Interspeech 2018
2-6 September 2018, Hyderabad

Binaural Speech Intelligibility Estimation Using Deep Neural Networks

Kazuhiro Kondo[1], *Kazuya Taira*[1 2], *Yosuke Kobayashi*[3]

[1]Yamagata University, Japan
[2]Currently with Yamamoto Electric Corp., Japan
[3]Muroran Instutute of Technology, Japan

kkondo@yz.yamagata-u.ac.jp, taira.kazuya@ydk.jp, ykobayashi@csse.muroran-it.ac.jp

Abstract

We attempted to estimate the speech intelligibility of binaural speech signal with additive noise. The assumption here was that both the target speech signal and the noise source are directional sources. In this case, when the speech and noise sources are located away from each other, the intelligibility generally improves since the human auditory system can potentially segregate these two sources. However since intelligibility tests are commonly conducted using monaurally recorded signals, the intelligibility is often under-estimated compared to live human listeners since this segregation capability is neglected. We have previously proposed to use binaurally recorded signals to estimate the speech intelligibility and compared the estimation accuracy of several machine learning methods on this signal. We showed that random forests (RF) combined with the better ear model and Mel filter banks gives the highest accuracy compared to other methods, such as the support vector machines or logistic regression. In this paper, we attempt to introduce deep neural networks (DNN) to this task. Initial evaluation results show that the use of DNN can provide a modest improvement over RF.
Index Terms: speech intelligibility, objective estimation, binaural speech, DNN

1. Introduction

With most of the adult population in developed and emerging countries owning a mobile phone, speech communication is conducted in all sorts of environments. For example, it may be conducted in a reverberant office space with ample background speech and noise, on a busy street crossing with automobile noise, or a quiet private office. Thus, techniques for efficient and accurate speech quality assessment is necessary to guarantee an acceptable level of communication quality over these networks. Speech intelligibility is a measure that quantifies the identification accuracy of the received speech content over a transmission channel and is a crucial measure of speech communication quality [1, 2].

Speech intelligibility measurement is conducted by having human subjects listen to degraded read speech samples, and having them identify the contents of the samples, measuring the accuracy of this identification. The contents of the sample are typically syllables, words or sentences. The test stimuli need to cover all aspects of the language being tested, such as the phonetic context. The test also needs to be conducted by a large panel of human listeners so that the variations in the responses by individuals are averaged out. Thus, speech intelligibility tests are generally time-consuming and expensive.

Accordingly, research into the estimation of speech intelligibility without the use of human listeners have been conducted. For example, the Articulation Index (AI) by French and Steinberg estimates the speech intelligibility from the perceptual average SNR measurements in critical bands [3]. Steeneken and Houtgast proposed the Speech Transmission Index (STI), which uses artificial speech signals transmitted over the channel under test, and estimates the speech intelligibility by measuring the weighted average modulation depth of the received signal [4].

However, these estimation methods base their measurements on monaural signals. In a realistic environment, for example in a crowded classroom or a large conference hall, human subjects listen to speech signals using both ears. This can potentially lead to better identification of the test signal since the human auditory system can potentially discriminate sources traveling from different directions. In other words, if the test speech signal travels from a different direction than the noise source, the human listener may be able to selectively listen to the test signal. However, speech intelligibility is often measured using monaural signals, recorded using a single microphone. This can lead to a significant underestimation of the speech intelligibility since this will ignore the human ability to discriminate speech from noise when these arrive from different directions.

There has been work on the estimation of speech intelligibility from binaural signals. For example, Wijngaarden et al. have proposed an extension of the STI for handling binaural signals [5]. They have shown that speech intelligibility estimation accuracy on binaural signals can be improved compared to conventional STI that use monaural signals.

