Technical Report for Clarity Prediction Challenge 2022

SpeechTeamAS

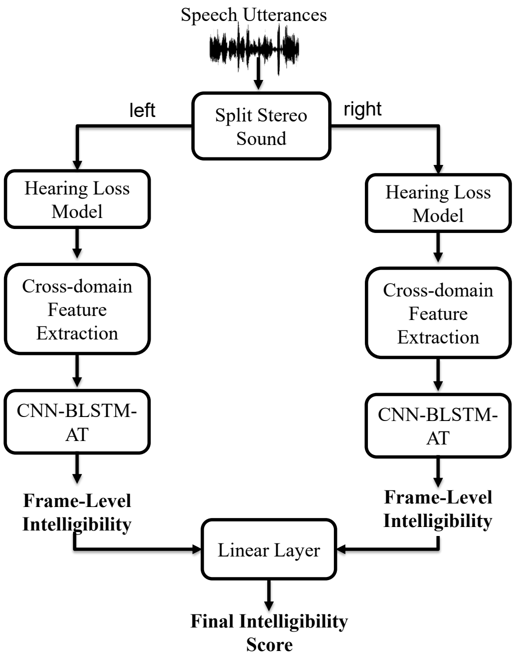
Speech Utterances
Split Stereo Sound
left
right
Hearing Loss Model
Hearing Loss Model
Cross-domain Feature Extraction
Cross-domain Feature Extraction
CNN-BLSTM-AT
CNN-BLSTM-AT
Frame-Level Intelligibility
Frame-Level Intelligibility
Linear Layer
Final Intelligibility Score

Figure 1: *Architecture of the proposed systems.*

1. Overview

In this Clarity Prediction Challenge 2022, We propose two different systems for performing speech intelligibility prediction. Specifically, both systems use the same cross-domain features (spectral and time-domain features). However, different self-supervised learning (SSL) pre-trained model is selected to deploy the latent representations feature. These three generated features are selected as the input features to the model. For Team 16, we selected Hubert [1] model to deploy the SSL feature, and for Team 33, we selected wavLM [2] to deploy the SSL feature. It is noteworthy that Hubert was developed to generate the feature for the ASR Task. While wavLM was developed to generate better features for the speech processing task. Therefore, we intend to deploy two different systems to analyze which SSL feature performs the best for speech intelligibility prediction on hearing aid. It is noteworthy that the speech utterance is first processed by the MSBG [3] hearing loss model before developing the cross-domain features.

For the model architecture, we followed our previous work [4, 5] It comprises a convolutional neural network and bidirectional long short-term memory (CNN-BLSTM) architecture for representation extraction, a multiplicative attention layer, and a fully-connected layer. As in this challenge, the speech utterance is a multi-channel; we intend to develop two different branches, where each branch corresponds to the specific channel. Next, the output of each channel is concatenated and finally optimized in the additional linear layer before finally getting the final prediction score. The detail of our system diagram is shown in the Figure. 1.

It is noteworthy that we did not perform any data augmentation during training the model. We only selected the available training samples to train the Track 1 and Track 2 models. Besides, only the corresponding intelligibility score is used as the model's ground truth, and we did not attach any additional label during training the model.

2. References

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[2] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," 2021. [Online]. Available: https://arxiv.org/abs/2110.13900

[3] Z. Tu, N. Ma, and J. Barker, "Optimising Hearing Aid Fittings for Speech in Noise with a Differentiable Hearing Loss Model," in *Proc. Interspeech 2021*, 2021, pp. 691–695.

[4] 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.

[5] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *arXiv:2110.02635*, 2022.