Technical Paper of E003 and E024: A Non-Intrusive Speech Intelligibility Prediction Using Binaural Cues and Time-Series Model with One-Hot Listener Embedding

Candy Olivia Mawalim, Xiajie Zhou, Shogo Okada, and Masashi Unoki

Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
{candylim, s2210112, okada-s, unoki}@jaist.ac.jp

Abstract

This paper describes our proposed speech intelligibility prediction system submitted to the second Clarity Prediction Challenge (CPC2). This challenge aims to facilitate the advancement of improved hearing aids through the automatic prediction of speech intelligibility. Our system was developed by integrating the equalization-cancellation model to mimic the central processing of binaural cues. Further, we trained a time-series model with a one-hot listener characteristic embedding layer to predict speech intelligibility. The preliminary evaluation of the development set showed that our proposed method could effectively predict speech intelligibility.

Index Terms: speech intelligibility, non-intrusive, one-hot embedding, equalization-cancellation model

1. Introduction

The speech intelligibility prediction method is a critical component in developing and optimizing hearing aids. By leveraging automated speech intelligibility methods, hearing aid manufacturers can create more effective and user-friendly devices that cater to individual needs and improve overall hearing experiences for individuals with hearing impairment. The second Clarity prediction challenge (CPC2) aims to find the optimal speech prediction method for hearing aids in a more realistic scenario[1]. The collected scenes in the CPC2 were derived from the first and the second Clarity Enhancement Challenge (CEC1 and CEC2). The CEC2 contains more variety of noise sources, the head is moving while talking, and the onset timing is less predictable.

Researchers have explored various approaches to automatically assess speech intelligibility in the first Clarity prediction challenge (CPC1) [1]. For instance, the method based on short-time objective intelligibility (STOI), namely the modified binaural STOI (MBSTOI) [2], was introduced as the baseline system. Further, machine learning approaches that simulated the process of automatic speech recognition systems and the related features, such as [3, 4], were also proposed and showed promising results.

This report describes our submitted proposed systems for predicting speech intelligibility in the second Clarity prediction challenge (CPC2). We report two systems: E003 and E024. The E003 is a lightweight method using a stackregressor with the input of the final embedding layer of pre-trained wavLM large [5]. Meanwhile, the E024 is our proposed method that utilizes a time-series processing with an equalization-cancellation model. The hearing loss condition was represented as a one-hot embedding layer of an audiogram.

[1] https://claritychallenge.org/docs/cpc2/cpc2_intro

2. Method

2.1. Feature Extraction

In the feature extraction, we corporate two types of waveform language model (wavLM) features, i.e., the final embedding layer output of the pre-trained model from the whole utterance and the time-series type (extracted from one-second overlap windowing). The wavLM is an extension of the HuBERT framework, which enables the pre-trained model to be used for speech recognition and related tasks. The choice of wavLM is based on its ability to extract fine-grained details from audio signals and its training with a language modeling objective, resulting in rich representations of speech and linguistic properties. Unlike traditional methods that use handcrafted features for predicting speech intelligibility, the wavLM model operates directly on raw waveform data, enabling it to capture complex patterns and dependencies present in the speech signal. This approach can be advantageous for tasks related to speech intelligibility.

2.2. Regression models

Figure 1 and figure 2 showed the block diagram of the E003 and the E024 systems, respectively. The main difference between E003 and E024 are from the regression models. In E003, we utilize a stackregressor to build a prediction model. In E024, we utilize a more complex model using Long short-term memory (LSTM) network with one-hot listener embedding layer. Additionally, we pass the improved SPIN waveform through an equalization-cancellation model [6] before extracting the wavLM features for as the input for the E024 system.

2.2.1. Stackregressor (E003)

The wavLM extracted from the mean of left and right signals were utilized as inputs for both the base-regressor and meta-regressor, which were used to calculate the ultimate speech intelligibility score. Although speech is a complex and multidimensional signal, the base-regressor is composed of linear regressor, support vector machines, and random forest to make predictions. While linear regression models have been successfully applied in various speech-related tasks, they may not be sufficient to capture the intricate patterns present in complex listening environments and individual listener variations that affect speech intelligibility. To address this, we employ support vector machines and random forests in addition to the linear regressor to capture a broader range of relationships and effectively handle non-linearities in the data. Finally, the predictions from each regressor were fed into the Ridge-CV meta-regressor,

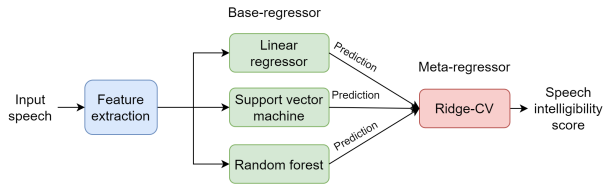

Figure 1: *Block diagram of E003*

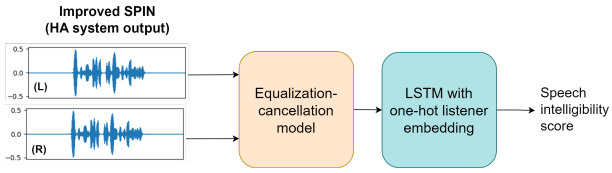

Figure 2: *Block diagram of E024*

creating an ensemble model that generates the final speech intelligibility score.

