amount of time, labor and cost. Therefore, research on an objective evaluation method is being conducted, which measures physical quantities representing the degree of quality deterioration from the observed signals and estimates the subjective speech intelligibility from these quantities.

The objective evaluation method is classified into two broad types. One is the full reference method that estimates the subjective evaluation values using both degraded speech that passed through the evaluation system and the original speech before degradation. The other is the non-reference method that estimates speech intelligibility from only degraded speech. Generally, the full reference method using the original speech as the reference signal boasts a higher estimation accuracy than the non-reference method.

If the reference signal is not available, the non-reference method, which does not require a reference signal, is used, but the estimation accuracy is generally lower than the full reference method. Therefore, we propose a method to estimate the original speech before deterioration by removing noise,

[1] txt21892@st.yamagata-u.ac.jp

[2] kkondo@yz.yamagata-u.ac.jp

and use it as a reference signal, as a method to obtain the same estimation accuracy as the full reference method even when there is no reference signal.

In this paper, we performed noise reduction and speech intelligibility estimation using Deep Neural Network (DNN) and evaluated its estimation accuracy. We also focus on degradation due to additive noise.

2. SPEECH INTELLIGIBILITY ESTIMATION METHOD

2.1 Intelligibility Estimation Flow

The flow of speech intelligibility estimation is shown in Figure 1. First, in order to train a DNN model for clean speech estimation, noise is superimposed on clean speech, and noisy speech is created. The first DNN (DNN1), which will remove the noise, is constructed by training with the noisy speech as input data, and the clean speech as supervisory data. Next, feature values are calculated from the noisy speech and the estimated clean speech obtained using DNN 1. Then, the second DNN (DNN2), which is trained with the calculated feature values as input data and the subjective speech intelligibility evaluation results as supervisory data, estimates the speech intelligibility.

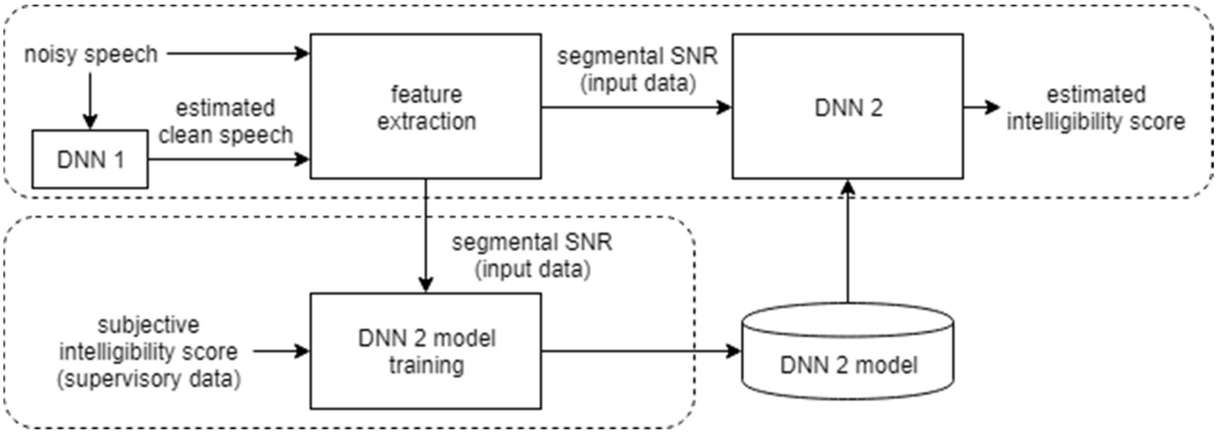

noisy speech
estimated clean speech
DNN 1
feature extraction
segmental SNR (input data)
DNN 2
estimated intelligibility score
segmental SNR (input data)
subjective intelligibility score (supervisory data)
DNN 2 model training
DNN 2 model

Figure 1 – Intelligibility estimation flow

2.2 Feature Value

The feature value used is the frequency weighted segmental SNR (fwSNRseg) [2]. The calculation is performed using the following formula,

$$\mathrm{fwSNRseg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)\log_{10}\frac{X^2(j,m)}{\{X(j,m)-\hat{X}(j,m)\}^2}}{\sum_{j=1}^{K} W(j,m)} \tag{1}$$

Here, j is the frequency band number, K is the number of frequency bands, and $W(j,m)$ is the weight of the j-th band. M is the total number of frames, $X(j,m)$ and $\hat{X}(j,m)$ are the amplitude spectrum of the clean speech and noisy speech in the j-th band in the m-th frame. Instead of using fwSNRseg as is, average processing is applied only in the time frame direction, and the weighted SNR by the band is taken as the input of DNN 2. In this case, mel scaling is used as the weight, and the number of frequency band divisions is set to 16.

