Lightweight Speech Intelligibility Prediction with Spectro-Temporal Modulation for Hearing-Impaired Listeners

Xiajie Zhou, Candy Olivia Mawalim, Huy Quoc Nguyen, Masashi Unoki

Graduate School of Advanced Science and Technology, JAIST, Japan

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

Abstract

Hearing loss leads to reduced frequency resolution and impaired temporal resolution, making it difficult for listeners to distinguish similar sounds and perceive speech dynamics in noise. To capture these perceptual degradations, we employ spectro-temporal modulation (STM) analysis as the core feature representation. This study proposes a speech intelligibility prediction framework that uses STM representations as input to lightweight convolutional neural network (CNN) models. We design two models: STM-CNN-SE (E020a), which incorporates squeeze-and-excitation (SE) block, and STM-CNN-ECA (E020b), which uses efficient channel attention (ECA) block and richer input features. Compared to the HASPI, experiments on the CPC3 development dataset show that E020a and E020b reduce root-mean-square error (RMSE) by 11.2% and 12.6%, respectively. These results demonstrate the effectiveness of STM-based CNN architectures for speech intelligibility prediction under hearing loss conditions.

Index Terms: hearing loss, spectro-temporal modulation, speech intelligibility

1. Introduction

Hearing-impaired listeners experience degraded frequency resolution and reduced temporal resolution, both of which are critical for understanding speech in noisy environments [1]. These perceptual deficits are attributed to impaired cochlear processing, leading to broader auditory filters and slower temporal integration. To capture these deficits, we employ spectro-temporal modulation (STM) analysis, which explicitly encodes how acoustic energy varies across both frequency and time. STM representations can capture the changes in modulation patterns caused by hearing loss and serve as a reliable basis for prediction model.

Building on this representation, we develop prediction models that predict speech intelligibility from STM representations. Two key insights guide our approach.

(1) We simulate hearing loss prior to feature extraction, allowing the resulting STM representations to reflect listener-specific degradations in frequency and temporal resolution.

(2) A lightweight convolutional neural network (CNN) with channel attention can effectively learn mappings from STM representations to predicted scores.

Based on these principles, we propose two models: STM-CNN-SE as E020a, which integrates a squeeze-and-excitation (SE) block, and STM-CNN-ECA as E020b, which replaces SE with an efficient channel attention (ECA) block. Both models take STM representations as input and are trained to predict normalized speech intelligibility scores.

2. Method

2.1. Audio Preprocessing and STM Representations

We convert stereo clean speech and SPIN into fixed-size STM representations, as illustrated in Fig. 1. This preprocessing pipeline follows our previous work [2]. First, to retain only voiced content, we apply energy-based voice activity detection (VAD) on the clean channel and extract the corresponding segments from both clean speech and SPIN. We then simulate listener-specific hearing loss using the MSBG model [3, 4], with parameters configured according to the severity level (Mild, moderate, or moderately severe). Next, we extract the temporal envelope using the Hilbert transform and apply a low-pass filter whose cutoff frequency depends on the severity level [5, 6]. Finally, the signals are passed through two-dimensional Gabor filters to obtain a four-dimensional STM representation [7]:

$$\mathrm{STM} \in \mathbb{R}^{N_S \times N_T \times C \times T},$$

where $N_S = 5$ and $N_T = 10$ are the numbers of spectral and temporal modulation channels, respectively; C is the number of ERB channels derived from the Gammatone filterbanks after hearing loss simulation; and T is the number of time frames.

2.2. Speech Intelligibility Models

We propose two CNN-based regression models—STM-CNN-SE (E020a) and STM-CNN-ECA (E020b)—to predict speech intelligibility scores from STM representations. As shown in Fig. 2, both models share a common architectural backbone.

Each model takes as input a multi-channel STM representations built from clean speech and SPIN. The input is first normalized across time to reduce input bias and improve training stability. The core of the model consists of three convolutional blocks, each containing two Conv2D layers with batch normalization and ReLU activation [8]. These blocks extract hierarchical STM features, while intermediate max-pooling reduces spatial resolution and retains dominant responses. An attention module follows the final ConvBlock to recalibrate channel-wise importance. Finally, features are flattened and fed into a fully connected head to output a normalized score in the range $[0, 1]$.

