Deep Learning-based Speech Intelligibility Prediction Model by Incorporating Whisper for Hearing Aids

Ryandhimas E. Zezario[1][2], *Chiou-Shann Fuh*[1], *Hsin-Min Wang*[2], *Yu Tsao*[2]

[1]National Taiwan University
[2]Academia Sinica

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

Abstract

Improving the effectiveness of hearing aid (HA) devices in assisting users to understand speech in noisy surroundings is of utmost importance. To achieve this, it is critical to create a metric that can accurately forecast speech intelligibility for HA users. In our previous research, we introduced a non-intrusive multi-branched speech intelligibility prediction model known as MBI-Net. Building upon the promising outcomes of MBI-Net, our goal is to further enhance its performance by incorporating a pre-trained weakly supervised model called Whisper to enrich the acoustic features. We propose two versions of MBI-Net with these enhancements, namely MBI-Net+ and MBI-Net++. MBI-Net+ maintains the same model architecture as MBI-Net, featuring two branches, each consisting of a hearing loss model, a cross-domain feature extraction module, a task-specific layer, and a linear layer that produces the final output. Unlike the original MBI-Net, which relies on a self-supervised learning (SSL) model for deploying cross-domain features, MBI-Net+ adopts Whisper to deploy the acoustic features. Similarly, MBI-Net++ also employs Whisper for deploying the cross-domain features but with a more elaborate design, consisting of two branches, where each branch aims to predict the frame-level scores of intelligibility and HASPI (Hearing Aid Speech Perception Index), respectively. The predicted frame-level scores from each corresponding score are concatenated and fused using two different linear layers to produce the final prediction scores for intelligibility and HASPI.

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

1. Introduction

An accurate metric for predicting speech intelligibility is crucial to assess the performance of applications related to speech. The most reliable and straightforward approach to conducting evaluations is by conducting human listening tests. However, such tests are costly and less practical. With the emergence of deep learning models, several studies have successfully adopted these models to create automatic speech intelligibility prediction models [1, 2, 3, 4].

In the field of predicting speech intelligibility for hearing aids, various approaches have demonstrated strong predictive performance. For example, in [3], the utilization of hidden layer representations from automatic speech recognition (ASR) models as acoustic features for predicting speech intelligibility scores is elaborated. Furthermore, [4] presents a multi-branched speech intelligibility prediction model (MBI-Net), in which each branch comprises a hearing loss model, a cross-domain feature extraction module, and layers of convolutional neural network-bidirectional long-short term memory with attention mechanism (CNN-BLSTM-ATT). The outputs of these branches are then concatenated and fused in a linear layer to produce the final prediction performance.

In this challenge, owing to the notable performances demonstrated by MBI-Net [4], our objective is to present an enhanced version of MBI-Net. This enhancement involves leveraging a weak-supervision model, namely Whisper [5], to utilize acoustic features. The initial enhanced version of MBI-Net, referred to as MBI-Net+, maintains the same model architecture as the original MBI-Net. This architecture employs a multi-branched module that combines the outputs of each branch as input for the linear layer, which calculates the final intelligibility score. However, unlike the original MBI-Net, MBI-Net+ combines power spectral (PS), learnable filterbank (LFB), and Whisper to deploy cross-domain features, while the original MBI-Net uses a combination of PS, LFB, and a pre-trained self-supervised learning (SSL) model for feature deployment.

For the subsequent system, named MBI-Net++, Whisper is also utilized for deploying cross-domain features. However, MBI-Net++ employs a more intricate design, consisting of two branches, where each branch is dedicated to task-specific modules that forecast frame-level scores of intelligibility and the Hearing Aid Speech Perception Index (HASPI). The projected frame-level scores from each corresponding predicted scores are concatenated and merged through two distinct linear layers. These layers generate the final prediction scores for both intelligibility and HASPI. Experimental results confirm that MBI-Net++, which employs the HASPI score as additional information, can achieve an overall better root-mean-square-error (RMSE) and correlation score.

2. Proposed Systems

In this section, we present two proposed systems for the challenge: MBI-Net+ (E016) and MBI-Net++ (E023). For the MBI-Net+ model, the overall architecture of MBI-Net+ is depicted in Fig. 1. As illustrated in the figure, given dual-channel utterance, the audio undergoes processing by the MSBG hearing loss model [6, 7]. This processing modifies the speech signal according to the HA pattern and acts as a simulator to mimic the hearing ability of HA users. The simulated audio from the MSBG hearing loss model is subsequently split into two monaural speech signals, with the first and second channels corresponding to the left and right ears, respectively.

