Integrating Linguistic and Acoustic Cues for Machine Learning-Based Speech Intelligibility Prediction in Hearing Impairment

Candy Olivia Mawalim[1], *Xiajie Zhou*[1], *Huy Quoc Nguyen*[1], *Masashi Unoki*[1]

[1]Graduate School of Advanced Science and Technology, JAIST, Japan

candylim@jaist.ac.jp, xiajie@jaist.ac.jp, hqnguyen@jaist.ac.jp, unoki@jaist.ac.jp

Abstract

Speech intelligibility prediction for individuals with hearing loss is paramount for advancing hearing aid technology. Leveraging recent breakthroughs in ASR foundation models, particularly Whisper, we fine-tuned a Whisper model for speech intelligibility prediction. Our approach incorporates data augmentation using impulse responses from diverse everyday environments. This study investigates the effective integration of linguistic and acoustic cues to enhance the prediction of fine-tune ASR models, aiming to compensate for both hearing loss and information loss during signal downsampling. Our goal is to improve speech intelligibility prediction, especially in noisy conditions. Experiments demonstrate that integrating these cues is beneficial. Furthermore, employing a weighted average ensemble model, which balances predictions from left and right audio channels and considers both stable and unstable linguistic and acoustic cues, significantly improved prediction performance, reducing the RMSE by approximately 2 and enhancing the Pearson correlation coefficient (ρ) by around 0.05.

Index Terms: speech intelligibility, hearing loss, ASR

1. Introduction

Accurate speech intelligibility prediction for individuals with hearing loss is vital for developing improved hearing aids, as current models frequently lack robustness, especially in noisy environments. Recent advances in automatic speech recognition (ASR), particularly with foundation models such as Whisper [1], Wav2vec2 [2], and WavLM [3], offer promising results in any speech processing tasks. Wav2vec2 excels in self-supervised representation learning in low-resource and transfer learning scenarios, WavLM provides robust speech representations and enhanced paralinguistic feature extraction (e.g., tone, emotion), and Whisper demonstrates strong performance, particularly in content-driven tasks and noisy conditions [4].

However, these foundation models are typically trained on 16-kHz signals and optimized for clean audio. Applying them to hearing loss scenarios, which often involve higher-fidelity signals, can lead to a significant loss of crucial acoustic and linguistic information during downsampling. This missing data, especially at higher frequencies, can severely compromise the accuracy of intelligibility predictions for impaired listeners.

In this study, we investigate how perception of linguistic and acoustic cues in hearing loss can be effectively integrated into improve these speech foundation models. Our goal is to compensate for both hearing loss and information lost during signal downsampling, thereby improving the prediction of speech intelligibility in noisy environments.

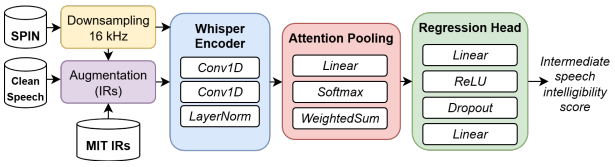

Figure 1: *Architecture of the Fine-tuned Whisper Model for Speech Intelligibility Prediction. SPIN is speech in noise.*

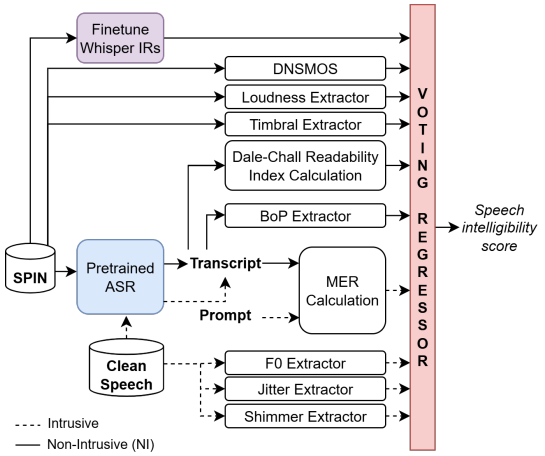

Figure 2: *Integration Model with Linguistic and Acoustic Cues.*

2. Model Architecture

2.1. Fine-tuned ASR Model

Figure 1 illustrates the architecture of our fine-tuned Whisper model. We specifically selected the `small.en` variant, which is pre-trained on an English-only dataset, owing to its favorable performance-to-parameter ratio. Our proposed model's architecture consists of three primary components: the Whisper Encoder, an Attention Pooling Layer, and a Regression Head.

To enhance the robustness of our model across diverse acoustic environments, we incorporated an augmentation layer during training. This layer generates augmented data by convolving the audio with randomly selected impulse responses obtained from The MIT McDermott dataset [5], which comprises 271 impulse responses recorded in various everyday locations. We generated 20% augmented data from our existing training set, aiming to expose the model to a wider range of environmental conditions beyond those inherently present in the original dataset. For training our regression model, we employed a hybrid loss function that combined MSE and Pearson correlation loss (referred to as MSE-PearsonLoss).