2.2.2. *LSTM with one-hot embedding (E024)*

The combination of LSTM and one-hot embedding offers several advantages for speech intelligibility prediction. First, it allows the model to work with discrete categorical data in a meaningful way, avoiding the risk of treating these categories as continuous variables, which could lead to inaccurate predictions. We extracted a one-hot embedding feature from the audiogram of the listener. We convert the power level in each frequency for the left and right ears into a vector with 16 units with a threshold of 60 dB. Second, the LSTM's ability to process sequential information ensures that the model can effectively capture temporal dependencies and dynamics present in speech signals, which are essential for determining intelligibility. The input of the LSTM model is the time-series wavLM feature from the equalization-cancellation model.

The hyperparameter settings were chosen carefully to optimize the model's performance. We set the number of LSTM units to 32 and used an embedding dimension of 16 to represent the audiogram. The sequence length was set to 8x1024 (4 seconds utterance) to capture sufficient context without overwhelming the model's computation. We employed a batch size of 4 for training efficiency. To enhance the performance of the model, we use a sequential self-attention with kernel regularization L2 and bias regularization L1. The attention regularization weight is set to 0.0001. The learning rate was set to 0.001, and we trained the model for 50 epochs using the Adam optimizer with mean squared error (MSE) loss function. These hyperparameter settings were chosen based on experimentation and yielded promising results in accurately predicting speech intelligibility.

3. Experiment

3.1. Dataset

We used the CPC2 dataset[2] for the experiment. The collected scenes were derived from CEC1 and CEC2. Each scene was simulated as a box-shaped room with one or multiple interference noises. The label was obtained from the subjective listening test from hearing-impaired people hearing sentences with 7 to 10 words spoken by a target speaker. There are three partitions in the dataset for cross-validation evaluation.

3.2. Evaluation setting

In the preliminary evaluation phase, we split the training data into training and development sets. We selected 30% listeners in the CEC2 subset for the development set and the remaining 70% listener of CEC2 and the whole CEC1 subsets for training the model. After obtaining the most optimized hyperparameters, we built the speech intelligibility prediction model using the entire train data to predict the evaluation data provided in the challenge.

We assess the performance of the model based on the regression task. To evaluate the model's accuracy, the typical regression metrics were utilized, including root mean squared error (RMSE), Pearson correlation (ρ_P), Spearman correlation (ρ_S), and R-squared (R^2). These metrics help quantify the difference between the predicted and actual correctness values obtained from the listening test.

3.3. Results

Table 1 shows the speech intelligibility prediction results in the development and testing phases. As mentioned earlier, we split the dataset by taking 30% listeners out of the training set from the CEC2 subset for testing our system in the development set. We compared the baseline HASPI and two proposed systems E003 and E024. In all development sets, the order of the most to the least accurate methods in terms of RMSE are E024, E003, and HASPI, respectively.

For the evaluation set by the CPC2 organizers, we obtained the overall prediction results as shown in the bottom row of Table 1. In this evaluation setting, the listener and the hearing aid system are unknown in the training data. The overall results showed that the E003 could predict a little bit better than the E024 in terms of RMSE and correlation but not significantly different. From these results, we can see the inconsistency between the results in the development and testing phases. It might happen due to the different settings in the evaluation (e.g., the prediction results of E024 might be more system-dependent than E003). More analysis of the prediction results will be performed in the future.

4. Limitation and Future Work

Our proposed methods were trained using only the embedding layer output of the wavLM pre-trained model. Thus, it introduces some disadvantages, such as the limitation in capturing other acoustic features that are prominent for predicting speech intelligibility. Additionally, the hearing loss model that might be beneficial to mimic the hearing perception of listeners with hearing impairment has not been included in the prediction model. The input of the equalization-cancellation model used in the E024 is the binaural waveform without any bandpass filter. Hence, the output waveform of the equalization-cancellation model used for feature extraction might be limited. These limitations with the considerably insufficient evaluation of this report will also be addressed in our future work.

5. Acknowledgment

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (No. 201605002),

[2] https://claritychallenge.org/docs/cpc2/cpc2_data

Table 1: *Speech intelligibility prediction results using development and evaluation sets*

Evaluation data	**Method**	ρ_P	ρ_S	**RMSE**	**stderr**	R^2
dev1	HASPI	0.6725	0.6780	30.1820	1.2435	0.4293
	E003	**0.7613**	**0.7671**	26.1529	1.0971	**0.5715**
	E024	0.7330	0.6248	**24.8069**	**0.5629**	0.5322
dev2	HASPI	0.7253	0.7288	27.2645	1.1422	0.5231
	E003	0.7110	0.7227	28.7021	1.1672	0.4715
	E024	**0.7316**	**0.6210**	**25.1148**	**0.5936**	**0.5240**
dev3	HASPI	0.6384	0.6235	30.2953	1.2734	0.4031
	E003	**0.7452**	**0.7403**	26.7215	1.1074	**0.5356**
	E024	0.7105	0.6017	**25.5201**	**0.6206**	0.5038
eval (all)	E003	**0.643**		**31.09**	**1.03**	
	E024	0.616		31.57	1.06	

a Grant-in-Aid for Scientific Research (B) (No. 21H03463), the Japan Society for the Promotion of Science (JSPS) KAKENHI grant (No. 22K21304, No. 22H04860, and No. 22H00536), and JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6).

6. References

[1] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. of Interspeech*. ISCA, 2022, pp. 3508–3512.

[2] 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.

[3] Z. Tu, N. Ma, and J. Barker, "Unsupervised uncertainty measures of automatic speech recognition for non-intrusive speech intelligibility prediction," in *Proc. of Interspeech*. ISCA, 2022.

[4] R. E. 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," pp. 3944–3948, 2022.

[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-training for Full Stack Speech Processing," 2021.

[6] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," *The Journal of the Acoustical Society of America*, vol. 136, pp. 768–776, 8 2014.