Following this, the first and second audio channels are processed by task-specific layers, which consist of cross-domain feature extraction and the CNN-BLSTM-ATT layer. These layers aim to predict frame-level intelligibility scores from the first and second channels, respectively. The detailed mechanism of the cross-domain feature extraction process is depicted in Fig. 2. This feature extraction module is composed of three components: (1) spectral features, obtained by converting speech

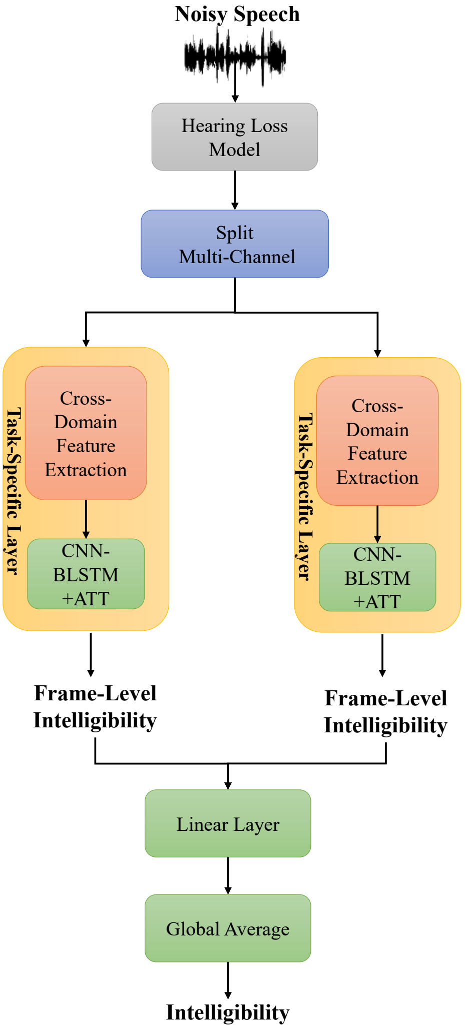

Figure 1: *Architecture of the MBI-Net+ model.*

waveforms through the short-time Fourier transform (STFT); (2) learnable filter bank (LFB) features [8]; (3) latent representations from the weakly supervised Whisper model [5].

Finally, the predicted frame-level intelligibility scores from the two branches are combined using a linear layer and global average pooling to obtain the final speech intelligibility score. To enhance training stability, the objective function for training MBI-Net+ comprises both frame-level and utterance-level scores, combined as follows:

$$
\begin{gathered}
O = \frac{1}{U}\sum_{u=1}^{U}[(I_u - \hat{I}_u)^2 + \frac{\alpha_m}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{im}_f)^2] + \\
L_{left} + L_{right} \\
L_{left} = \frac{\alpha_l}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{il}_f)^2 \\
L_{right} = \frac{\alpha_r}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{ir}_f)^2
\end{gathered}
\tag{1}
$$

where the terms L_{left} and L_{right} represent the frame-level loss associated with the left and right branches (referring to the ears). The symbols $I_u, \hat{I}u$ denote the actual and predicted intelligibility scores at the utterance level. The variable U denotes the total count of training utterances, while Fu represents the number of frames in the u-th training utterance. Additionally, $\hat{im}_f$, $\hat{il}_f$, and $\hat{ir}_f$ denote the predicted frame-level intelligibility scores of the main branch, left branch, and right branch respectively, for the f-th frame. The coefficients α_m, α_l, and α_r determine the balance between the losses at the utterance and frame levels.

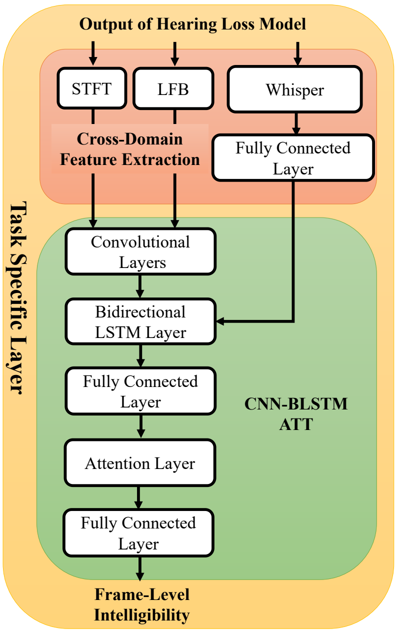

Figure 2: *Illustration of extraction cross-domain feature and obtaining frame-level intelligibility score on CNN-BLSTM+AT architecture.*

Furthermore, the architecture of the MBI-Net++ model is depicted in Figure 3. In general, MBI-Net++ adopts the same model architecture as the MBI-Net+ model. Additionally, within each task-specific layer, this module doesn't solely predict frame-level intelligibility scores, but also predicts frame-level HASPI scores. We assume that the supplementary information from HASPI might enhance the model's overall generalization capability. Subsequently, the corresponding frame-level scores are combined and integrated through two linear layers. This combination generates the final predictive scores for both intelligibility and HASPI ratings. The training objective for MBI-Net++ is defined as follows:

$$
\begin{gathered}
O = L_{Int} + L_{HASPI} \\
L_{Int} = \frac{1}{U}\sum_{u=1}^{U}[(I_u - \hat{I}_u)^2 + \frac{\alpha_m}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{im}_f)^2] + \\
L_{left-int} + L_{right-int} \\
L_{HASPI} = \frac{1}{U}\sum_{u=1}^{U}[(H_u - \hat{H}_u)^2 + \frac{\alpha_m}{F_u}\sum_{f=1}^{F_u}(H_u - \hat{hm}_f)^2] + \\
L_{left-haspi} + L_{right-haspi}
\end{gathered}
\tag{2}
$$

where $L_{left-int}$ and $L_{right-int}$ represent the frame-level loss of the left branch and right branch for estimating frame-level intelligibility, respectively; $L_{left-HASPI}$ and $L_{right-HASPI}$ represent the frame-level loss of the left branch and right branch for estimating frame-level HASPI, respectively. $\{H_u, \hat{H}_u, \hat{hm}_f\}$ denote the true utterance level score,

