Non-intrusive Speech Intelligibility Prediction Model for Hearing Aids using Multi-domain Fused Features

Guojian Lin[1], *Fei Chen*[1]

[1] Southern University of Science and Technology, Shenzhen, China
12432635@mail.sustech.edu.cn, fchen@sustech.edu.cn

Abstract

Automatic speech intelligibility prediction system plays an important part in the development of hearing aids. In the second Clarity Prediction Challenge (CPC2), speech foundation models (SFMs) have shown remarkable performance in the task of speech intelligibility prediction. In this report, we propose a no-reference speech intelligibility assessment model for hearing aids that fuses multi-domain SFMs representations. Our model employs left and right ear branches to process input speech signals, fusing frame-level representations from three pretrained SFMs (Hubert, Whisper, and M2D-CLAP). Moreover, the proposed model utilizes Bi-LSTM and Multi-Head Self-Attention to process the fused representations in both frame and feature dimensions, calculating frame-level intelligibility scores. Finally, the model outputs the predicted intelligibility score through global average pooling of the frame-level scores. We evaluated the RMSE and correlation of models using representations from individual models and multi-domain fused representations on the development set of the third Clarity Prediction Challenge (CPC3). Experimental results show that the multi-domain fused features outperform any single-domain features.

Index Terms: speech intelligibility prediction, speech foundation model, multi-domain fused features

1. Introduction

Developing accurate automatic speech intelligibility assessment systems is crucial for the design of hearing aid algorithms. In recent years, several studies have utilized deep neural networks (DNNs) to create non-intrusive speech assessment models for hearing aids [1], [2], [3], [4]. Speech foundation models (SFMs) trained on broad and large-scale data have demonstrated robust generalization in downstream tasks. In the second Clarity Prediction Challenge (CPC2), researchers have found that adapting SFMs for intelligibility prediction achieved outstanding performance [5], [6]. We referenced the binaural branch and prediction module of the non-intrusive assessment model MBI-Net+ [7] in CPC2 and proposed our model for Clarity Prediction Challenge 3 (CPC3) by combining cross-domain representations extracted from three pretrained models (Hubert [8], Whisper [9], M2D-CLAP [10]) in SFMs. Experimental results demonstrate that the multi-domain feature fusion method effectively enhances performance on the development set compared to the use of single-domain features.

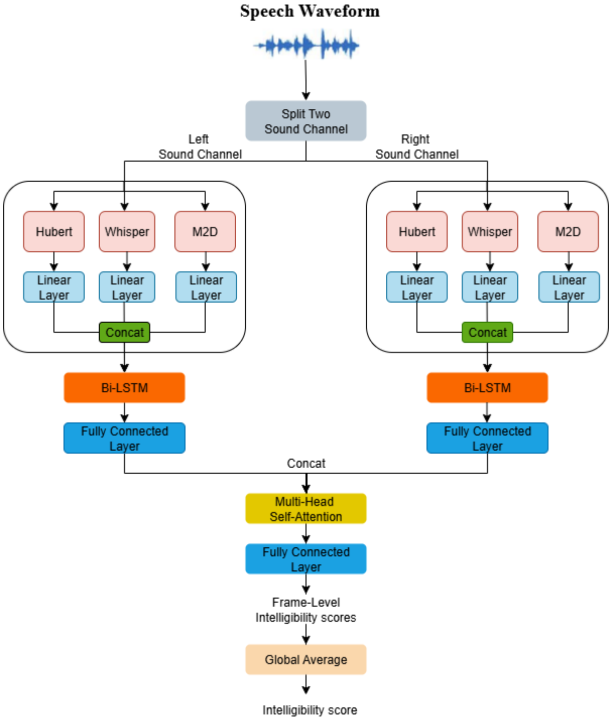

Figure 1: *Architecture of the proposed model*

2. Proposed Model

2.1. Model Architecture

The overall architecture of our proposed model is shown in Figure. 1. Dual-channel speech waveforms are input, representing the signals received by the left channel and right channel respectively. The signals are processed through the encoders of three pretrained models: HuBert-Large, Whisper-small, and M2D-CLAP, extracting self-supervised learning (SSL) representations, automatic speech recognition (ASR) representations, and general audio representations. Then, we used linear layers to project all representations to a unified dimension of 384. All the representations are concatenated along the frame-level to generate 384-dimensional multi-domain fused features. Then, fused features are fed into one Bi-LSTM layer with an input size of 384 units for temporal modeling, followed by a linear layer and a dropout layer. Subsequently, representations from the left and right ear branches are concatenated along the feature dimension. Then, the concatenated features are processed by a Multi-Head Self-Attention layer followed by a fully connected layer, which outputs the frame-level intelligibility scores. Finally, global average pooling is applied to obtain the final predicted intelligibility score.