2.3 Japanese Diagnostic Rhyme Test (JDRT)

The speech intelligibility used in this paper is calculated using the JDRT [3]. DRT is an intelligibility evaluation method performed by listening to a word-pair that differs only in the first phoneme [4]. The subject listens to only one word in the word-pair and chooses which one was heard from a choice of two words. Word speech used in the JDRT is classified into six types according to the phonemic feature of the first phoneme in the word. In this paper, we use the result obtained for the word-pairs in the feature "sustention" which is considered to be most suitable for training data because the effect on the subjective intelligibility shown by additive noise is the largest with this feature. In order to exclude the chance level in the number of correct answers in the two-to-one response, the correct answer rate is calculated using the following formula,

$$S = (R\text{-}W) / T \qquad (2)$$

Here, S is the correct answer rate (i.e., the intelligibility), R is the number of correct answers, W is the number of incorrect answers, and T is the total number of trials. The correct answer rate calculated using equation (2) may show negative values. However, in this paper, negative values are forced to 0 so that the correct answer rate never falls below 0.

3. EXPERIMENTS

3.1 Sound Source

The sound source used for training is the read-word speech in the JDRT word list. This list is a list composed of word-pairs differing only in its initial phoneme and is divided into six attributes according to phonemic features. Speech is read words of 120 words in total, 60 word-pairs by 1 female speaker. All 120 words are used for the training of DNN 1. The word-pairs used for the training of DNN2 are the pairs of words classified into the "sustention" feature. The subjective intelligibility for the words in this feature has been evaluated by past intelligibility tests. The noise added to these words is selected from the JEIDA-NOISE noise database0 [5]. Ten noise types selected from JEIDA - NOISE were used for the training of DNN 2, and three were used for testing. The breakdown of the noise types used is summarized in Table 1. The sampling frequency for all samples is 16 kHz, the number of quantization bits is 16 bits, and the number of channels is monaural.

Table 1 – Noise used for training and testing

Noise	Training DNN 1	Training DNN 2	Testing DNN 2
1. Exhibition (booth)			✓
2. Exhibition (aisle)	✓	✓	
3. Public telephone booth		✓	
4. Factory	✓	✓	
5. Sorting facility		✓	
6. Heavy traffic road	✓	✓	
7. Crowd		✓	
8. Train (bullet express)		✓	
9. Train (local line)	✓		✓
10. Computer room		✓	
11. Air conditioner	✓		✓
12. Ducts	✓	✓	
13. Elevator halls	✓	✓	

3.2 Training Conditions

In the training of DNN 1, logarithmic power spectrogram of noisy speech was used as input data. A frame for which noise is desired to be removed, and five frames before and after this frame are taken as one extended input frame. One frame of the logarithmic power spectrogram of clean speech corresponding to the frame for which noise is desired to be removed was used as the supervisory data.

In the training of DNN 2, fwSNRseg calculated using logarithmic power spectrogram of noisy speech and clean speech was used as input data. The spectrogram output from DNN 1 is used to calculate fwSNRseg without using the clean speech, i.e., no reference signal is used. The phase signal is not required in this calculation. Therefore, the influence of the noisy phase can be neglected here. For supervisory data, we used speech intelligibility calculated from the results of the JDRT. The intelligibility is calculated using equation (2) using the subject's response to the 20-word speech with additive noise.

The short-time Fourier transform for obtaining the spectrogram was performed with a frame length of 320 samples, a Hann window for the window function, with 50% overlap. The training was carried out under the conditions shown in Table 2 using this model.

Table 2 – Training conditions

Item	DNN 1	DNN 2
Preprocessing of training data	Standardization	Standardization
Hidden layer activation function	ReLU	ReLU
Output layer activation function	Linear	Sigmoid
Dropout	0.5	None
Loss function	MSE	MSE
Learning rate	0.01	0.01
Optimizer	Adam	Adam

3.3 Evaluation methods

For the evaluation, we use RMSE between the subjective intelligibility and the estimated intelligibility, defined by equation (3).

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{3}$$

Here, x_i is the subjective intelligibility, and y_i is the estimated intelligibility. N is the number of samples in the test data set. In addition, Pearson's product-moment correlation coefficient is also used for evaluation. The formula is defined as follows,

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{4}$$

Here, $\bar{x}$ and $\bar{y}$ are the average of x_i and y_i respectively.

4. RESULTS AND DISCUSSIONS

Table 3 shows the RMSE and correlation coefficient between subjective intelligibility and estimated intelligibility when (1) using real clean speech, and (2) using estimated clean speech as the reference signal. In the closed test (training data), the correlation coefficient was 0.9721 and the RMSE was 0.0743, showing the same degree of accuracy as when the real clean speech was used. However, in the open test (test data), the correlation coefficient and RMSE were 0.8307 and 0.1839, respectively, and the accuracy deteriorated significantly. This is still relatively high estimation accuracy, given the fact that a clean reference signal is not used here. However, we believe there is still much room for improved estimation accuracy.