Compared to E020a, E020b introduces three modifications.

(1) *Input features:* E020a uses the STM difference map, while E020b additionally includes the log ratio mask, yielding a richer representation of signal degradations

(2) *Training strategy:* E020b applies data augmentation using Spectrogram Augmentation (SpecAugment) and Mixup regularization to enhance robustness to modulation loss and reduce overfitting in noisy conditions.

(3) *Attention mechanism:* E020a employs a SE block to model global inter-channel dependencies [9], while E020b uses an ECA block to capture fine-grained local interactions with lower complexity [10]. Both blocks are illustrated in Fig. 3.

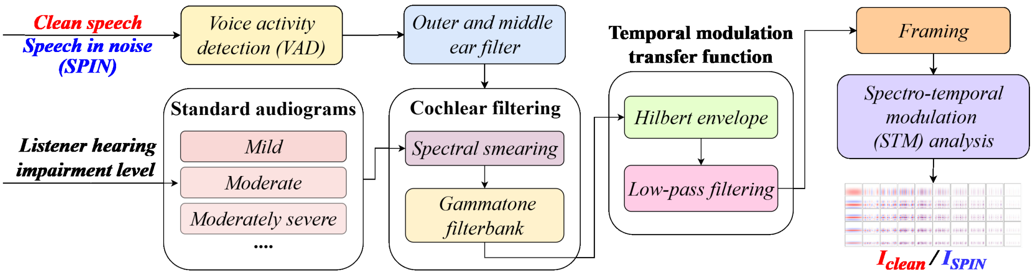
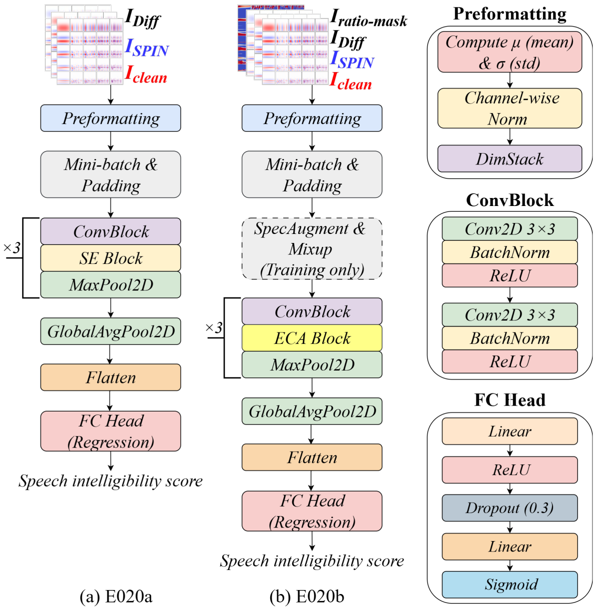


Figure 1: *Overview of the audio preprocessing and STM representation pipeline. The input includes stereo clean and SPIN, along with listener hearing impairment severity levels, and the output is a fixed-size STM representation.*



(a) E020a (b) E020b

Figure 2: *Model architectures of E020a (left) and E020b (right).*

3. Experiments

3.1. Experimental Setup

To simulate listener-specific hearing loss, we apply the MSBG model and adopt the standard audiograms defined in the HASPI.

Both models are trained using the AdamW optimizer with weight decay of 1×10^{-5} and an initial learning rate of 3×10^{-4}. For E020b, the learning rate is dynamically scheduled using OneCycleLR with a peak of 1×10^{-3}. The loss function is a weighted combination of mean squared error and Pearson correlation, with a regularization factor $\lambda = 0.1$, to jointly promote prediction accuracy and monotonicity.

