Domain-Adapted Automatic Speech Recognition with Deep Neural Networks for Enhanced Speech Intelligibility Prediction

Haeseung Jeon[1], *Jiwoo Hong*[2], *Saeyeon Hong*[1], *Hosung Kang*[2], *Bona Kim*[1], *Se Eun Oh*[2], *and Noori Kim*[*3]

[1]Ewha Womans University, Division of Artificial Intelligence and Software, South Korea
[2]Ewha Womans University, Dept. of Computer Science and Engineering, South Korea
[3]Purdue University, School of Engineering Technology, United States

{haeseungjeon, hjiwoo0914, saeyeonhong}@ewha.ac.kr, kanghsung717@ewhain.net,
{kimbn, seoh}@ewha.ac.kr, kim4147@purdue.edu

Abstract

While previous studies have shown that adapting Automatic Speech Recognition (ASR) models can outperform intrusive methods, many existing approaches still rely on pre-trained ASR models without domain-specific adaptation. In this work, we investigate the effect of fine-tuning ASR models using a domain-specific signal dataset to improve representation quality. Furthermore, we conduct a comparative evaluation of two prominent Deep Neural Network (DNN) architectures for audio modeling, such as Convolutional Neural Networks (CNNs) and Transformers. Notably, both models outperform the Hearing Aid Speech Perception Index (HASPI) score, with the Transformer-based model demonstrating higher performance due to its ability to capture global contextual information.

Index Terms: domain adaptation, deep neural network (DNN), speech recognition, computational paralinguistics

1. Introduction

Predicting speech intelligibility, a core focus of the Clarity Prediction Challenge (CPC), is crucial for improving communication and understanding individuals with hearing impairments. Kate et al. suggest an intrusive way for the prediction, the Hearing Aid Speech Perception Index (HASPI), which is calculated by comparing the coherence between the reference signal and the signal processed through a hearing aid, using an auditory model that simulates both normal-hearing and hearing-impaired listeners [1]. However, for the speech intelligibility task, the HASPI demonstrates certain limitations. Figure 1 presents the relationship between the HASPI score and the corresponding correctness (the percentage of correctly identified words) in the CPC3 dataset. Although the overall trend resembles a sigmoid curve, there remains considerable variation in the correctness. Furthermore, HASPI's predictive mechanism is predominantly bottom-up (signal-driven) rather than integrating top-down (context-driven) cognitive processes and also requires a clean, unprocessed reference speech signal for comparison against the degraded or processed signal, a condition often absent in real-world scenarios.

To address the latter issue, the previous work developed non-intrusive systems with pre-trained large acoustic models, such as Automatic Speech Recognition (ASR) [2]. One such attempt was suggested by Mogridge et al. [3], which proposed a model based on Whisper [4], a large pre-trained ASR model to extract rich features from the signal. They enhanced prediction performance by applying a Bidirectional LSTM (Bi-LSTM) with attention pooling to these features, showing significantly improved intelligibility estimation compared to HASPI.

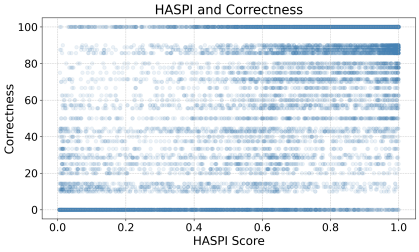

Figure 1: *Comparison of correctness and the HASPI score.*

Moreover, Cuervo et al. [5] further improved the performance by leveraging cross-attention mechanisms to integrate ASR-derived features with binaural information. Subsequently, Cuervo et al. [6] have shown that such cross-attention architectures significantly boost prediction performance, especially when combined with diverse ASR foundation models. However, most of the works adapted the pre-trained ASR by freezing the parameters of the model.

Inspired by these advances, our study proposes a two-stage framework: (1) We perform fine-tuning Whisper on the hearing aid output (*signal*) and the corresponding response from individuals with hearing impairment (*response*). By leveraging a pre-trained large-scale ASR model, we adapt its representations to align with the CPC domain. (2) Then, we utilize the extracted features from Whisper as inputs to either a Convolutional Neural Network (CNN) or a Transformer-based model for a regression task. While both architectures significantly outperform HASPI-based metrics, the Transformer-based model demonstrates distinguished performance due to its ability to capture global contextual information, such as phonemes affected by hearing loss.

2. Data Preprocessing and Model Architecture

2.1. ASR Fine-Tuning

Among the widely used pre-trained models in the CPC challenge [2], such as Whisper [4] and WavLM [7], we adopted Whisper as our ASR backbone. To align with the encoder-decoder architecture of Whisper, we preprocessed the CPC3 training dataset accordingly. In the original dataset, user responses were either empty or composed solely of '#' symbols when the audio was not recognized. However, for ASR training, it is necessary to provide sufficient tokens as decode input for predicting tasks. To address this, let the true label (*prompt*) contain N_t tokens, denoted as $\mathbf{T} = \{t_1, t_2, \ldots, t_{N_t}\}$, and the user answer (*response*) contain N_a tokens, denoted as

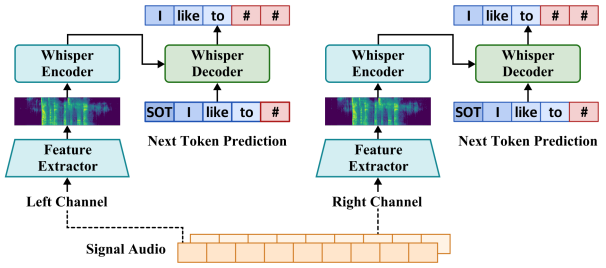

Figure 2: *The overview of the Whisper fine-tuning.*

$\mathbf{A} = \{a_1, a_2, \ldots, a_{N_a}\}$. To standardize the input length for ASR training, when $N_t > N_a$, we pad the response $\mathbf{A}$ with '#' tokens so that $N_t = N_a$, as shown in Figure 2.

