Non-Intrusive Speech Intelligibility Prediction for Hearing-Impaired Listeners Using Whisper

Matthew Daly

Johns Hopkins University

mdaly14@jhu.edu

Abstract

This technical report presents our submission for the 2nd Clarity Prediction Challenge. We propose a deep learning approach to performing non-intrusive speech intelligibility prediction for hearing-impaired listeners by fine-tuning the Whisper foundation model.

Index Terms: speech intelligibility, non-intrusive, hearing loss, foundation model, Whisper, deep learning

1. Introduction

Speech intelligibility measures how well a listener can understand the speech in an audio signal. This is a useful metric to evaluate speech enhancement systems, whose goal is to improve the intelligibility of speech signals. Oftentimes, measuring the intelligibility of speech involves listener tests, where test subjects listen to the signals and report what they hear. Unfortunately, this is a time-consuming process that can inhibit the development of speech enhancement algorithms, thus motivating the development of objective metrics which can measure intelligibility without requiring human listener tests.

Objective speech intelligibility metrics can be intrusive or non-intrusive. Intrusive metrics, such as the short-time objective intelligibility (STOI) metric [1], measure the intelligibility of a noisy signal by comparing it to a clean reference signal. Non-intrusive metrics do not use a reference signal and only take the noisy signal as input.

The goal of the 2nd Clarity Prediction Challenge (CPC2) is to develop systems for objectively predicting speech intelligibility of signals for hearing-impaired listeners. Listener-specific objective measurement of intelligibility can benefit the development of speech enhancement for hearing aids. CPC2 provides a training data set of speech signals labeled with a hearing-impaired listener's reported intelligibility for the signal. The training data also includes the listeners' audiograms, which characterize the listener's specific hearing loss.

Our proposed system for CPC2 leverages deep learning, which has been very successful in applications in the audio domain, especially speech processing. The Transformer neural network architecture, introduced in [2], has proven to be powerful in learning from audio signal data to perform speech related tasks such as automatic speech recognition (ASR). Foundation models that use the Transformer architecture have become popular for speech processing applications. These foundation models have been trained on extremely large quantities of audio data, and they have been shown to effectively fine-tune on smaller domain-specific data sets.

For our CPC2 submission, we propose a deep learning system that fine-tunes the Whisper ASR foundation model [3] to predict speech intelligibility. Our proposed system is non-intrusive.

2. Foundation Models

Foundation models are deep learning models that have been pre-trained on large amounts of data in order to learn generalized representations that can be applied to a variety of downstream tasks. Foundation models often utilize self-supervised or semi-supervised learning. In the audio domain, popular foundation models include wav2vec 2.0 [4], HuBERT [5], WavLM [6], and Whisper [3]. The useful features that these models have learned from large-scale training have been applied to tasks such as speaker verification [7][8], speaker recognition [9], speech emotion recognition [10][11], and Alzheimer's disease detection [12].

The Whisper model, which we use for our speech intelligibility prediction system, uses the Transformer architecture. It was trained on 680,000 hours of multilingual audio data using supervised multitask training. This large-scale multitask training on varied data allow the model to effectively generalize to new data sets and tasks.

3. Method

We propose a deep learning approach to non-intrusive intelligibility prediction that leverages the Whisper foundation model by fine-tuning it on the CPC2 data set. We implement our network architecture and training recipes using the SpeechBrain toolkit [13].

3.1. Network Architecture

We use the "base" pre-trained Whisper model, which has 74 million parameters. Since Whisper uses an encoder-decoder Transformer architecture, we select only the pre-trained encoder module to extract feature embeddings from the input audio. The Whisper encoder accepts input in the form of 80-channel log-magnitude Mel spectrograms computed from audio sampled at 16 kHz.

We perform temporal pooling on the output of the encoder by taking the mean of the features across the time dimension. After this pooling, the output is passed to a fully-connected neural network. A sigmoid activation function is applied to the output of the fully-connected network to obtain an intelligibility score between 0 and 1.

We incorporate the listener audiogram information into the network by adding a convolutional neural network module that takes in the audiogram and the Mel spectrogram and outputs an encoding that is added back to the original input Mel spectrogram before being passed to the Whisper encoder. The goal of this module is to try to learn listener-specific hearing loss characteristics.

3.2. Training

We train our model using AdamW optimization with learning rate annealing. The initial learning rate for the parameters in the Whisper encoder is smaller than that of the parameters in the added layers. We perform data augmentation while training by using the SpecAugment method [14], which is popular for training deep learning models for audio tasks.

We train our model and run experiments in Google Colaboratory using A100 GPUs.

4. Results

This section will contain the performance results of our model on CPC2 data.

5. Conclusions

This section will contain our conclusions drawn from our experiments and results.

6. References

[1] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[2] 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.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[4] 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.

[5] 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.

[6] 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," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[7] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," 2021.

[8] J. Peng, T. Stafylakis, R. Gu, O. Plchot, L. Mošner, L. Burget, and J. Černocký, "Parameter-efficient transfer learning of pre-trained transformer models for speaker verification using adapters," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7967–7971.

[10] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.

[11] L.-W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[12] J. Li, K. Song, J. Li, B. Zheng, D. Li, X. Wu, X. Liu, and H. Meng, "Leveraging pretrained representations with task-related keywords for alzheimer's disease detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680