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

Ryandhimas E. Zezario[12], Chiou-Shann Fuh[2], Hsin-Min Wang[1], Yu Tsao[1]

[1]National Taiwan University
[2]Academia Sinica

Introduction

- An accurate metric for predicting **speech intelligibility is crucial** to assess the performance of applications related to speech.
- The **most direct measure** of speech intelligibility is the **subjective listening test.**
- However, **such tests are costly and less practical.**

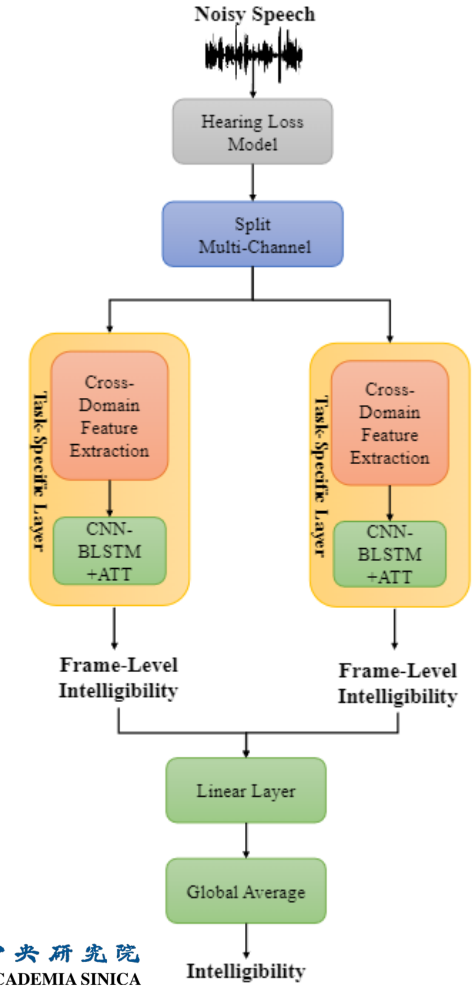
Introduction

- With the emergence of deep learning models, several studies have successfully adopted these models to create automatic speech intelligibility prediction models:

  1. Non-intrusive speech intelligibility prediction using convolutional neural network [1]
  2. STOI-Net: A deep learning based non-intrusive speech intelligibility assessment mode [2]
  3. Deep Learning-Based Non-Intrusive Multi-Objective Speech Assessment Model With Cross-Domain Features [3]
  4. Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listener [4]
  5. MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids [5]

Introduction

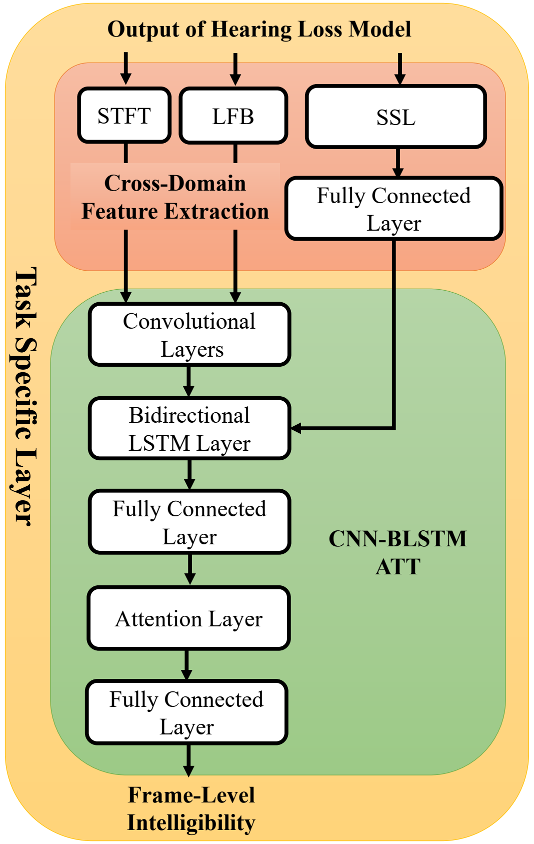
- In this challenge, owing to the notable performances demonstrated by MBI-Net [5], our objective is to present an enhanced version of MBI-Net by proposing MBI-Net+ and MBI-Net++.



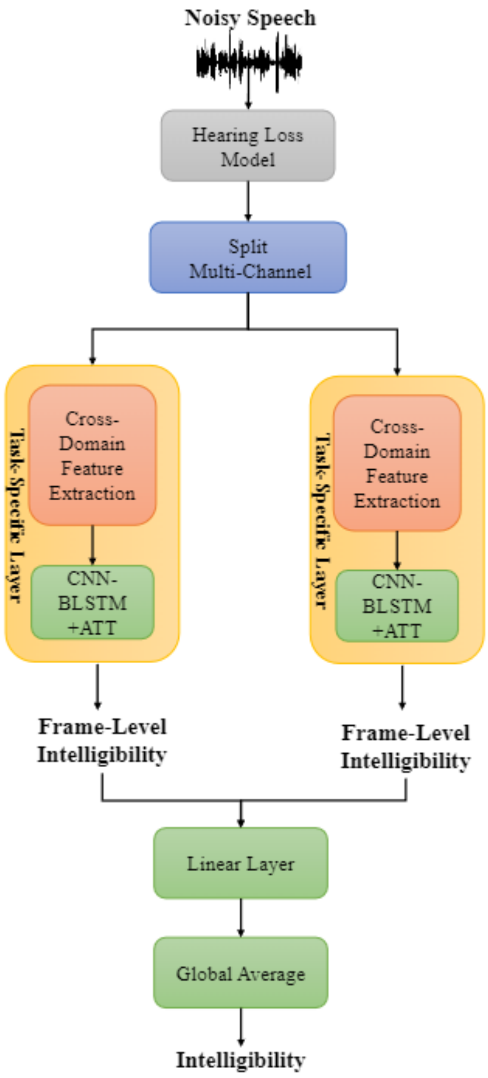
$$O = \frac{1}{U}\sum_{u=1}^{U}\left[(I_u - \hat{I}_u)^2 + \frac{\alpha_m}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i}_f)^2\right] + L_{left} + L_{right}$$

$$L_{left} = \frac{\alpha_l}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{il}_f)^2$$

$$L_{right} = \frac{\alpha_r}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{ir}_f)^2$$



