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

Candy Olivia Mawalim, Xiajie Zhou, Huy Quoc Nguyen, Masashi Unoki
Japan Advanced Institute of Science and Technology, Japan

BACKGROUND

Hearing Aids (HA)

Assist hearing-impaired listener

Speech Intelligibility Assessment

Difficulty to assess speech intelligibility especially in noisy environment

RECENT TREND

SPEECH FOUNDATION MODEL
(e.g., Whisper [1], Wav2vec2 [2], and WavLM [3])

Downstream Tasks
(ASR, audio classification, etc.)

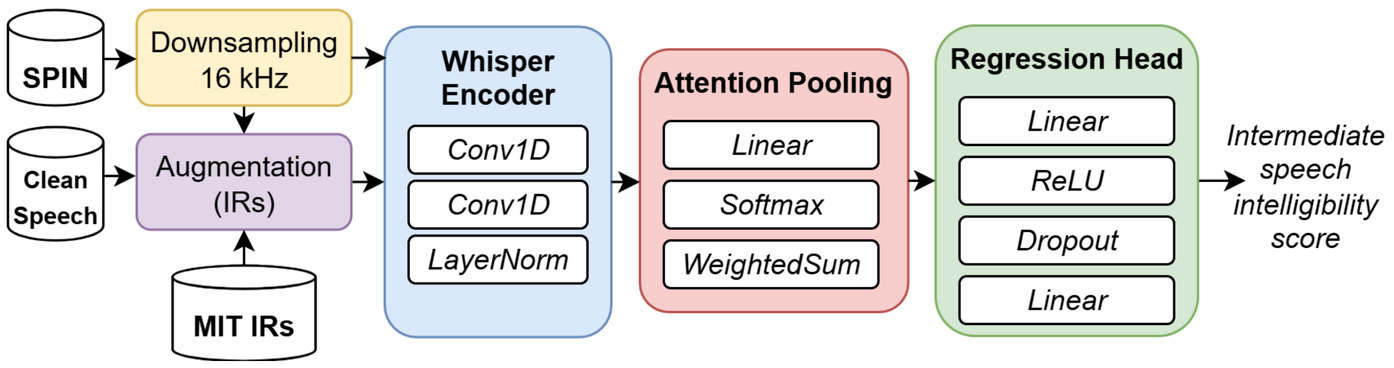
RESEARCH QUESTIONS

1. How effectively do foundational models predict speech intelligibility in hearing aids?

2. How can we integrate linguistic and acoustic cues to improve speech intelligibility prediction?

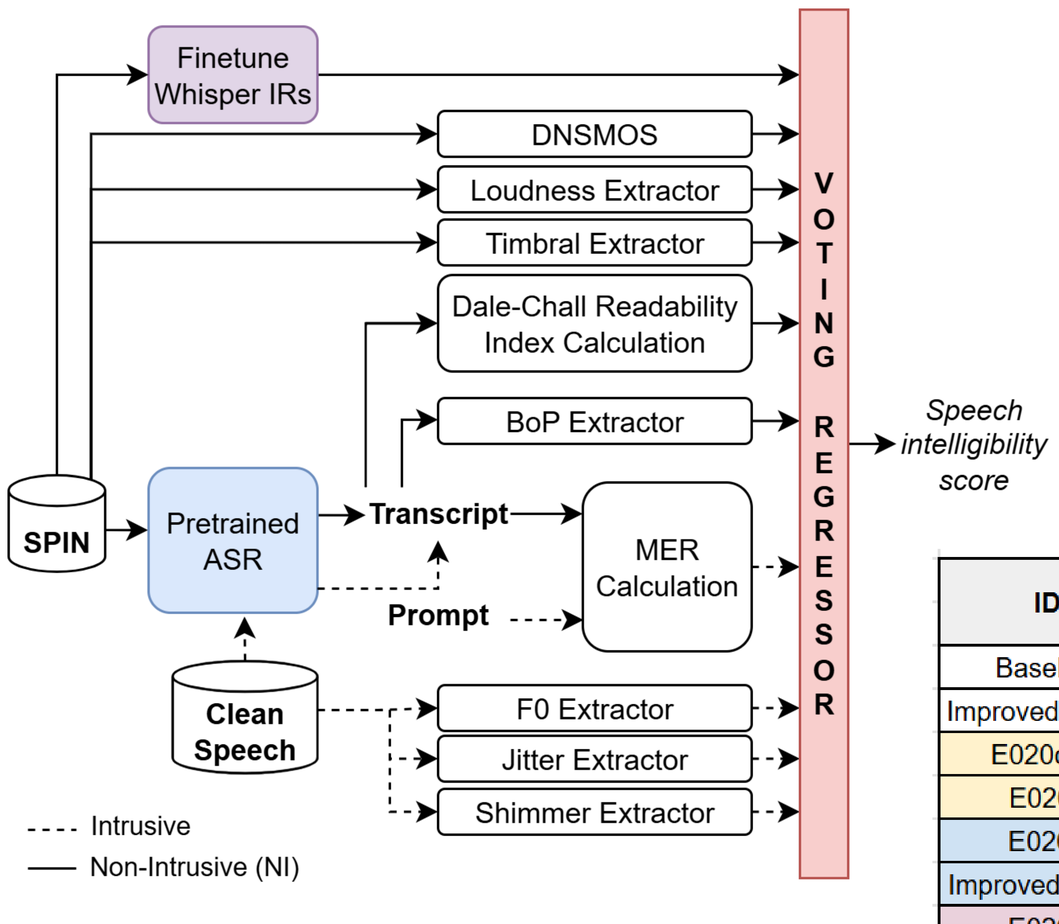
EXPERIMENT 1: FINE-TUNING SPEECH FOUNDATION MODEL

- ❑ Fine-tuning on individual foundation model shows that Whisper gives the most promising results.
- ❑ We selected the `small.en` variant for its strong performance and efficient design.



- ❑ To improve robustness, an augmentation layer by convolving the audio with impulse responses (IRs) obtained from The MIT McDermott dataset [4] (271 IRs recorded in various everyday locations).
- ❑ **Limitation:** Original model was trained on 16 kHz dataset → a significant loss at higher frequencies.

EXPERIMENT 2: LINGUISTIC AND ACOUSTIC CUES INTEGRATION



SPIN: Speech-in-Noise, **IRs**: Impulse Responses,
DNSMOS: Deep Noise Suppression Mean Opinion Score,
BoP: Bag-of-Phonemes, **MER**: Match Error Rate

- ❑ **Linguistic Cues:** MER, BoP, Dale-Chall readability index
- ❑ **Acoustic Cues:** F0, Jitter, Shimmer, Loudness, Timbral (Hardness, Brightness, Sharpness), DNSMOS
- ❑ **Voting Regressor:** GradientBoosting, RandomForest, and Linear Regressors
- ❑ **Final Ensemble Predictions:** weighted average of
  1) Left and right (LR) channels with stable cues
  2) LR channels with all cues
