A Non-intrusive Binaural Speech Intelligibility Prediction for Clarity-2023

Katsuhiko Yamamoto[1]

[1]AI Lab, CyberAgent, Inc., Japan

katsuhiko@ieee.org

Abstract

A non-intrusive speech intelligibility (SI) prediction method is proposed for the second Clarity Prediction Challenge (CPC2). The model employs self-attention mechanisms and two types of multi-task learning to estimate speech segments and predict the SI of a target speech without the corresponding reference signal. The shared layer consists of latent representations of a deep neural network extracted from outputs of non-linear auditory filterbanks with individual hearing-impaired listeners' audiograms. Evaluation results with the development dataset for CPC2 show the proposed method outperforms the baseline, which needs the corresponding reference signal.

Index Terms: binaural speech intelligibility prediction, auditory model, deep-neural network, multi-task learning

1. Introduction

The Clarity Prediction Challenge (CPC) aims to investigate speech intelligibility (SI) prediction methods for enhanced speech through hearing-impaired (HI) listeners. In the second challenge (CPC2), the committee provides a dataset, including binaural audio processed by hearing aids (HA) for speech-in-noise, the corresponding clean reference speech signals, listeners' characteristics, and measured SI scores from listening tests.

This report proposes a non-intrusive SI prediction model with auditory-based binaural processing and deep neural network (DNN) architecture that directly converts acoustic representations from an auditory filterbank to the SI scores.

2. Proposed model

Figure 1 shows the overall architecture of the proposed model. Inputs of the proposed model are a stereo-enhanced speech processed by hearing aids and a listener's audiogram. The output is the predicted SI score of the input signal. The model has three stages: a pre-processing part, a shared layer, and a multi-task layer.

2.1. Pre-processing part

The input stereo speech signal is separated into the left and right channels and temporarily upsampled to $48,000$ Hz. An implementation version of the Gammachirp filterbank analyzes each monaural signal based on the characteristics of a HI listener and decomposes it into frame-based 100-ch excitation patterns [1]. Two parameters can be applied to the auditory filter: a listener's audiogram and a health factor of the compression characteristics. The individual health factor was determined stepwise from the average of the audiogram patterns at all frequencies, e.g., "NOTHING," "MILD," "MODERATE," and "SEVERE."

The auditory spectrogram is resampled to $10,000$ Hz and normalized using mean- and variance information. The frame length to normalizations is almost the same as 384 ms [2]. The normalization process with the long time-average frame enhances the slow fluctuations corresponding to amplitude modulation by speech signals.

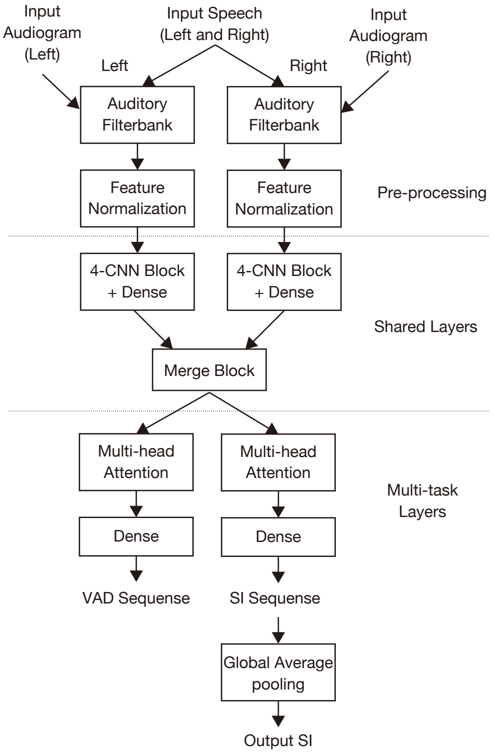

Figure 1: *Architecture of the proposed model.*

To summarize the above, two normalized auditory spectrograms (L-frame × 100-ch) of the left and right channels are used for the inputs of the proposed DNN architecture to train and predict a SI of a speech signal.

2.2. Shared layer

The shared layers are designed with four convolutional neural network (CNN) blocks and a merge block. A CNN block consists of three 2-D convolutional layers with a kernel size of 3×3 and rectified linear units (ReLU) activations, and the stride length of the final layer is 1×3 [3]. Finally, the output from the four 4-CNN blocks is flattened and converted to $L \times 128$ dimensions by a dense layer.

A merge block combines two outputs from the left and right channels of the CNN blocks. In this model, two latent representations are concatenated and fused by a dense layer with 128 ReLU nodes. Then, a dropout layer with a rate of 0.3 is added to the end of the block. The fused representations are split into two tasks, voice activity detection (VAD) and SI prediction.

2.3. Multi-task layer

Previous research shows that a multi-task learning (MTL) method improves the accuracy of speech intelligibility predictions [4]. Therefore, a simple VAD task is set as an MTL with the SI prediction task. The proposed model uses a multi-head attention mechanism for each task to gather task-specific information from shared features. In the proposed model, the number of heads is set to 128. Once the attention has been applied, we use a dense layer with one node activated with a sigmoid function. For each frame, the dense layer produces two predictions, the VAD probability $\hat{V}_{u_l}$, and the SI score $\hat{I}_{u_l}$ in each frame l of speech utterances. Finally, a global average pooling layer obtains the final SI score $\hat{I}_u$.

