MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids

Ryandhimas E. Zezario[12], Fei Chen[3], Chiou-Shann Fuh[2], Hsin-Min Wang[1], Yu Tsao[1]

[1]National Taiwan University
[2]Academia Sinica
[3]Southern University of Science and Technology of China

Introduction

- A fair way to assess speech intelligibility is **critical for a variety of speech-related applications.**
- The **most direct measure** of speech intelligibility is the **subjective listening test.**
- However, **conducting large-scale hearing tests is prohibitive.**

Introduction

- A series of **speech intelligibility measures based on signal processing** have been proposed:

  - ❑Speech intelligibility index (SII)
  - ❑Extended SII (ESII)
  - ❑Speech transmission index (STI)
  - ❑Short-time objective intelligibility (STOI)
  - ❑Modified binaural short-time objective intelligibility (MBSTOI)

Introduction

- With the advent of deep learning (DL) models, several studies have used D**L models to deploy non-intrusive speech intelligibility prediction** models.

  - ❑To predict STOI [1,2,3]
  - ❑To predict subjective listening test results [4,5]

- **Few studies have focused on designing speech intelligibility prediction** models for HA users.
  - ❑ HASA-Net [6]: formulates the hearing loss pattern as a vector, which is combined with speech signals.

Introduction

- In our previous study, a multi-objective speech assessment model **(MOSA-Net) [7] was proposed to predict objective quality and intelligibility** metrics for normal hearing individuals

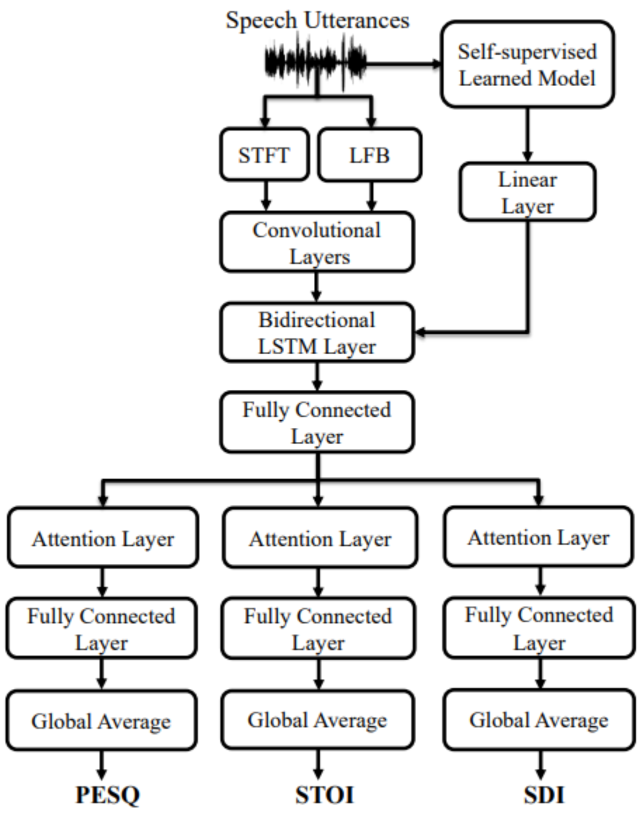

Fig. 1. Architecture of the MOSA-Net model.

Introduction

- In this study, we extend MOSA-Net and develop a speech intelligibility prediction model for HA, called the **multi-branched speech intelligibility prediction model (MBI-Net).**

MBI-Net

- MBI-Net consists of **two branches of model**, each characterizing one channel of speech signals in a binaural HA system.
- Each branch of MBI-Net consists of an MSBG model [8], a cross-domain feature extraction module, and a frame-level speech intelligibility prediction model.
- The MSBG model **modifies the speech signal according to the HA pattern** and serves as a **simulator to simulate the hearing ability** of HA users

