Non-Intrusive Multi-Branch Speech Intelligibility Prediction using Multi-Stage Training

Ryandhimas E. Zezario[1], *Szu-Wei Fu*[2], *Dyah A.M.G. Wisnu*[13], *Hsin-Min Wang*[1], *Yu Tsao*[1]

[1]Academia Sinica
[2]NVIDIA
[3]National Chengchi University

ryandhimas@citi.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw

Abstract

We propose three systems based on an improved multi-branch speech intelligibility prediction model (iMBI-Net) for the third edition of the Clarity Prediction Challenge. The base system integrates spectral, waveform, and Whisper-based features with severity-level audiogram information, processed through a multi-branch convolutional-bidirectional module with attention mechanisms. The second system, iMBI-Net-R, adds a single refinement module on top of the base model. The third system, iMBI-Net-E, includes three refinement modules, each using different acoustic inputs, with score-level ensembling across the branches. Experimental results confirm that multi-stage training with diverse objectives boosts performance, and iMBI-Net-E achieves the best results among all our submitted systems, demonstrating the effectiveness of our approach.

Index Terms: speech intelligibility, hearing aid, hearing loss, self-supervised learning, cross-domain features

1. Introduction

With the launch of two Clarity Prediction Challenges [1, 2], there has been growing interest in the development of non-intrusive speech intelligibility prediction models for hearing aids (HAs) [3, 4, 5, 6]. Results from both challenges show that incorporating richer acoustic features leads to overall better prediction performance. For example, in the first challenge, using hidden layer features from ASR models [6] and SSL models [7] resulted in better performance. In the second challenge, Whisper-based features were particularly effective, with all three top-performing non-intrusive systems leveraging Whisper for acoustic feature extraction [3, 4, 5].

Recently, the third edition of the Clarity Prediction Challenge aims to address a more challenging scenario, where only severity level information is available, without access to specific listener audiograms in the input features. Building on the notable performance of our previous MBI-Net+ model [5], we propose an improved multi-branch speech intelligibility prediction model, referred to as iMBI-Net. Specifically, we deploy three versions of the model. The first system is the base version of iMBI-Net, which integrates three types of acoustic features—spectral features, waveform features, and Whisper-based features—along with audiogram information representing hearing loss severity. The second system, named iMBI-Net-R, extends the base model by incorporating a single refinement module for score adjustment. The third system, iMBI-Net-E, employs three separate refinement modules, each utilizing distinct acoustic inputs to better capture diverse signal characteristics. Additionally, score-level ensembling is applied across the refinement outputs. Experimental results confirm that applying multi-stage training with different objective functions to the base module improves performance over intrusive baseline systems. The addition of refinement modules yields further gains, with iMBI-Net-E achieving the best performance among all our submitted systems, highlighting the effectiveness of our proposed approach.

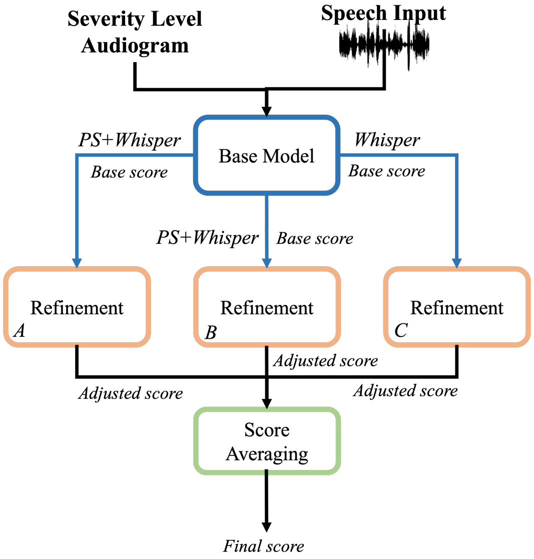

Figure 1: *Architecture of the iMBI-Net-E model.*

2. Proposed Systems

In this section, we present three proposed systems for the challenge: iMBI-Net (E011A), iMBI-Net-R (E011B), and iMBI-Net-E (E011C). The overall model architecture is shown in Fig. 1. The base model of iMBI-Net follows the architecture of our previous MBI-Net+ model [5]. However, unlike MBI-Net+, which incorporated the MSBG hearing loss model, our iMBI-Net directly integrates the relative severity level information into the model. Furthermore, unlike the original MBI-Net+, which used only mean square error (MSE) as the objective function, we introduce a multi-stage objective function. This includes a rank-based contrastive loss in the second stage, and a combination of MSE and Pearson correlation loss as the final objective. This strategy allows us to improve prediction performance without increasing the model size.

