Abstract

This work introduces a novel non-intrusive speech intelligibility prediction model developed for the Clarity Prediction Challenge 3 (CPC3) dataset. The proposed system leverages multi-level representations extracted from three complementary sources: OpenAI's Whisper ASR model, Microsoft's WavLM self-supervised learning model, and a Convolutional Routing Transformer Attention (CRTA) mechanism. These representations are fused within a multi-branch architecture to predict intelligibility scores as perceived by hearing-impaired individuals, without access to a clean reference signal. The system was trained on CPC3 data and evaluated using listener-derived correctness scores. It achieved a Pearson correlation of 0.78 and an RMSE of 26.28 on development set. The model generalizes well to unseen listeners and hearing aid configurations, making it suitable for clinical applications.

1. Introduction

Hearing loss is a rapidly growing public health concern, currently affecting 466 million people worldwide and projected to impact over 630 million by 2030 due to global aging [1]. Older adults are particularly susceptible to hearing degradation due to physiological changes in the cochlea and auditory pathways [2, 3]. As the condition progresses, speech intelligibility (SI) especially in noisy environments becomes significantly impaired, leading to social, emotional, and cognitive decline [4]. To address this, hearing aid (HA) technologies have advanced, yet robust evaluation methods for their effectiveness are still lacking [5, 6]. Traditional human listening tests are costly and labor-intensive [7, 8], while existing objective metrics like HASPI [9] depend on access to clean reference signals, which limits real-world applicability [10].

Recent research has focused on non-intrusive models that estimate intelligibility from HA-processed audio without requiring reference signals [11]. The use of deep learning, particularly self-supervised and ASR-based models, has shown promise [12]. For example, HASA-Net and HASA-Net+ estimate SI by integrating acoustic signals with listener profiles [13]. Other approaches incorporate pre-trained speech foundation models (SFMs), as in Cuervo and Marxer [14] and Mogridge et al. [15] or combine intelligibility and quality estimation from HA outputs [16]. The CPC3 dataset provides a robust platform to test such systems, simulating realistic HA scenarios with diverse acoustic environments and listener data.

2. Methodology

Our proposed system is built around a multi-branch neural architecture designed to capture low-level acoustic, mid-level phonetic, and high-level linguistic information relevant to human perception. It consists of three parallel branches. The first is inspired by OpenAI's Whisper ASR model, and processes 40-bin log-Mel spectrograms using convolutional layers followed by Transformer encoder blocks. Adapter modules are included to fine-tune the ASR features for intelligibility prediction while minimizing overfitting. The second branch uses raw audio waveforms as input to a pre-trained WavLM Base model [17], which extracts robust, generalizable features learned from large-scale unlabeled speech data. Adapter layers again help align these representations to the downstream task. The third branch incorporates a CRTA module with sparse attention over Mel-spectrograms, selectively enhancing frames that contribute most to intelligibility. This attention output is processed through a convolutional feedforward network.

After feature extraction, outputs from all three branches are concatenated and passed through a bi-directional LSTM for temporal modeling, followed by two multi-head attention blocks that refine feature interactions. The final layers include fully connected blocks with GELU activations and a sigmoid output to map predictions between 0 and 1. Adaptive average pooling ensures consistency across variable-length input. During inference, the better-ear strategy is used: the ear (left or right channel) with the higher predicted intelligibility score is selected, reflecting the well-known auditory better-ear effect [18].

3. Experimental Setup

The system was trained using the CPC3 dataset, which contains 15,520 stereo audio samples sampled at 32 kHz. The training set included 12,410 samples derived from CEC1 and CEC2, while the validation set consisted of 3,110 samples from the acoustically diverse CEC3. Each audio sample is associated with a structured listener response, including transcription accuracy, which serves as the ground truth target. All preprocessing and model evaluation were done on stereo .wav files. During evaluation, intelligibility scores are predicted separately for left and right channels, and the highest score is selected per sample.

The model was trained for 200 epochs using L1 loss (mean absolute error), a batch size of 1, and the Adam optimizer. This setup was chosen to handle variability in signal content and duration. Training was performed on a single NVIDIA A100 GPU (40 GB) and completed in approximately 18 hours. No external data beyond CPC3 was used, and only officially released versions of the Whisper and WavLM Base models were employed, fully aligning with CPC3's non-intrusive track requirements.

4. Results

Our system achieved a Pearson correlation of 0.78 and an RMSE of 26.28 on the CPC3 development set, outperforming the CPC3-provided HASPI baseline model (correlation 0.72, RMSE 28.7). HASPI computes intelligibility using auditory modeling and clean reference signals and applies logistic regression to produce percentage correctness predictions. While this intrusive method offers high performance in controlled settings, it is less suited for clinical or embedded applications where reference signals are unavailable.