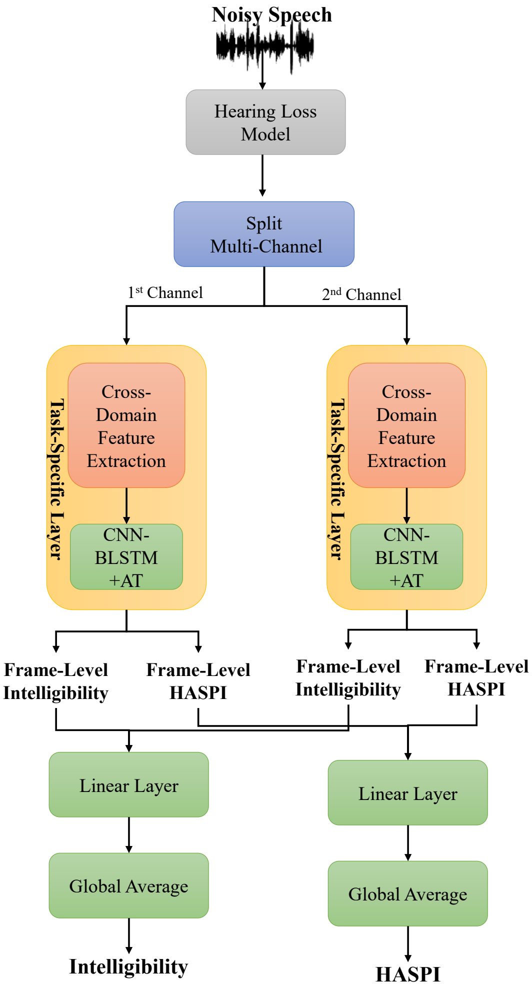

Figure 3: *Architecture of the MBI-Net++ model.*

predicted utterance-level score, and predicted frame level score of the HASPI, respectively.

3. Experiments

In this section, we present the experimental setup and results of MBI-net+ and MBI-Net++ on the Clarity Prediction Challenge 2023 dataset.

3.1. Experimental Setup

The Clarity Prediction Challenge (CPC) dataset for 2023 comprises numerous systems carried over from the preceding Clarity Enhancement Challenge in 2022. To elaborate, this dataset is categorized into three distinct tracks, and from within these tracks, we employ three speech assessment models. Additionally, our model was trained entirely on the CPC 2023 dataset while simultaneously deploying the MBI-Net+ and MBI-Net++ models. Two evaluation metrics, namely root mean square error (RMSE), and linear correlation coefficient (LCC), were used to evaluate the performance of MBI-Net. Lower RMSE indicates that the predicted scores are closer to the ground-truth scores (lower is better). In contrast, a higher LCC score indicates that the predicted score has a higher correlation to the ground-truth score (higher is better).

Table 1: *RMSE and LCC scores of MBI-Net+ and MBI-Net++*

Systems	Total Params	RMSE	LCC
MBI-Net+	3,441,863	26.79	0.754
MBI-Net++	3,540,686	**26.39**	**0.763**

3.2. Experimental Results

As indicated in Table 1, both MBI-Net+ and MBI-Net++ demonstrate the capability to attain notably low RMSE scores. This achievement underscores the advantage of incorporating Whisper for the deployment of cross-domain features. Interestingly, through the utilization of supplementary information from the HASPI score, MBI-Net++ achieves superior performance when contrasted with the MBI-Net+ model.

4. Conclusion

In this Clarity Prediction Challenge 2023, we have proposed two novel systems which leverage the acoustic features from Whisper, namely, MBI-Net+ and MBI-Net++. By leveraging Whisper embedding feature to deploy cross-domain features, our proposed systems can notably maintain a low RMSE score. In addition, by incorporating HASPI as an additional assessment metric, MBI-Net++ can achieve overall better prediction performance than the MBI-Net+ model.

5. References

[1] A. H. Andersen, J. M. D. 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.

[2] R. E. Zezario, S.-W. Fu, C.-S. Fuh, Y. Tsao, and H.-M. Wang, “STOI-Net: A deep learning based non-intrusive speech intelligibility assessment model,” in *Proc. APSIPA ASC*, 2020, pp. 482–486.

[3] 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*, 2022, pp. 3488–3492.

[4] 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 *Proc. Interspeech 2022*, 2022, pp. 3944–3948.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.

[6] T. Baer and B. C. J. Moore, “Effects of spectral smearing on the intelligibility of sentences in noise,” *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1229–1241, 1993.

[7] ——, “Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech,” *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2277–2280, 1994.

[8] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with sincnet,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.