  3) Mean of LR channels with stable cues
  4) Mean of LR channels with all cues

ID	NI	Model	Validation		Development		Evaluation	
			RMSE	ρ	RMSE	ρ	RMSE	ρ
Baseline	No	be-HASPI	29.33	0.6623	28.00	0.7200	29.47	0.6973
Improved HASPI		HASPI	**22.28**	**0.8221**	**24.71**	**0.7973**	**25.97**	**0.7756**
E020c-NI	Yes	FT-Whisper	24.82	0.7787	30.16	0.7092	32.74	0.6469
E020c	No	FT-Whisper	**21.86**	**0.8293**	**24.81**	**0.7972**	**26.14**	**0.7733**
E020a	No	STM-CNN-SE	24.71	0.7754	24.86	0.7859	28.13	0.7442
Improved E020a		STM-CNN-SE	**22.47**	**0.8193**	**23.15**	**0.8179**	**25.60**	**0.7835**
E020b	No	STM-CNN-ECA	24.16	0.7872	24.46	0.7944	27.88	0.7469
Improved E020b		STM-CNN-ECA	**22.32**	**0.8222**	**23.47**	**0.8123**	**26.02**	**0.7769**

[5] E020a and E020b

Scan for more information about our paper

Conclusion

1) Our research proposed integrated models for robust speech intelligibility prediction in individuals with hearing loss.
2) Fine-tuning fundamental models (e.g., Whisper) gave limited performance in speech intelligibility prediction.
3) By integrating linguistic and acoustic cues, we significantly enhanced predictive accuracy of a finetuned Whisper model and other speech intelligibility models.
4) Our experiments consistently demonstrated improved RMSE by about 2 and correlation by about 0.05.

References
[1] A. Radford, et al., "Robust speech recognition via large-scale weak supervision," in Proc. of ICML 2023.
[2] A. Baevski, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. of NeurIPS 2020.
[3] S. Chen, et al., "WavLM: Largescale self-supervised pre-training for full stack speech processing," IEEE J. Sel. Top. Signal Process. 2022.
[4] https://mcdermottlab.mit.edu/Reverb/IR_Survey.html
[5] X. Zhou, et al., "Lightweight Speech Intelligibility Prediction with Spectro-Temporal Modulation for Hearing-Impaired Listeners," Clarity Workshop 2025.

Email: {candylim, xiajie, hqnguyen, unoki}@jaist.ac.jp