Recently, Liu et al. have proposed using blind source separation (BSS) to separate the target speech source and the masker in binaural signals, and estimate the intelligibility using objective estimation measures of the separated target speech [6]. They fed the separated target speech source to three binaural estimation methods; the binaural distortion-weighted glimpse proportion (BiDWGP), the binaural speech intelligibility index (BiSII), and the binaural speech transmission index (BiSTI). They have shown that relatively high accuracy is possible, but artifacts of the BSS block tend to affect the accuracy of this estimation. The same group of authors also recently proposed a non-intrusive method to estimate the binaural speech intelligibility [7]. They use deep neural networks (DNN) to blindly separate the target speech source from the competing sources, concurrently apply blind-source localization (BSL), and apply intrusive binaural estimation methods to the separated speech. The DNN is applied to the log-power spectra calculated from the Short-Time Fourier Transform (STFT) coefficients of the binaural signal. The separated speech source is fed to the BiDWGP and the Binaural Short-Term Objective Intelligibility (BiSTOI). They have shown very high estimation accuracy with both of these estimation measures for speech mixed with relatively stationary noise (babble, speech-shaped noise, and speech-modulated noise).

We have also been attempting to improve the estimation accuracy of binaural speech intelligibility [8, 9, 10]. We showed

10.21437/Interspeech.2018-27

that relatively high estimation is possible using a 16-band Mel filter-bank (Mel-16), a better ear model, and using Random Forests (RF) to map these objective measures to speech intelligibility [10]. In this paper, we further attempt to improve the estimation accuracy using DNN to estimate the intelligibility directly from the outputs of the Mel-16 filter-banks.

This paper is organized as follows. The next section outlines the estimation method for binaural speech intelligibility. This is followed by a brief introduction of the newly-introduced DNN used to estimate the binaural intelligibility in this paper. We then describe the evaluation experiment of the binaural speech intelligibility with the DNN and the previously-studied RF. Finally, we give the conclusion and plans for further research.

2. Estimation of binaural speech intelligibility

Figure 1 shows a block diagram of the binaural speech intelligibility estimation method being investigated. In this method, we attempt to estimate the binaural intelligibility of a mixed speech and noise source arriving from various directions. We assume that not only the target speech, for which we are measuring the intelligibility but also the noise source is a directive source. An example of such directive noise source is babble from a group of bystanders talking loudly from a specific direction or an automobile passing by from one direction to the other. Obviously, such directional noise source can potentially have a more profound effect on the target speech compared to diffuse noise sources.

We trained a mapping function between an objective measure calculated using the binaural signal to the subjective intelligibility. In order to do so, we first compiled a database of target speech traveling from various directions by convolving monaural target speech samples with the corresponding Head Related Transfer Functions (HRTFs). We also prepared noise sources from different directions by convolving this with the same HRTFs. Then, these two sources were mixed to compile a database of localized speech and noise with various azimuth combinations.

We compared the following three objective measures to model the binaural signal.

- The Better-Ear (BE) Model: Fig. 2 depicts the BE model, which selects either the left or the right channel based on the channel-wise SNR. The channel selection is conducted frame by frame. The SNR is calculated in sub-bands.
- The Band-wise Better-Ear (BBE) Model: Fig. 3 depicts the BBE model, which selects either channel for each of the sub-band based on the sub-band SNR of left and right channel. The selection is also conducted in each temporal frame.
- The Pooled Channel Model: Fig 4 depicts the pooled channel model, which simply splits the binaural signal into sub-channels, calculates the SNR by sub-band, and simply pools all SNR values in all sub-bands for both channels.

Two configurations for the sub-bands were used in the calculation of objective models. First, 25 critical bands used in the AI standard [3] was used as a reference. This sub-band configuration is commonly used in other measures such as the frequency-weighted SNR [11].

We also attempted to use the Mel-16 filter bank, where the frequency scale is converted into the Mel scale, and divided into equal Mel frequency bands. In previous work [10], we found that the use of 16 bands gives the highest prediction accuracy, and so we will be using this configuration here.