For iMBI-Net-R, the model consists of the base model and a refinement module, specifically using the type B refinement

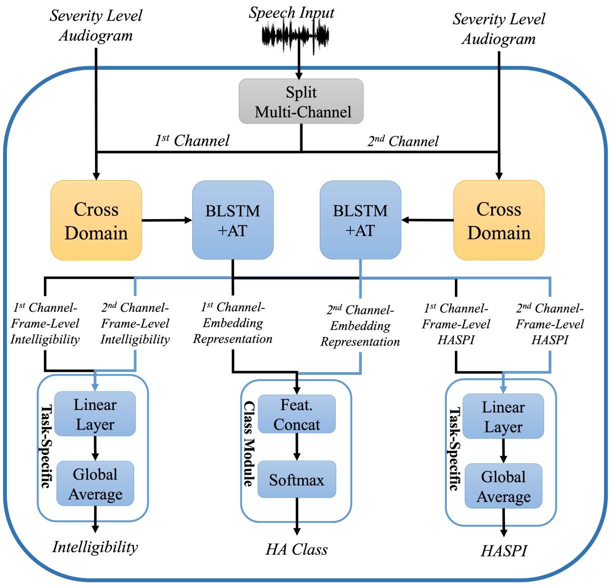

Figure 2: *Detailed architecture of Base-Model*

as shown in Fig. 3. In this system, the base model is fixed, and we use its estimated score and corresponding acoustic features to guide the refinement. The refinement module is designed to estimate the residual score, i.e., the difference between the base prediction and the ground-truth score. The training objective includes both MSE and Pearson correlation coefficient (PCC) losses for accurate residual estimation.

In the final model, iMBI-Net-E, we introduce two additional refinement branches, referred to as refinement A and refinement C. These differ in their input features and pooling strategies. For example, refinement A does not use an adapter layer and directly concatenates power spectral (PS) features with Whisper features, which are then passed through a BLSTM module. The rest of the architecture follows Fig. 3. In refinement C, only Whisper features are used as input to the BLSTM. Instead of attentive pooling, we apply mean pooling before concatenating the result with the base score and feeding it into an MLP, similar to refinement B in Fig. 3. Each refinement module is trained individually to ensure better stability. Finally, we stack the three refinement modules on top of the base model and perform score averaging to produce the final prediction.

3. Experiments

In this section, we present the experimental setup and results of iMBI-Net on the Clarity Prediction Challenge 2025 dataset.

3.1. Experimental Setup

The Clarity Prediction Challenge (CPC) 2025 dataset includes numerous systems carried over from the preceding Clarity Enhancement Challenge. Unlike the previous edition, only severity level information is available in the current edition. We split the dataset into 13,968 samples for training and 1,552 samples for additional validation. Furthermore, three evaluation metrics—root mean square error (RMSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC)—are used to assess the performance of MBI-Net. A lower RMSE indicates that the predicted scores are closer to the ground-truth scores (lower is better), while higher LCC and SRCC values indicate stronger correlations between the predicted and ground-truth scores (higher is better).

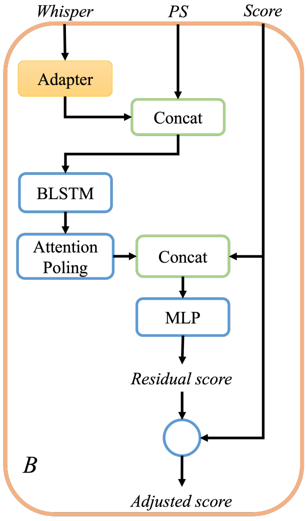

Figure 3: *Detailed Architecture of Refinement Model*

Table 1: *Evaluation scores of MBI-Net+ on our validation set.*

Systems	RMSE	SRCC	LCC
iMBI-Net (E011A)	20.7773	0.7985	0.8498
iMBI-Net-R (E011B)	20.5648	0.8031	0.8550
iMBI-Net-E (E011C)	20.4848	0.8076	0.8555

3.2. Experimental Results

As shown in Table 1, all variants of iMBI-Net achieve notably low RMSE scores. Furthermore, by adding a refinement module, iMBI-Net-R outperforms the base version of iMBI-Net. Interestingly, with the addition of two more refinement modules and the application of score averaging, iMBI-Net-E achieves the best overall performance, further demonstrating the advantages of using an ensemble model with appropriately designed refinement modules. Additionally, the base version of iMBI-Net currently submitted to the challenge development set leaderboard achieves an RMSE of 22.80, ranking third out of 33 systems.

4. Conclusion

In this paper, we introduce iMBI-Net, an improved multi-branch speech intelligibility prediction model, along with two enhanced variants: iMBI-Net-R and iMBI-Net-E. Through the integration of diverse acoustic features and multi-stage training strategies, our models demonstrated notable performance on the Clarity Prediction Challenge 2025 dataset. The addition of refinement modules and score-level ensembling further improved prediction accuracy, with iMBI-Net-E achieving the best results among our iMBI-Net systems. These findings highlight the potential of multi-branch and refinement-based architectures for advancing non-intrusive intelligibility prediction in hearing aid applications.

5. 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. INTERSPEECH*, 2022.

[2] J. Barker, M. A. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. ICASSP*, 2024, pp. 11 551–11 555.

[3] S. Cuervo and R. Marxer, "Speech Foundation Models on Intelligibility Prediction for Hearing-Impaired Listeners," in *Proc. ICASSP*, 2024, pp. 1421–1425.

[4] 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 *Proc. ICASSP*, 2024, pp. 306–310.

[5] 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 *Proc. INTERSPEECH*, 2024, pp. 3844–3848.

[6] Z. Tu, N. Ma, and J. Barker, "Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-Impaired Listeners," in *Proc. INTERSPEECH*, 2022, pp. 3488–3492.

[7] R. E. Zezario, F. Chen, C. S. Fuh, H.-M. Wang, and Y. Tsao, "Mbinet: a non-intrusive multi-branched speech intelligibility prediction model for hearing aids," in *Proc. INTERSPEECH*, 2022, pp. 3944–3948.