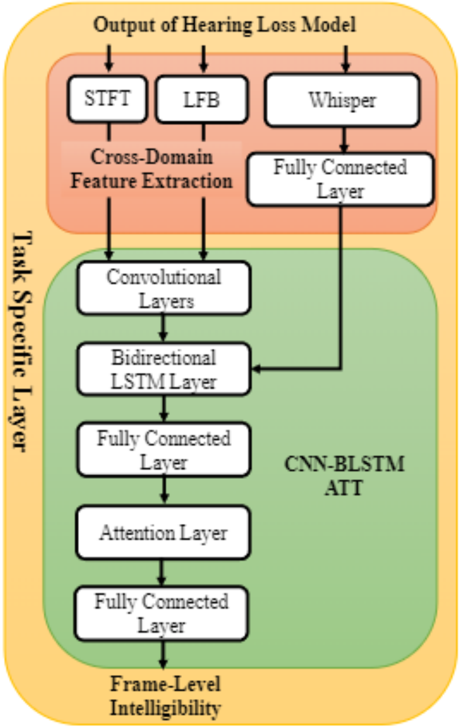
MBI-Net+



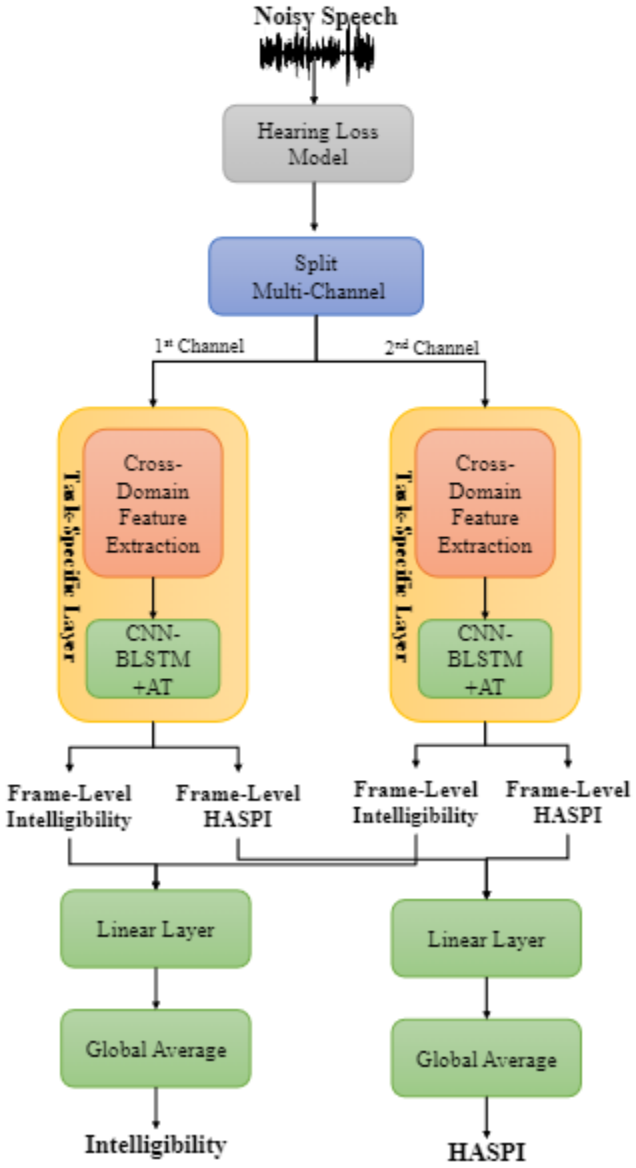
$$O = \frac{1}{U}\sum_{u=1}^{U}\left[(I_u - \hat{I}_u)^2 + \frac{\alpha_m}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i_{m_f}})^2\right] + L_{left} + L_{right}$$

$$L_{left} = \frac{\alpha_l}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i_{l_f}})^2$$

$$L_{right} = \frac{\alpha_r}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i_{r_f}})^2$$



MBI-Net++



$$O = L_{Int} + L_{HASPI}$$

$$L_{Int} = \frac{1}{U}\sum_{u=1}^{U}[(I_u - \hat{I}_u)^2 + \frac{\alpha_m}{F_u}\sum_{f=1}^{F_u}(I_u - \hat{i_{m_f}})^2] + L_{left-int} + L_{right-int}$$

$$L_{HASPI} = \frac{1}{U}\sum_{u=1}^{U}[(H_u - \hat{H}_u)^2 + \frac{\alpha_m}{F_u}\sum_{f=1}^{F_u}(H_u - \hat{h_{m_f}})^2] + L_{left-haspi} + L_{right-haspi}$$

Experiments

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.

Experiments

Experimental Results

RMSE: Root Mean Square Error
STDERR: Standard Deviation Error
LCC: Linear Correlation Coefficient

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**

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 intelligi-bility assessment model,” in Proc. APSIPA ASC, 2020, pp. 482–486.

[3] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, “Deep learning-based non-intrusive multiobjective speech assessment model with cross-domain features,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 54-70, 2023.

[4] 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.

[5] 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,” in Proc. Interspeech, 2022, pp. 3944–3948

Thank You