We conducted subjective intelligibility evaluations using the compiled database to collect a database of subjective intelligibility to be used as supervisory signals in the training. Twelve subjects, all in their early twenties with normal hearing, participated in these evaluations. The objective measure of each of the mixed signals is calculated, and the mapping function from this measure to the supervisory subjective intelligibility is trained. In [8, 9, 10], we attempted the use of some popular machine learning techniques, such as Support Vector Regression (SVR) and RF, to improve the mapping accuracy at all SNR ranges compared to the conventional Logistic Regression (LR).

The trained functions were then used to estimate the intelligibility of speech mixed with unknown noise.

3. DNN for speech intelligibility estimation

In this paper, we introduce a DNN for speech intelligibility estimation of binaural signals. Objective measures were calculated using each of the three methods described in the previous section. For the BE and BBE methods, one value for each of the sub-band is used as input to the DNN. For critical bands, this comes to 25 values, and for the Mel-16 sub-bands, 16 values. For the Pooled Channel model, one value for each of the stereo channel is output for each sub-bands, i.e. 50 values for the critical sub-bands, and 32 for the Mel-16 sub-bands.

From initial experiments, we found that the use of 2 hidden layers gave the best estimation accuracy. The number of units for each of the layers for the critical sub-bands were 128, 128, 128, and 1 for the pooled measure, and 64, 128, 64, and 1 for the other two measures. Likewise, for the Mel-16 sub-bands, they were 64, 128, 64, and 1 for the pooled measure, and 32, 64, 32, and 1 for the other two measures. These were also empirically chosen through initial experiments. The activation functions used were the exponential linear unit (ELU) for the hidden layers and the sigmoid for the output layer. The input was scaled using the RobustScaler, and batch normalization was used. The optimizer used was the Adamax optimizer. The dropout rate was set to 0.5, and the batch size to 64. The training generally stopped at about 100 epochs with the early stopping option.

4. Comparison of estimation accuracy using DNN with other machine learning methods

The speech intelligibility estimation accuracy was evaluated for localized speech mixed with localized competing noise at various azimuth combinations on the horizontal plane.

4.1. Experimental conditions

We selected 60 words (30 word-pairs) out of the Japanese Diagnostic Rhyme Test (DRT) word list [12], which is a Japanese version of the DRT, a two-to-one forced selection word intelligibility test [13, 14]. The words were read by one female speaker. Three noise samples were used. Two were selected from the JEIDA noise database [15]; A/C fan coil, and local train. In addition to these, babble noise was selected from the Signal Processing Information Base (SPIB) database [16]. We used speech mixed with babble and A/C fan coil noise at SNR lev-

