Conformer-based fusion of text, audio, and listener characteristics for predicting speech intelligibility of hearing aid users

Naoyuki Kamo[1], Kenichi Arai[1], Atsunori Ogawa[1], Shoko Araki[1], Tomohiro Nakatani[1], Keisuke Kinoshita[1], Marc Delcroix[1], Tsubasa Ochiai[1], and Toshio Irino[2]

[1]NTT Corporation, Japan
[2]Faculty of Systems Engineering, Wakayama University, Japan

naoyuki.kamo.ka@hco.ntt.co.jp

Abstract

We propose a speech intelligibility (SI) prediction method for the first Clarity Prediction Challenge (CPC1), which combines Conformer-based deep neural networks (DNNs) and the CPC1 baseline system. The DNN receives text, speech audio, and listener characteristics (audiogram, etc.) as its inputs and directly estimates the SI scores of the given speech. Then, we take an ensemble average of the SI scores obtained with the 10-best DNN models selected using our defined development set and the CPC1 baseline system. In experiments using the development set, the proposed method outperforms the baseline for both track 1 and track 2 scenarios.

1. Introduction

The first Clarity Prediction Challenge (CPC1) explores methods to predict speech intelligibility (SI) scores of noisy speech processed by hearing aids. The SI score is defined as the correct recognition rate (correctness) of words by hearing aid users. CPC1 organizers collected SI scores of the processed noisy speech samples from the users, and CPC1 participants develop methods to predict the SI scores using listener characteristics of the users. For the development, the participants can also use clean and processed speech audio signals, correct transcriptions, and the recognized word sequences by the users. This report describes our proposed method, a learning-based deep neural network model, that directly predicts the SI scores by fusing all the data provided for CPC1 except for the hearing loss model.

2. Proposed method

To predict the SI scores, our proposed DNN model receives four kinds of inputs, transcriptions, outputs from hearing aid (HA), clean speech signals convolved with the anechoic binaural room impulse responses (AE), and listener characteristics. Figure1 shows the structure of the DNN model. To achieve better prediction, we take an ensemble average of prediction scores from the multiple DNN models, which were trained by varying hyper parameters, and that from the baseline system [1].

2.1. Conformer-based word correct/incorrect prediction

A characteristic of our DNN model is to receive a sequence of word IDs corresponding to a correct transcription. With a sequence, since the DNN is aware of the number of words, the DNN model can predict binary labels for each word which indicate whether a word in a transcription is recognized correctly or incorrectly by an user. We refer this label as correct/incorrect label.

[1]https://github.com/claritychallenge/clarity_CC/tree/master/clarity_CPC1

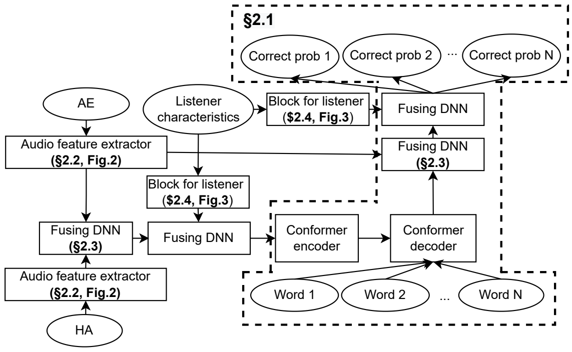

Figure 1: *Structure of prediction model*

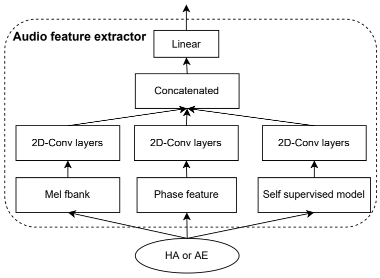

Figure 2: *Audio feature extractor: It integrates three features of Mel-fbank, phase, and output of self supervised model.*

Regarding the model structure, we adopted Conformer [1], also known as the state-of-the-art end-to-end automatic speech recognition model. It can handle two feature sequences with different lengths, i.e., word IDs and an audio signal (HA). The encoder receives HA and the decoder receives word IDs, then our DNN model yields a sequence representing the probability of each word being correct/incorrect.