2.4. Objective function and Optimization

The loss function for each task can be defined as a combination of the mean-squared error (MSE) and binary cross-entropy using the following equations:

$$O = \frac{1}{U}\sum_{u=1}^{U}[(I_u - \hat{I}_u)^2 + \frac{1}{L_u}\sum_{l=1}^{L_u}(I_u - \hat{I}_{u_l})^2 - \frac{1}{L_u}\sum_{l=1}^{L_u}\{(V_{u_l}\log\hat{V}_{u_l}) + (1 - V_{u_l})\log(1 - \hat{V}_{u_l})\}], \quad (1)$$

where U is the total number of training speech utterances, L_u is the total number of frames, I_u is the correct SI for a single utterance, and V_{u_l} is an ideal VAD probability in each frame labeled as binary. In the study, we used the Adam optimizer with a learning rate of 0.0001 for the optimization process.

3. Experiments

3.1. Dataset

The data provided by CPC2 was used to train the proposed DNN model. To define the training (train) and development (dev) sets for each dataset, the CPC2 dataset was divided into the train or dev dataset. The data sizes of train and dev data were 2449/272 for CEC2.train.1, 2501/277 for CEC2.train.2, 2494/277 for CEC2.train.3, 5191/576 for CEC1.train.1, 4774/530 for CEC1.train.2, and 4598/510 for CEC1.train.3 respectively.

In addition, we made sequence label data of an ideal VAD probability V_{u_f} to support the MTL. The ideal VAD label was defined as the binary sequence in that the positive value starts at 2 sec and ends at 4 sec. Note that the ideal VAD label was only used for training.

3.2. Computational requirements

For training, we used AMD™ EPYC 7763 64-Core@2.45 GHz (total memory of 120 GB) and NVIDIA-A2 Tensor Core GPU. Training typically lasts between three and eight hours, using a batch size 2048.

3.3. Baseline model

The baseline model for CPC2 is an extended version of HASPI version 2 [5]. The model uses the input stereo speech, individual audiograms (left and right ear), and the corresponding reference speech. The input signal is separated into left and right channels for monaural processing by HASPI. Finally, the better SI score predicted by each HASPI is chosen as the final score.

Table 1: *Experimental results for development sets*

Dataset	Model	RMSE	NCC	KT
CEC2.train.1	Baseline	29.82	0.66	0.50
	Proposed	**28.23**	**0.73**	**0.56**
CEC2.train.2	Baseline	30.06	0.68	0.51
	Proposed	**27.47**	**0.76**	**0.58**
CEC2.train.3	Baseline	30.35	0.67	0.50
	Proposed	**27.09**	**0.75**	**0.52**
CEC1.train.1	Baseline	26.56	0.68	**0.43**
	Proposed	**21.62**	0.68	0.36
CEC1.train.2	Baseline	26.62	**0.69**	**0.43**
	Proposed	**22.63**	0.56	0.34
CEC1.train.3	Baseline	26.48	**0.67**	**0.43**
	Proposed	**22.19**	0.58	0.29

4. Results and discussions

Table 1 shows the results of experiments with the root-mean-squared error (RMSE), Normalized cross-correlation (NCC), and Kendall's tau (KT) between sets of the percent correct and the predicted SI. As a result, the proposed model predicted SI with less RMSE than the baseline system for all datasets of CEC2 and CEC1. The NCC and KT of results by the proposed method were higher than the baseline for datasets of CEC2.

The baseline model is designed as a "better ear" model to predict SI scores for each ear and select the higher value. However, the result indicates that DNN-based models may combine and use binaural representations processed by each auditory filter with individual hearing loss information.

5. Conclusions

A non-intrusive model is proposed to predict the SI of datasets for CPC2. The proposed model consists of a non-linear auditory filterbank, binaural sharing layers, and two types of attention layers for MTL. The evaluation results show the proposed non-intrusive method outperforms the intrusive baseline model.

6. Acknowledgements

The author would like to thank his colleagues for their valuable discussions and the CPC2 committees for providing essential datasets and analysis environments.

7. References

[1] T. Irino, "Hearing impairment simulator based on auditory excitation pattern playback: WHIS," *IEEE Access*, doi: 10.1109/ACCESS.2023.3298673, 2023.

[2] A. H. Andersen, J. M. de Haan, Z.-H. 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, 2018.

[3] R. Edo-Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids," in *Proceesdings of Interspeech 2022*, pp. 3944–3948, 2022.

[4] H.-T. Chiang, Y.-C. Wu, C. Yu, T. Toda, H.-M. Wang, Y.-C. Hu, and Y. Tsao, "HASA-Net: A Non-Intrusive Hearing-Aid Speech Assessment Network," in *proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 907–913, 2021.

[5] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI) Version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.