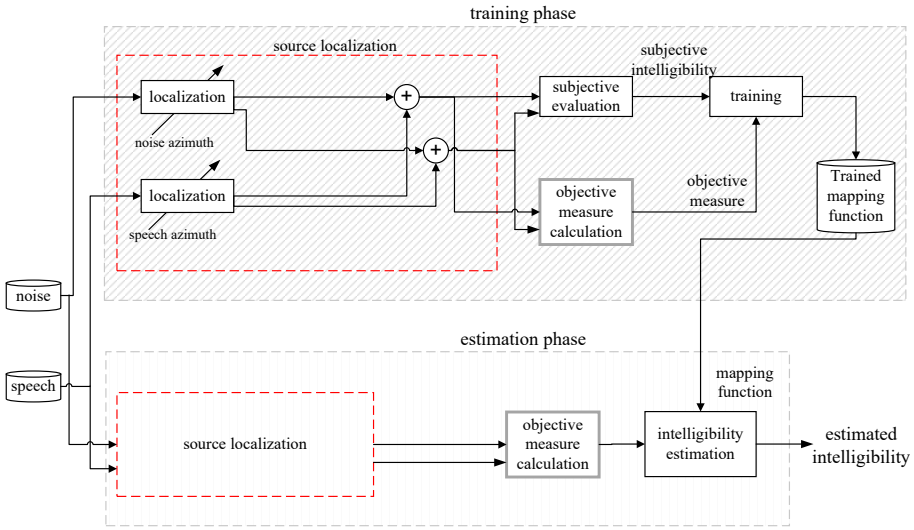

Figure 1: *Block diagram of speech intelligibility estimation*

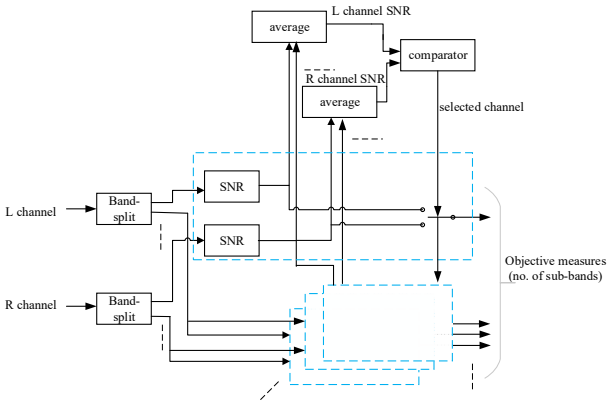

Figure 2: *Better-Ear objective measure*

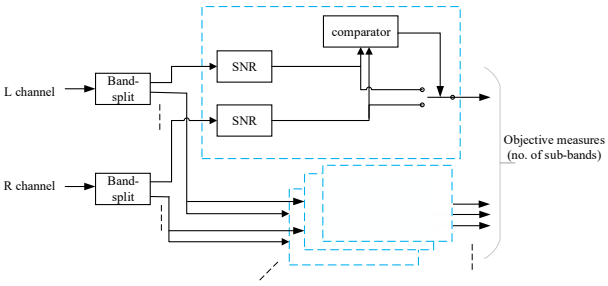

Figure 3: *Band-wise Better-Ear objective measure*

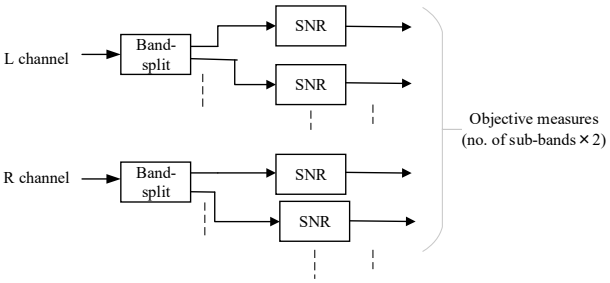

Figure 4: *Pooled-Channel Objective Measure*

els -6, -12, and -18 dB, respectively to train the models, and then tested the trained models on the local train noise, mixed at the same SNR levels.

Both the speech and noise samples were localized at various azimuths by convolving with the KEMAR HRTF, available from MIT [17]. The azimuths for the speech and noise sources were either 0 (directly in front of the listener), ± 45, or ± 90 degrees (positive degrees to the right, negative degrees to the left of the listener). All sources were located on the horizontal plane, at the same height as the listeners' ears. These signals were then split into sub-bands, its output fed to either the BE, the BBE, or the Pooled Channel model, and finally, the intelligibility is estimated from these using RF or DNN.

As the baseline for conventional estimation, we also estimated the intelligibility using a monaural signal. Thus, the stereo signals were mixed down to a single channel signal. Then, this signal was split into 25 critical bands. Classic LR was then applied to this measure to estimate the subjective intelligibility.

Estimation for this baseline and RF was reported in [10] and recited here for comparison.