For inference, we count the number of words with the probability of "correct" exceeding 0.5, and divide it by the total number of words to obtain the predicted correctness.

2.2. Audio feature extractor

Our audio feature extractor consists of three components, Mel filterbank, phase feature extractor, and self-supervised model (Wav2Vec2 [2] or HuBERT [3]). Figure 2 shows how three features are integrated.

We adopted a pre-trained model in s3prl[2] for the self-supervised model and froze its parameters during our DNN model training.

[2]https://github.com/s3prl/s3prl

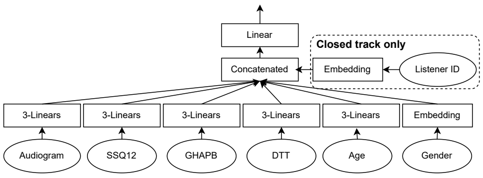

Figure 3: *DNN Block to input listener characteristic*

For the phase features, we used the phase difference between two microphones of each ear, i.e, $\sin(\theta_{1,f,t} - \theta_{0,f,t})$ and $\cos(\theta_{1,f,t} - \theta_{0,f,t})$, and the phase change over time, i.e., $\sin(\theta_{0,f,t+1} - \theta_{0,f,t})$ and $\cos(\theta_{0,f,t+1} - \theta_{0,f,t})$, where $\theta_{c,f,t}$ is the phase of the audio signal in time-frequency domain, and $c \in \{0, 1\}$, f, t denote the microphone channel at each ear, frequency index and time-frame index, respectively.

We concatenated the above three features extracted from each ear microphone, and put them into the convolution layer.

2.3. DNN structure to input clean speech

We also fed the AE into our DNN. As the DNN layer to input the AE, we compared two cases, before the conformer encoder or after the conformer decoder. In the former case, HA and AE features were simply concatenated, because their feature sequences had the same length. In the latter case, we use an additional conformer DNN to handle two series of different lengths, the AE feature sequence and the conformer decoder output.

2.4. DNN structure to input listener characteristic

We used all six listener characteristics (audiogram, SSQ12, GHAPB, DTT, age, gender) provided by CPC1. Moreover, in case of track 1 (listener seen), we also utilize the listener ID. The details of each characteristic are described in the documentation of CPC1 challenge [3].

Figure 3 illustrates the DNN structure for these listener characteristics. For audiogram, SSQ12, GHAPB and age, we used the given value as they are as features. For DTT, the threshold value is used as a 1-dim. feature, and Gender is treated as a binary label and input to DNN using an embedding layer.

As with the way of the AE input, we considered two input layers, i.e., before the conformer encoder or after the conformer decoder.

2.5. Multi task learning

In addition to correct/incorrect label prediction for each word (§2.1), we also introduce a loss to predict a scalar correctness directly, which is called as correctness loss, and system label classification loss to exploit multi task learning.

Correct/incorrect classification loss: We input the probability of correct/incorrect for each word to a classification loss function.

Correctness loss: We take the mean of the probability of ”correct” for each word as the predicted correctness and input it to a regression loss function.

System label classification loss: A system label distinguishes ten hearing aid systems used in CPC1. Since the response by each user is different depending on the audio quality of each system, we also make our DNN being aware of the system difference. We derive a predicted system label by inputting the Conformer encoder output averaged over time frames to a Linear DNN predictor and compute a classification loss for the system label.

.

2.6. Ensemble method

We created approximately 5,000 models with different hyperparameters, and selected the 10-best models from them using the evaluation results for our development set.

We averaged the predicted correctness for each sample by the selected 10-bests models and the baseline model[4] with equal weights to ensemble these models. This is our submission system, “Baseline + 10 best ens.”.