MBI-Net

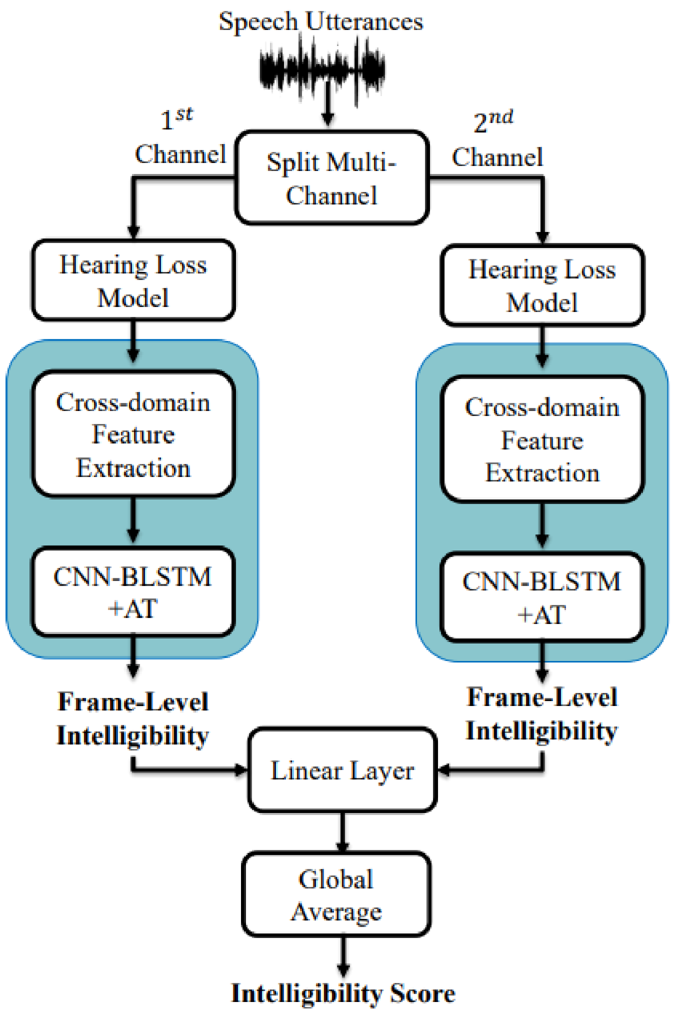

Figure 1: *Architecture of the MBI-Net model.*

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

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

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

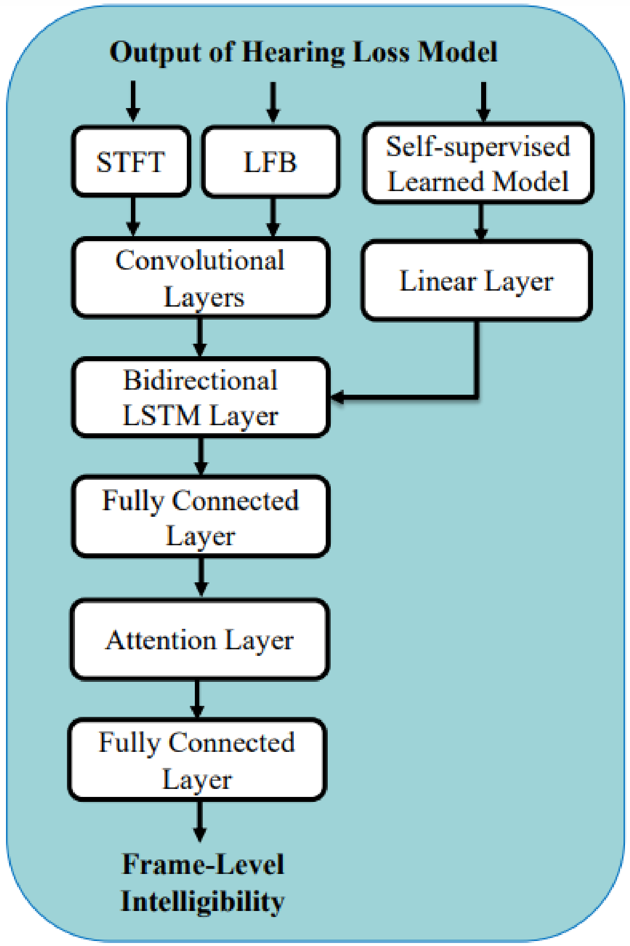

Figure 2: *Illustration of extraction cross-domain feature and obtaining frame-level intelligibility score on CNN-BLSTM+AT architecture.*

Experiments

Experimental Setup

- The Clarity Prediction Challenge dataset 2022 included **ten HA systems** from the previous Clarity Enhancement Challenge 2021 [9].

- Twenty-five HA users participated in the listening test, and each **listener was asked to answer what she/he heard** from a played speech sample.

- The **intelligibility score ranges from 0 to 100** (the higher the better).

- The training set consisted of two tracks, Track 1 and Track 2. Track 1 consisted of 4863 training utterances, and Track 2 consisted of of 3580 training utterances.

Experiments

Experimental Results

Table 1: *RMSE, Standard Deviation, and LCC scores of Let-Branch, Right-Branch, MBI-Net (Ave), and MBI-Net (Lin) on the closed-set (Track 1) dataset.*

Systems	RMSE	STDERR	LCC
Left-Branch	25.33	0.51	0.73
Right-Branch	26.24	0.52	0.72
MBI-Net (Ave)	25.12	0.51	0.74
MBI-Net (Lin)	**24.65**	**0.50**	**0.74**

Experiments

Experimental Results

Table 2: *RMSE, Standard Deviation, and LCC scores of Baseline, MBI-Net (Hub), and MBI-Net (WavLM) on the closed-set (Track 1) dataset.*

Systems	RMSE	STDERR	LCC
Baseline	28.52	0.58	0.62
MBI-Net (Hub)	24.65	0.50	0.74
MBI-Net (WavLM)	24.06	0.49	0.75
MBI-Net (WavLM+)	**23.05**	**0.46**	**0.78**

Table 3: *RMSE, Standard Deviation, and LCC scores of Baseline, MBI-Net (Hub), and MBI-Net (WavLM) on the open-set (Track 2) dataset.*

Systems	RMSE	STDERR	LCC
Baseline	36.52	1.35	0.53
MBI-Net (Hub)	30.72	1.22	0.59
MBI-Net (WavLM)	28.90	1.09	0.65
MBI-Net (WavLM+)	**24.36**	**0.96**	**0.75**