4.2. Results and discussion

Tables 1 and 2 shows the RMSE and Pearson's correlation between the subjective and the estimated intelligibility with the various combinations of binaural objective measure calculation and mapping functions. As can be seen, DNN with Mel-16 sub-bands give the best results, with an RMSE of 0.115 and correlation of 0.938. This is modestly better than the best result with RF and Mel-16 sub-bands, which we reported in [10]. Thus, it seems that DNN can potentially improve the intelligibility estimation accuracy over other machine learning methods. However, we suspect that the amount of data used for training is rather limited in our experiments in order to decently train a DNN of this magnitude. We believe that the estimation accu-

Table 1: *RMSE Between Subjective and Estimated Intelligibility*

ML Function	Filter Bank	Front-end			
		BE	BBE	Pooled	Mono
LR (baseline)	Critical	-	-	-	0.208
RF	Critical	0.140	0.150	0.159	-
	Mel-16	**0.115**	**0.114**	0.126	-
DNN	Critical	0.141	0.122	0.139	-
	Mel-16	0.126	**0.115**	0.137	-

Table 2: *Pearson's Correlation Between Subjective and Estimated Intelligibility*

ML Function	Filter Bank	Front-end			
		BE	BBE	Pooled	Mono
LR (baseline)	Critical	-	-	-	0.568
RF	Critical	0.879	0.861	0.792	-
	Mel-16	0.920	0.921	0.891	-
DNN	Critical	0.903	0.914	0.882	-
	Mel-16	0.920	**0.938**	0.884	-

racy, as well as the generalization to other types of noise, can be further improved with more training data that include a wider range of noise types.

Fig. 5 plots the subjective vs. estimated intelligibility. Although there are some outliers, most of the estimations are close to the equal rate line (diagonal line) even for unseen noise data. On the other hand, Fig. 6 shows the plot for the baseline condition; use of a monaural signal with the critical band and logistic regression. As can be seen, the DNN estimation is a large improvement over this baseline condition. In fact, the baseline condition seems to fail to predict the lower SNR conditions completely. This may be because LR cannot generalize to noise conditions outside of the trained noise conditions, while DNN does a much better job of generalization.

5. Conclusion

We introduced DNN to estimate the speech intelligibility of binaural signals. Both the speech and competing noise were assumed to be directional sources, traveling from varying azimuths on the horizontal plane. Binaural signals were split into sub-bands, better ear models were applied, and DNN was used to map these output to estimated speech intelligibility. Estimation accuracy using DNN was compared to the accuracy using RFs, which we have shown in the previous study to give the most accurate results when combined with Mel frequency filter banks and the better ear model, which selects either left or the right channel with higher estimated SNR. It was shown that DNN can potentially improve the estimation accuracy compared to estimation using RFs, with an RMSE of 0.115 and Pearson Correlation of 0.938.

We believe we still have not trained the DNN with enough training data, and can still improve its accuracy with additional data, even though we already have a very accurate estimator. For starters, we would like to add more noise types to the degradation. Also, we currently use simple SNR values of each filter bank output for the objective measure. However, we would like to test with other distance measures that are motivated to simulate the auditory characteristics more accurately, such as the Log Area Ratio or Weighted Spectral Slope.

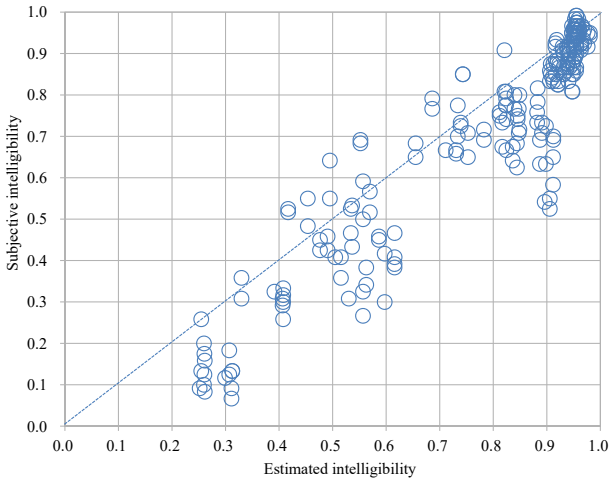

Figure 5: *Distribution of subjective vs. estimated Intelligibility using the band-wise better ear model and DNN*