3. Data and resources

3.1. Data set

Our DNN model was trained using the data provided by CPC1. As an external corpus, LibriSpeech corpus [5] was used for the training of self supervised model in our feature extractor (§2.2).

CPC1 has closed(track1) and open(track2) scenarios. Evaluation data in the closed track were obtained only from systems (HAs) and listeners that are seen in the training dataset. In contrast, either one of or both systems and listeners in those of the open track are unseen in the training dataset.

To define the training (train) and development (dev) sets for each track, we divided the training/development dataset provided by CPC1 into two sets, so that both sets include all combinations of hearing aid systems and listeners. This means that our dev set was a closed condition (both listener and system were seen in our train set), even for evaluation of track2. The data sizes of our train and dev sets were 2510 and 2353 for track 1, and 1847 and 1733 for track 2, respectively.

The correct/incorrect labels were obtained based on DP matching between the word sequences of the correct transcriptions and the recognized word sequence by the users.

Speech data were downsampled to 16 kHz.

3.2. Data augmentation

To train our DNN model, we adopted three data augmentation techniques, speed/volume perturbation, time/frequency masking in STFT domain, and channel shuffling.

For speech and volume perturbation, we used sox [5] to modify the speed by a factor between 0.95 and 1.05 and the amplitude by a factor between 0.8 and 1.0. We used the same factors for the augmentation of HA and AE.

As for time/frequency masking, we followed a method in SpecAug. We set mask widths for time and frequency at 20 and 30 and the number of masks at 2. Time/frequency masking was applied only to HA.

We performed channel shuffling by switching left and right channels in HA, AE, and audiograms.

3.3. Computational requirements

For training, we used Intel®Xeon®CPU E5-2630 v4 @ 2.20 GHz (total memory of 378 GB) and GeForce RTX 2080 Ti. It takes approximately 3 hours for training. For inference, it takes about 5 seconds per sample when using CPU.

[3]https://claritychallenge.github.io/clarity_CPC1_doc

[4]Regarding the baseline, we determined the parameters of the logistic function for mapping the MBSTOI [4] measure to the speech intelligibility score by using our train set.

[5]http://sox.sourceforge.net/

Table 1: *Experimental results for dev-set*

	Track 1 (closed)			Track 2 (open)		
	RMSE	PCC	SRC	RMSE	PCC	SRC
Baseline	29.34	0.62	0.55	28.21	0.68	0.57
1 best	26.29	0.71	0.62	26.16	0.73	0.60
10best ens.	26.00	0.72	0.64	24.93	0.75	0.65
Baseline + 10best ens.	**25.69**	**0.73**	**0.65**	**24.77**	**0.76**	**0.66**

4. Result

Table 1 summarizes the root mean square error (RMSE) of the predicted SI score. We also evaluated Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRC). In Table 1, "Baseline + 10 best ens." is our submission model.

From the result, it can be seen for both tracks that the performance of the 1 best model exceeds the baseline performance in all evaluation measures. Moreover, the ensemble of 10 best models further improves the performance, and our submission model, "Baseline + 10best ens.", provides the best prediction of the SI score.

Note that our dev set for track 2 was not an open set as described in Sec. 3.1. We have already confirmed that our proposed method outperforms the baseline system by using an open set, which consists of track 1 samples of the systems and listeners that are not present in track 2. Details are omitted due to space limitation.

5. Conclusion

We proposed Conformer based prediction model fusing audio, transcription, and all provided listener characteristic (audiogram, SSQ12, GHAPB, DTT, age, gender) and our model outperformed the CPC1 baseline system for our development set.

We created approximately 5000 models with different hyper-parameters and achieved better result using 10-best ensemble models averaging the predicted correctness by each model.

Finally, we obtained further improvement appending the baseline system to these DNN models and we reported this ensemble model as our best model.

6. References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[3] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[4] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," Speech Communication, vol. 102, pp. 1–13, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639317302947

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210.