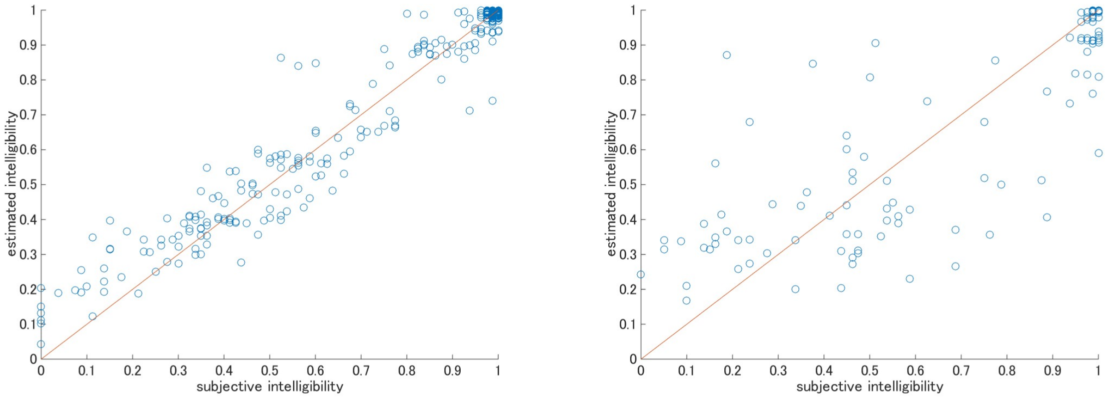
Figure 2 shows a scatter diagram of the subjective intelligibility and the estimated intelligibility on the training data (closed test), and on the test data (open test). In the closed test, it is shown that the plot points are well grouped along the diagonal line, at which the subjective and the estimated intelligibility match, and the speech intelligibility is being estimated with high accuracy. The open set test shows that the deviation of the plot points from the diagonal line is large, and the estimation accuracy is rather low. The following two points can be mentioned as the reasons for the low estimation accuracy of speech intelligibility in the open set test.

The first one is the low accuracy of DNN 1 constructed for noise removal. The model constructed at this time is a model tuned specifically for the read-word speech of a single female speaker. Noise reduction performance may be relatively low for (1) speech with low SNR that is completely masked by noise, and also (2) speech that is mixed with speech from other speakers, i.e., babble noise. It seems that the noise cannot be eliminated completely, and also the original speech is significantly distorted in these extreme cases.

The second cause is considered to be a lack of data used for training DNN 2 to estimate intelligibility. Subjective intelligibility used as supervisory data for training DNN is difficult to collect in large quantities because it involves a great deal of effort when making an evaluation. It seems that the lack of training data caused over-fitting, making it difficult to extract accurate feature values, in addition to the low noise removal performance.

Table 3 – RMSE and Pearson correlation between measured vs. estimated intelligibility

Ref. speech	Criteria	Training data	Test data
Real clean speech	RMSE	0.0674	0.1178
	Correlation	0.9759	0.9441
Estimated clean speech	RMSE	0.0743	0.1839
	Correlation	0.9721	0.8307



(a) Training data

(b) Test data

Figure 2 – Distribution of subjective vs. estimated speech intelligibility

5. CONCLUSIONS

In this paper, we investigated a non-reference speech intelligibility estimation using noise reduction. DNN was used for noise removal and speech intelligibility estimation. The noise component was removed from the logarithmic power spectrogram of the noisy speech, and fwSNRseg was calculated using the estimated signal as a reference signal. This calculated fwSNRseg was then used to train the DNN to estimate the speech intelligibility. The estimation accuracy of the trained DNN was evaluated using the RMSE and the Pearson's product-moment correlation coefficient. As a result, the estimation accuracy on the training data was comparable to when using the clean speech, but accuracy on the test data (unseen data) deteriorated compared to the case using clean speech. The low estimation accuracy may be due to low noise removal performance and low training data volume of the DNN for intelligibility estimation. We plan to improve the generalization of the model on unseen data by introducing a more sophisticated network topology to the DNN model used in the noise reduction instead of a simple densely connected topology model. We will also attempt to increase the amount of training data for intelligibility estimation.

ACKNOWLEDGEMENTS

This work was supported in part by the JSPS KAKENHI 17K00223.

REFERENCES

1. K. Kondo, K. Taira, Y. Kobayashi, "Binaural Speech Intelligibility Estimation Using Deep Neural

Networks", Proc. Interspeech, 2-6 September 2018, Hyderabad, Vol. 74 (1), pp.1858-1862

2. J. Ma, Y. Hu, and P. C. Loizou, J. Acoust. Soc. Am., Vol. 125, No. 5, pp. 3387-3405, May 2009.
3. K. Kondo, et al. "Two-to-one selection-based Japanese speech intelligibility test", J. of Jap. Acoust. Soc., vol. 63, no.4, p.196-205, 2007.4
4. W. D. Voiers, "Speech intelligibility and speaker recognition," Stroudsburg (PA): Dowden, Hutchinson & Ross, 1977. Ch. Diagnostic evaluation of speech intelligibility, p. 374–387.
5. JEIDA Noise Database, http://research.nii.ac.jp/src/en/JEIDANOISE.html