Experiments

Experimental Results

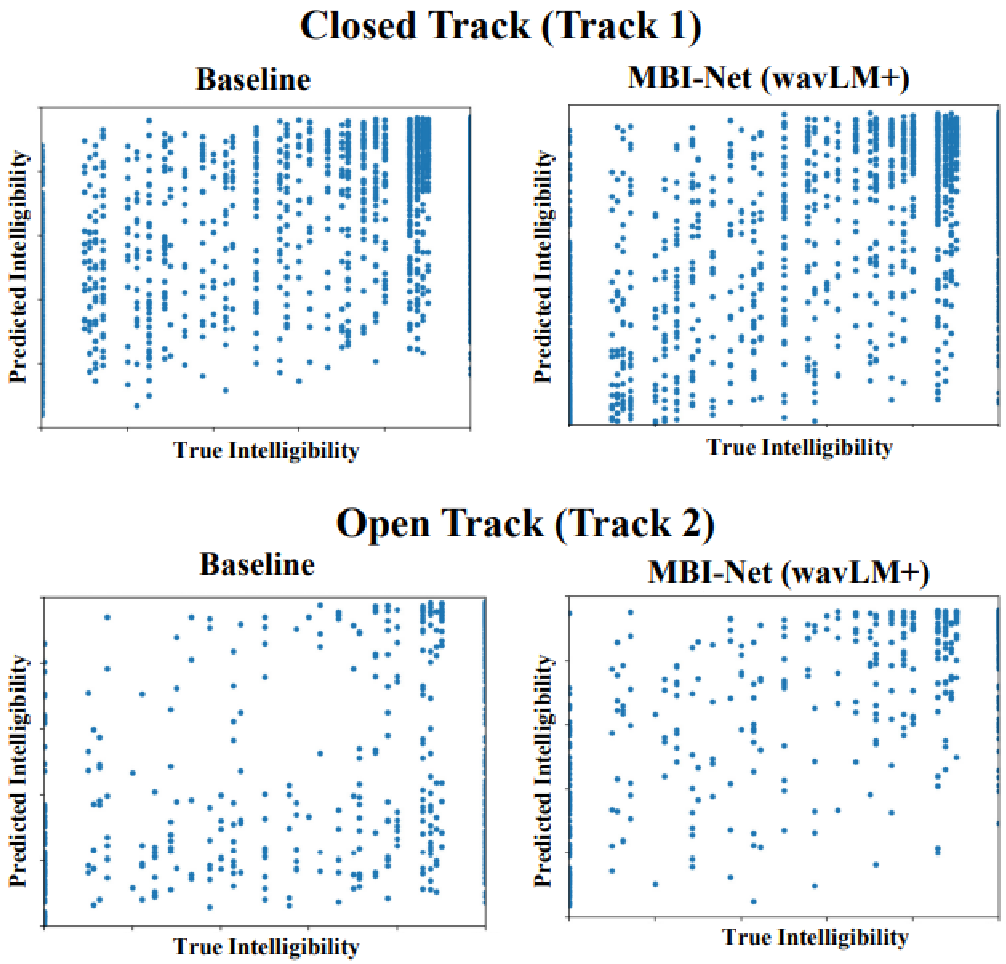

Figure 3: *Scatterplots of two speech intelligibility prediction models: Baseline and MBI-Net (WavLM+).*

Conclusion

- In this study, we presented MBI-Net, a multi-branched speech intelligibility prediction model for binaural HA users.

- MBI-Net adopts **two-branches of models corresponding to two speech channels** of the binaural HAs.

- Each branch of MBI-Net consists of an **MSBG model, a cross-domain feature extraction module,** and **the CNN-BLSTM+AT model architecture**.

- The outputs of the **two branches are then fused through a linear layer** to obtain the final speech intelligibility score.

Conclusion

- Experimental results from both Track 1 and Track 2 have **confirmed the advantages** of implementing the **multi-branched model** and using **cross-domain features** for achieving a better intelligibility prediction score.

- Furthermore, experimental results confirm the **advantages of WavLM in deploying representative SSL features.**

References

[1] X. Jia and D. Li, "A deep learning-based time-domain approach for non-intrusive speech quality assessment," in Proc. APSIPA ASC, 2020, pp. 477–481.

[2] X. Dong and D. S. Williamson, "An attention enhanced multitask model for objective speech assessment in real-world environments," in Proc. ICASSP, 2020, pp. 911–915.

[3] 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.

[4] A. H. Andersenan, J. M. 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.

[5] M. B. Pedersen, A. H. Andersen, S. H. Jensen, and J. Jensen, "A neural network for monaural intrusive speech intelligibility prediction," in Proc. ICASSP, 2020, pp. 336–340.

[6] 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), 2021, pp. 907–913.

[7] 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," arXiv:2110.02635, 2022.

[8] 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

[9] S. Graetzer, J. Barker, M. A. T. J. Cox, J. F. Culling, G. Naylor, E. Porter, and R. V. Munoz, "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing," in Proc. Interspeech, 2021.

Thank You