Using preprocessed text data and waveform audio sampled at 16,000 Hz — the default setting of the Whisper's feature extractor — we trained the encoder and decoder for the next token prediction task. In line with prior work [5, 3], we split the audio into left and right channels and trained separate ASR models for each channel. The models were trained with a batch size of 16, a learning rate of 1×10^{-5}, and for 500 epochs.

2.2. CNN and Transformer Pre-Training

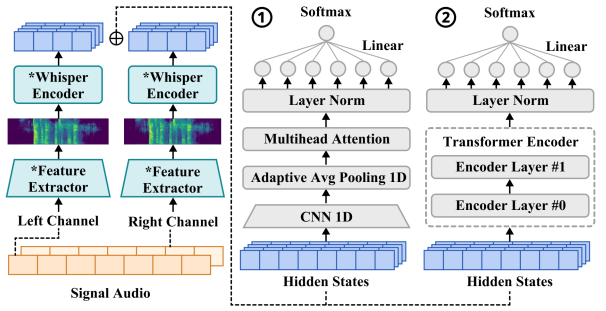

Figure 3: *The overview of the CNN/Transformer-based model.*

Using the frozen fine-tuned ASR model, we extracted its hidden states and employed them as input features for both CNN and Transformer-based models in the regression task. The overall architecture is illustrated in Figure 3, where ① represents the CNN-based model and ② displays the Transformer-based model.

The architecture of the CNN-based model is inspired by audio embedding models such as Wav2vec [8], HuBERT [9], and WavLM [7]. It combines a CNN layer followed by a self-attention mechanism to effectively capture local temporal features while modeling global relationships within the audio.

On the other hand, we also constructed a Transformer-based model adopting two layers of Transformer [10] encoder to more effectively capture global dependencies across long-form transcription. The experimental results for the two models are explained in the Section 3.3.

3. Experiments and Results

3.1. Experimental Setting

All experiments were conducted on a single NVIDIA RTX A6000 GPU (40GB VRAM), with 1.0T RAM, and a storage system comprising three 7TB SATA SSDs and two NVMe SSDs. The software environment included PyTorch 2.6.0, running on Ubuntu 20.04 with CUDA 11.4. The FLOPs, parameter counts, and estimated GPU memory requirements of the models were computed using the THOP profiler.

Table 1: *Comparison of parameters and FLOPs of CNN-based and Transformer-based models.*

Approach	Param. (M)	FLOPs (G)
CNN-based	175.18	263.22
Transformer-based	174.01	261.45

Table 1 summarizes the estimated resource requirements. While both architectures exhibit similar computational complexity, the CNN-based model shows slightly higher FLOPs and parameter counts than the Transformer-based model.

3.2. Dataset

Our experiments are conducted exclusively on the CPC3 dataset, without integrating any external data. During training, we use only the signal data from the training subset, along with the corresponding sentence data representing human responses.

3.3. Results

Table 2: *Comparison of HASPI, CNN-based and Transformer-based approaches in terms of correlation and RMSE.*

Approach	Correlation (↑)	RMSE (↓)
HASPI (Baseline)	0.72	28.00
CNN-based	0.81	23.37
Transformer-based	**0.82**	**22.95**

Table 2 presents the correlation and Root Mean Squared Error (RMSE) for the HASPI (baseline), CNN-based, and Transformer-based models on the *dev* dataset. Both the CNN and Transformer variants significantly outperform the baseline, demonstrating the efficacy of a non-intrusive, fine-tuned ASR approach. Notably, the Transformer-based model achieves the lowest RMSE score, which can be attributed to its ability to aggregate global spatial information via self-attention, in contrast to the CNN's reliance on local receptive fields [11].

This experimental result reflects the characteristics of the domain and dataset, as the frequencies that individuals with hearing loss struggle to perceive are not localized but rather globally distributed, which makes the Transformer architecture more suitable for capturing such patterns. The Transformer's self-attention mechanism effectively modeled individual differences in hearing loss through attention weights.

4. Conclusions and Future Works

We developed and compared a domain-specific fine-tuned ASR model, augmenting its architecture with the DNNs, such as CNN or transformer layers, for enhancing speech intelligibility prediction. To the best of our knowledge, this is the first attempt to apply such an approach in CPC. Experimental results demonstrate that the proposed method significantly outperforms the HASPI baseline, highlighting the effectiveness of our proposed method. However, the current model does not incorporate audiogram information, which represents individualized hearing loss profiles. As part of future work, we plan to investigate the impact of incorporating phoneme-level weighting informed by audiogram data to further enhance prediction performance.

5. Acknowledgements

This work was supported by Ralph W. and Grace M. Showalter Research Trust (No. 41001449), Clifford B. Kinley Trust, and Health of the Forces by Purdue University.

6. References

[1] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[2] J. Barker, M. A. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 551–11 555.

[3] R. Mogridge, G. Close, R. Sutherland, S. Goetze, and A. Ragni, "Pre-trained intermediate asr features and human memory simulation for non-intrusive speech intelligibility prediction in the clarity prediction challenge 2," *evaluation*, vol. 1, no. 6, pp. 6–21, 2023.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[5] S. Cuervo and R. Marxer, "Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction," *Proc. ISCA Clarity-2023*, 2023.

[6] ——, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1421–1425.

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] 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*, vol. 29, pp. 3451–3460, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in neural information processing systems*, vol. 34, pp. 12 116–12 128, 2021.