Table 1: *Results on the development set in terms of RMSE and correlation*

Model	RMSE	Correlation
Hubert	0.75	27.45
Whisper	**0.80**	24.74
M2D-CLAP	0.77	26.66
Hubert + Whisper + M2D-CLAP	**0.80**	**24.18**

3. Experiments

3.1. Experiment setup

We trained the model on the training data of the CPC3 dataset and validated its performance on the development set, using Root Mean Square Error (RMSE) and correlation as performance evaluation metrics.

Training data was derived from CEC1 and CEC2, including 20 enhancement systems. For each speech enhancement system in training data, 90% of data were used for training and 10% for validation. Then, we merged data from all enhancement systems as train set and validation set, ensuring that the train set included all enhancement systems.

During training, we used MSE loss to calculate the errors between the predicted utterance scores, frame-level scores, and their corresponding ground truths. Adam optimizer was employed with an initial learning rate of 1e-4, along with dynamic learning rate decay strategy (patience = 10, factor = 0.1), and the training process involved 50 epochs.

3.2. Results and discussion

Table 1 shows the performance of using single-domain pretrained features and multi-domain fused features on the development set. We report the evaluation results using representations from single models and multi-domain fused representations, respectively. The results show that the evaluation performance using Whisper embedding representations alone is significantly superior to that using M2D-CLAP and Hubert features individually, while the performance of Hubert features alone is relatively poorer compared to the other two features. Multi-domain fused features integrating Whisper, M2D-CLAP, and Hubert features achieve the best performance among all model combinations, with an RMSE of 24.12 and a correlation of 0.80. It demonstrates that fusing self-supervised, automatic speech recognition (ASR), and general audio features can enhance the model's representation capability in complex environments and effectively improve model performance. It is worth noting that the performance of using multi-domain fused representations does not show a significant improvement compared to using Whisper representations alone, indicating that the fused representations mainly rely on the representation capability of Whisper.

We submitted the prediction results of the multi-domain fused features (Hubert + Whisper + M2D-CLAP) on the test set data of CPC3.

4. Conclusion

In the third Clarity Prediction Challenge, we proposed a non-intrusive intelligibility evaluation model using multi-domain fused features extracted from Hubert, Whisper and M2D-CLAP. Experimental results show that Whisper outperforms Hubert and M2D-CLAP in terms of intelligibility assessment performance on the development set. Furthermore, fusing the multi-domain representations of Whisper, Hubert, and M2D-CLAP enhances the robustness and adaptability to various acoustic conditions and effectively improves the performance of intelligibility prediction.

5. References

[1] 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 *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 907–913.

[2] 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," *arXiv preprint arXiv:2204.03305*, 2022.

[3] H.-T. Chiang, S.-W. Fu, H.-M. Wang, Y. Tsao, and J. H. Hansen, "Multi-objective non-intrusive hearing-aid speech assessment model," *The Journal of the Acoustical Society of America*, vol. 156, no. 5, pp. 3574–3587, 2024.

[4] R. Liang, Y. Xie, J. Cheng, C. Pang, and B. Schuller, "A non-invasive speech quality evaluation algorithm for hearing aids with multi-head self-attention and audiogram-based features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2166–2176, 2024.

[5] S. Cuervo and R. Marxer, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1421–1425.

[6] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate asr features and human memory models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 306–310.

[7] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Non-intrusive speech intelligibility prediction for hearing aids using whisper and metadata," in *Interspeech 2024*, 2024, pp. 3844–3848.

[8] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[10] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto, "M2d-clap: Masked modeling duo meets clap for learning general-purpose audio-language representation," *arXiv preprint arXiv:2406.02032*, 2024.