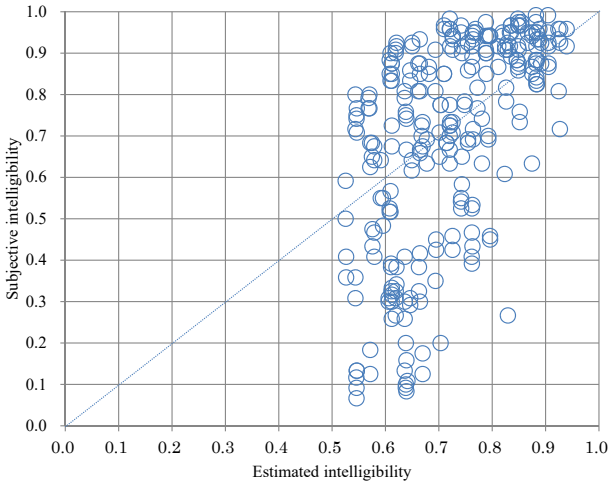

Figure 6: *Distribution of subjective vs. estimated Intelligibility using the monaural model and LR*

6. Acknowledgment

Initial DNN simulation environment was set up by Koki Yamada of Muroran Institute of Technology. The authors thank him for his expertise.

This work was supported in part by the JSPS KAKENHI Grant Numbers 25330182, 17K00223, and also the Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University (H29/A18).

7. References

[1] S. R. Quackenbush, T. P. B. III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[2] K. Kondo, *Subjective Quality Measurement of Speech*. Heidelberg, Germany: Springer-Verlag, 2012.

[3] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, 1947.

[4] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[5] S. J. Wijngaarden and R. Drullman, "Binaural intelligibility prediction based on the speech transmission index," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4514–4523, June 2008.

[6] Q. Liu, Y. Tang, P. J. B. Jackson, and W. Wang, "Predicting binaural speech intelligibility from signals estimated by a blind source separation algorithm," in *Proc. of Interspeech*, San Francisco, CA, Sept. 2016, pp. 140–144.

[7] Y. Tang, Q. Liu, W. Wang, and T. J. Cox, "A non-intrusive method for estimating binaural speech intelligibility from noise-corrupted signals captured by a pair of microphones," *Speech Communication*, vol. 2018, no. 96, pp. 116–128, 2018.

[8] K. Kondo and K. Taira, *Smart Innovation, Systems and Technologies*. Cham, Switzerland: Springer International, 2016, vol. 63, ch. Introduction and Comparison of Machine Learning Techniques to the Estimation of Binaural Speech Intelligibility, pp. 167–174.

[9] K. Taira and K. Kondo, "Estimation of binaural speech intelligibility based on the better ear model," *J. Acoust. Soc. Am.*, vol. 140, no. 4, Nov. 2016.

[10] K. Kondo and K. Taira, "Estimation of binaural speech intelligibility using machine learning," *Applied Acoustics*, vol. 2018, no. 129, pp. 408–416, 2018.

[11] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[12] M. Fujimori, K. Kondo, K. Takano, and K. Nakagawa, "On a revised word-pair list for the Japanese intelligibility test," in *Proc. Int. Symp. on Frontiers in Sp. and Hearing Res.*, Tokyo, Japan, Mar. 2006.

[13] W. D. Voiers, *Speech Intelligibility and Speaker Recognition*. Stroudsburg, PA: Dowden, Hutchinson & Ross, 1977, ch. Diagnostic Evaluation of Speech Intelligibility, pp. 374–387.

[14] ——, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, pp. 30–39, 1983.

[15] S. Itahashi, "A noise database and Japanese common speech data corpus," *J. Acoust. Soc. Japan*, vol. 47, no. 12, pp. 951–953, Dec. 1991, in Japanese.

[16] D. H. Johnson and P. N. Shami, "The signal processing information base," *IEEE Signal Processing Magazine*, vol. 10, pp. 36–42, Oct. 1993.

[17] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab, MIT Media Lab Perceptual Computing - Technical Report #280, May 1994, http://sound.media.mit.edu/resources/KEMAR/hrtfdoc.txt.