We train for up to 30 epochs with a batch size of 16 (E020a) and 8 (E020b), and select the best model based on validation root-mean-square error (RMSE). All models are trained on a single NVIDIA RTX 4090 GPU with 24 GB memory, taking 3–5 minutes per epoch. All experiments are reproducible.

3.2. Results and Analysis

Table 1 summarizes the performance of different models on the validation and development sets. Among single-branch models, E020a and E020b both outperform the HASPI, with E020b showing consistent improvements on the averaged scores. Specifically, E020a and E020b reduce RMSE on the development set by 11.2% and 12.6%, respectively.

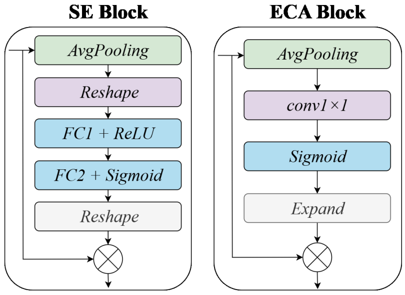


Figure 3: *Attention blocks used in E020a (SE block) and E020b (ECA block).*

Table 1: *Performance comparison of models on the validation and development sets.*

Model	**Ear**	**Validation Set**		**Development Set**	
		RMSE ↓	Pearson ↑	RMSE ↓	Pearson ↑
Baseline (HASPI)	–	–	–	28.00	0.7200
E020a	Left Right Avg	25.40 25.50 24.71	0.7605 0.7610 0.7754	– – 24.86	– – 0.7859
Improved E020a	–	**22.47**	**0.8193**	**23.15**	**0.8179**
E020b	Left Right Avg	24.52 24.91 **24.16**	0.7811 0.7721 **0.7872**	– – **24.46**	– – **0.7944**
Improved E020b	–	**22.32**	**0.8222**	**23.47**	**0.8123**

Notably, the *Improved* E020a system achieves the best overall performance, with a validation RMSE of 22.47 and a development RMSE of 23.15—corresponding to a 17.3% improvement over the baseline. This system integrates and ensembles STM-CNN-SE (E020a) with linguistic and acoustic cues, combined via ensemble methods.

4. Conclusion

This paper proposed two CNN-based models for speech intelligibility prediction using spectro-temporal modulation (STM) representations. By jointly capturing frequency and temporal resolution and incorporating attention mechanisms, the STM-CNN-SE and STM-CNN-ECA models achieve substantial improvements in prediction accuracy. On the CPC3 development set, E020a and E020b reduce RMSE by 11.2% and 12.6%, respectively, compared to HASPI. These results highlight the effectiveness of STM-based modeling for predicting speech intelligibility in realistic listening conditions.

5. Acknowledgements

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

6. References

[1] B. C. J. Moore, “Frequency selectivity and temporal resolution in normal and hearing-impaired listeners,” *British Journal of Audiology*, vol. 19, no. 3, pp. 189–201, 1985.

[2] X. Zhou, C. O. Mawalim, and M. Unoki, “Modeling multi-level hearing loss for speech intelligibility prediction,” *arXiv preprint arXiv:2507.22599*, 2025.

[3] Y. Nejime and B. C. J. Moore, “Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise,” *J. Acoust. Soc. Am.*, vol. 102, no. 1, pp. 603–615, 1997.

[4] G. F. Smoorenburg, “Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram,” *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 421–437, 1992.

[5] J. G. Desloge, C. M. Reed, L. D. Braida, Z. D. Perez, and L. A. Delhorne, “Temporal modulation transfer functions for listeners with real and simulated hearing loss,” *J. Acoust. Soc. Am.*, vol. 129, no. 6, pp. 3884–3896, 2011.

[6] S. P. Bacon and R. M. Gleitman, “Modulation detection in subjects with relatively flat hearing losses,” *J. Speech Lang. Hear. Res.*, vol. 35, no. 3, pp. 642–653, 1992.

[7] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, “Speech intelligibility prediction using spectro-temporal modulation analysis,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 210–225, 2020.

[8] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[9] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[10] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “Eca-net: Efficient channel attention for deep convolutional neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 534–11 542.