Table 1: *Overall Speech Intelligibility Prediction Results: Comparison of Models, Linguistic and Acoustic Cues Integration (Int.), and Ensemble (Ens.) Methods.*

NI	Model	Int.	Ens.	Linguistic Cues*			Acoustic Cues**						Validation		Development		Evaluation	
				MER	BoP	DC	F_0	Jit	Shim	Loud	Timb	MOS	RMSE	ρ	RMSE	ρ	RMSE	ρ
No	be-HASPI	No	No	-	-	-	-	-	–	-	-	-	29.33	0.6623	28.00	0.7200	29.47	0.6973
	HASPI	Yes	Yes	v	v	v	v	v	v	v	v	v	**22.28**	**0.8221**	**24.71**	**0.7973**	**25.97**	**0.7756**
Yes	FT-Whisper	Yes	No	-	v	v	-	-	-	-	-	v	24.83	0.7787	31.04	0.6991	32.98	0.6498
	FT-Whisper	Yes	Yes	-	v	v	-	-	-	v	v	v	24.82	0.7787	30.16	0.7092	32.74	0.6469
No	FT-Whisper	Yes	No	v	v	v	v	v	v	-	-	v	22.34	0.8164	26.24	0.7711	28.30	0.7341
	FT-Whisper	Yes	Yes	v	v	v	v	v	v	v	v	v	**21.86**	**0.8293**	**24.81**	**0.7972**	**26.14**	**0.7733**
No	STM-CNN-SE	No	No	-	-	-	-	-	-	-	-	-	24.71	0.7754	24.86	0.7859	28.13	0.7442
	STM-CNN-SE	Yes	No	v	v	v	v	v	v	-	-	v	22.51	0.8136	23.43	0.8125	26.00	0.7770
	STM-CNN-SE	Yes	Yes	v	v	v	v	v	v	v	v	v	**22.47**	**0.8193**	**23.15**	**0.8179**	**25.60**	**0.7835**
No	STM-CNN-ECA	No	No	-	-	-	-	-	-	-	-	-	24.16	0.7872	24.46	0.7944	27.88	0.7469
	STM-CNN-ECA	Yes	No	v	v	v	v	v	v	-	-	v	22.40	0.8203	24.10	0.8004	26.35	0.7717
	STM-CNN-ECA	Yes	Yes	v	v	v	v	v	v	v	v	v	**22.32**	**0.8222**	**23.47**	**0.8123**	**26.02**	**0.7769**

* MER: match error rate (Whisper (`medium.en`), WavLM-large, wav2vec2-large), BoP: Bag-of-Phonemes, DC: Dale-Chall Readability Index

** Jit: Jitter, Shim: Shimmer, Loud: Loudness (obtained using ITU-R BS.1770-4 standard, via pyloudnorm), Timb: Timbral (Hardness, Brightness, Sharpness), MOS: Quality scores obtained by DNSMOS.

(Intrusive models) **E020a**: STM-CNN-SE, **E020b**: STM-CNN-ECA, **E020c**: FT-Whisper, all with integration and ensemble models.

(Non-intrusive models) **E020c-NI**: FT-Whisper with integration and ensemble models.

2.2. Linguistic and Acoustic Cues Integration

The crucial impact of hearing loss on speech perception is typically visualized within an audiogram's "speech banana" area. Our method directly integrates various linguistic and acoustic cues that are particularly relevant to this perceptual impact. Figure 2 shows the integration of the prediction obtained by the finetune Whisper model with linguistic and acoustic cues.

For linguistic cues, we first leverage the match error rate (MER) derived from pre-trained ASR models as an initial indicator of intelligibility. Beyond this, we utilized "bag-of-phonemes" approach to to quantify the occurrence and distribution of difficult phonemes existing in the text. We also integrated a readability index to represent the overall ease of understanding the spoken text. Specifically, we utilized the Dale-Chall readability formula [6].

For acoustic cues, we incorporate fundamental frequency (F_0), jitter, shimmer, loudness, and timbral features. Studies consistently show that low-frequency cues, particularly those related to F_0, significantly improve speech recognition in noisy environments [7]. While jitter and shimmer are less directly linked to speech-in-noise perception, they are often combined with other cues and are strongly associated with voice pathology [8]. Regarding loudness, simply amplifying sound to restore normal loudness can paradoxically lead to excessively loud sensations at certain frequencies or sound levels, thereby complicating SPIN understanding for hearing aid users [9]. Furthermore, research indicates that individuals with hearing loss exhibit diminished timbre discrimination in both quiet and noisy conditions, a consequence of broadened auditory filters and reduced sensitivity to fine spectral and temporal details [10]. Noise is another significant factor influencing intelligibility. Since SNR values are unavailable, we utilized DNSMOS (Deep Noise Suppression Mean Opinion Score) [11] as a proxy to estimate speech quality in noisy conditions.

