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

Ryandhimas E. Zezario¹², Fei Chen³, Chiou-Shann Fuh², Hsin-Min Wang¹, Yu Tsao¹

¹National Taiwan University
²Academia Sinica
³Southern 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 DL 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.
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.
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

![Diagram of MBI-Net model](image)

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

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

The objective function is given by:

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

where:

- \( O \): Overall loss
- \( U \): Number of utterances
- \( I_u \): Actual intensity
- \( \hat{i}_u \): Predicted intensity
- \( \hat{i}_f \): Predicted intensity for frequency bin \( f \)
- \( F_u \): Number of frequency bins for utterance \( u \)
- \( L_{left} \) and \( L_{right} \): Loss functions for left and right channels, respectively

The loss functions are:

- \( L_{left} = \frac{\alpha_u}{F_u} \sum_{f=1}^{F_u} (I_u - \hat{i}_f)^2 \)
- \( L_{right} = \frac{\alpha_u}{F_u} \sum_{f=1}^{F_u} (I_u - \hat{i}_f)^2 \)
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 3580 training utterances.
### Experiments

#### Experimental Results

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

<table>
<thead>
<tr>
<th>Systems</th>
<th>RMSE</th>
<th>STDERR</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left-Branch</td>
<td>25.33</td>
<td>0.51</td>
<td>0.73</td>
</tr>
<tr>
<td>Right-Branch</td>
<td>26.24</td>
<td>0.52</td>
<td>0.72</td>
</tr>
<tr>
<td>MBI-Net (Ave)</td>
<td>25.12</td>
<td>0.51</td>
<td>0.74</td>
</tr>
<tr>
<td>MBI-Net (Lin)</td>
<td><strong>24.65</strong></td>
<td><strong>0.50</strong></td>
<td><strong>0.74</strong></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Systems</th>
<th>RMSE</th>
<th>STDERR</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>28.52</td>
<td>0.58</td>
<td>0.62</td>
</tr>
<tr>
<td>MBI-Net (Hub)</td>
<td>24.65</td>
<td>0.50</td>
<td>0.74</td>
</tr>
<tr>
<td>MBI-Net (WavLM)</td>
<td>24.06</td>
<td>0.49</td>
<td>0.75</td>
</tr>
<tr>
<td>MBI-Net (WavLM+)</td>
<td><strong>23.05</strong></td>
<td><strong>0.46</strong></td>
<td><strong>0.78</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Systems</th>
<th>RMSE</th>
<th>STDERR</th>
<th>LCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>36.52</td>
<td>1.35</td>
<td>0.53</td>
</tr>
<tr>
<td>MBI-Net (Hub)</td>
<td>30.72</td>
<td>1.22</td>
<td>0.59</td>
</tr>
<tr>
<td>MBI-Net (WavLM)</td>
<td>28.90</td>
<td>1.09</td>
<td>0.65</td>
</tr>
<tr>
<td>MBI-Net (WavLM+)</td>
<td><strong>24.36</strong></td>
<td><strong>0.96</strong></td>
<td><strong>0.75</strong></td>
</tr>
</tbody>
</table>
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


Thank You