To further analyze model robustness, we examined specific validation samples with mismatched subjective and HASPI predictions. For instance, CEC1_E005_S08837_L0219 had a HASPI score of 0.648 but a subjective correctness score of 0, indicating a clear perceptual mismatch. Conversely, CEC2_E008_S08625_L0201 had a HASPI of just 0.036 yet was rated at 71.4% by listeners. These discrepancies highlight the limitations of intrusive metrics in capturing perceptual intelligibility, particularly in the presence of HA processing artifacts.

5. Discussion

The superior performance of our non-intrusive model emphasizes the value of multi-stream feature fusion, particularly when combining ASR, SSSR, and sparse attention mechanisms. Our architecture captures both local and global speech dynamics and is capable of adapting to diverse noise conditions and HA processing styles. The ability to generalize to unseen listeners and unseen HA algorithms is critical, given the increasing personalization in hearing healthcare. The disjoint validation setup—where listeners and HA systems are not seen during training, ensured our model's generalization capacity aligned with CPC3 expectations.

6. Conclusion

We present a non-intrusive intelligibility prediction model that combines pre-trained ASR (Whisper), self-supervised (WavLM), and attention-based (CRTA) representations. Our system meets all CPC3 challenge rules and significantly outperforms the HASPI baseline. With no need for clean references and demonstrated generalization to unseen listeners and processing pipelines, our method is well-suited for clinical deployment and HA-integrated intelligibility estimation. This work contributes a scalable, perceptually aligned, and technically robust approach to speech intelligibility prediction.

References

[1] World Health Organization, "Addressing the Rising Prevalence of Hearing Loss," 2018, ISBN: 9789241550260.

[2] J. Rennies, S. Goetze, and J.-E. Appell, "Personalized Acoustic Interfaces for Human-Computer Interaction," in Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications, M. Ziefle and C.R¨ocker, Eds., chapter 8,pp. 180–207. IGI Global, 2011.

[3] World Health Organisation, "Ageing and Health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, Accesssed: 2023-07-26.

[4] Lin FR, Metter EJ, O'Brien RJ, Resnick SM, Zonderman AB, Ferrucci L. Hearing loss and incident dementia. Archives of neurology. 2011 Feb 14;68(2):214-20.

[5] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 18–30, 2015.

[6] S. Goetze, F. Xiong, J. Rennies, T. Rohdenburg, and J. Appell, "Hands-Free Telecommunication for Elderly Persons Suffering from Hearing Deficiencies," in IEEE Int. Conf. on E-Health Networking, Application and Services (Healthcom'10), 2010.

[7] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and Limitations of Existing Tools," IEEE Signal Processing Magazine, vol. 32, no. 2, pp. 114–124, 2015.

[8] A. Warzybok, I. Kodrasi, J. Jungmann, E. Habets, T. Gerk- mann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective Speech Quality and Speech Intelligibility Evaluation of Single-Channel Dereverberation Algorithms," in Int. Workshop on Acoustic Signal Enhancement (IWAENC 2014), France, Sep. 2014.

[9] J. M. Kates and K. H. Arehart, "The Hearing-aid Speech Perception Index (HASPI) Version 2," Speech Communication, vol. 131, pp. 35–46, 2021.

[10] Y. Feng, and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," Biomedical Signal Processing and Control, 71, pp.103204, 2022.

[11] C.O. Mawalim, B.A. Titalim, S. Okada, and M. Unoki, "Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss," Applied Acoustics, 214, pp.109663, 2023.

[12] S.W. Fu, Y. Tsao, H.T. Hwang, and H.M. Wang, "Quality-Net: An end- to-end non-intrusive speech quality assessment model based on BLSTM," arXiv preprint arXiv:1808.05344, 2018.

[13] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Enhancing Two-Stage Finetuning for Speech Emotion Recognition Using Adapters," In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 11316-11320, 2024.

[14] Chiang HT, Fu SW, Wang HM, Tsao Y, Hansen JH. Multi-objective non-intrusive hearing-aid speech assessment model. The Journal of the Acoustical Society of America. 2024 Nov 1;156(5):3574-87.

[15] R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, A. Ragni, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate ASR features and human memory models," In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 306-310, 2024.

[16] Ashkanichenarlogh V, Folkeard P, Parsa V. Towards Clinically Feasible Nonintrusive Quality and Intelligibility Indices for Hearing Aids. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2025 Apr 6 (pp. 1-5). IEEE.

[17] Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, Li J, Kanda N, Yoshioka T, Xiao X, Wu J. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing. 2022 Jul 4;16(6):1505-18.

[18] I. Gibbs, Bobby E., J. G. W. Bernstein, D. S. Brungart, and M. J. Goupell, "Effects of better-ear glimpsing, binaural unmasking, and spectral resolution on spatial release from masking in cochlear-implant users," The Journal of the Acoustical Society of America, vol. 152, no. 2, pp. 1230–1246, 08 2022.