3. Experiments

3.1. Experimental Setup

First, a correlation analysis was conducted between the cue and correctness to choose the most influential linguistic and acoustic cues with a threshold of 0.1. The voting regressor used for our integration model is composed of a GradientBoosting Regressor, RandomForest, and Linear Regression, all utilizing their default sklearn implementations.

Recognizing statistical discrepancies in cue averages and standard deviations across datasets, we developed a weighted ensemble model. This ensemble combines predictions from four sub-models: (1) left and right (LR) channels with stable cues (excluding loudness and timbral), (2) LR with all cues, (3) the mean of LR channels with stable cues, and (4) the mean of LR channels with all cues. Higher weights were assigned to sub-models utilizing stable acoustic cues and incorporating both left and right channel predictions.

3.2. Results

Table 1 presents the speech intelligibility prediction results obtained using the Clarity Prediction Challenge (CPC3) dataset. Our fine-tuned Whisper model, referred to as FT-Whisper, was compared against the HASPI [12], demonstrating the adaptability of our linguistic and acoustic cue integration approach across different prediction methodologies. We also applied this integration strategy and an ensemble model to the STM-CNN-based methods, as detailed in [13]. For CPC3 evaluation, we submitted specific configurations: the integrated and ensembled Convolutional Neural Network with Spectral-Temporal Modulation input and a Squeeze-and-Excitation (SE) block (STM-CNN-SE) as **E020a**; with an Efficient Channel Attention (ECA) block (STM-CNN-ECA) as **E020b**; and FT-Whisper as **E020c**. An improved non-intrusive version of **E020c**, designated **E020c-NI**, was also included. Due to space limitations, Table 1 presents only representative combinations of these evaluations. Our results indicate that integrating and ensembling models generally improved predictions, with RMSE reduced by about 2 and ρ increasing by roughly 0.05. These gains were even more significant when applied to the HASPI model, where RMSE decreased by over 4 and ρ improved by 0.15.

4. Conclusion

This paper proposed integrated models for robust speech intelligibility prediction in individuals with hearing loss. By integrating linguistic and acoustic cues with a weighted ensemble strategy, we significantly enhanced predictive accuracy of a fine-tuned Whisper model and other comparative models. Our experiments consistently demonstrated improved RMSE by about 2 and correlation by about 0.05, highlighting the potential of these combined techniques to advance speech intelligibility prediction for accurate hearing aid development.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers (20KK0233, 21H03463, 25H01139 and 25K21245).

6. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. of ICML 2023, 23-29 July, Honolulu, Hawaii, USA*, vol. 202. PMLR, 2023, pp. 28 492–28 518.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NeurIPS 2020, December 6-12, virtual*, 2020.

[3] 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 J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] H. Yang, J. Zhao, G. Haffari, and E. Shareghi, "Investigating pre-trained audio encoders in the low-resource condition," in *Proc. of Interspeech 2023, Dublin, Ireland*. ISCA, 2023, pp. 1498–1502.

[5] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[6] E. Dale and J. Chall, "A formula for predicting readability," *Educational Research Bulletin, 27*, pp. 1–20, 1948.

[7] J. Carroll, S. Tiaden, and F.-G. Zeng, "Fundamental frequency is critical to speech perception in noise in combined acoustic and electric hearing," *J Acoust Soc Am*, vol. 130, no. 4, pp. 2054–2062, Oct 2011.

[8] J. Kreiman and B. R. Gerratt, "Perception of aperiodicity in pathological voice," *J Acoust Soc Am*, vol. 117, no. 4 Pt 1, pp. 2201–2211, Apr 2005.

[9] D. Oetting, V. Hohmann, J.-E. Appell, B. Kollmeier, and S. D. Ewert, "Restoring perceived loudness for listeners with hearing loss," *Ear Hear*, vol. 39, no. 4, pp. 664–678, Jul/Aug 2018.

[10] S. Emiroglu and B. Kollmeier, "Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions," *Brain Res*, vol. 1220, pp. 199–207, Jul 2008.

[11] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. of ICASSP 2022, Virtual and Singapore*. IEEE, 2022, pp. 886–890.

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

[13] X. Zhou, C. O. Mawalim, H. Q. Nguyen, and M. Unoki, "Lightweight Speech Intelligibility Prediction with Spectro-Temporal Modulation for Hearing-Impaired Listeners," *The 3rd Clarity Prediction Challenge Technical Report*, 2025.