Intrusive Intelligibility Prediction with ASR Encoders

Hanlin Yu[1], *Haoshuai Zhou*[2], *Linkai Li*[2,3], *Boxuan Cao*[2], *Changgeng Mo*[2], *Shan Xiang Wang*[3]

[1]Department of Electrical Engineering, University of British Columbia, Canada
[2]Orka Labs Inc., China
[3]Department of Electrical Engineering, Stanford University, United States

hyu29@student.ubc.ca, {haoshuai,dicky.mo,boxuan.cao}@hiorka.com,
{linkaili,sxwang}@stanford.edu

Abstract

Sentence-level speech intelligibility predictors have plateaued with RMSEs above 20 on CPC2 [1]. Without a clean reference, it is hard to indicate *when* and *what* words or phonemes fail for hearing-impaired listeners. We hypothesize that modeling the *deviation* between a noisy utterance and its clean counterpart can reveal the spans that break comprehension and improve overall prediction. We participate in the Clarity Prediction Challenge 3 (CPC3) and ask whether incorporating clean *reference* signals can improve sentence-level intelligibility prediction for hearing-impaired listeners.

Index Terms:intrusive speech intelligibility assessment, multi-scale CNN,SFM layer selection

1. Introduction

Progress in this task hinges on two decisions. First, we must choose a feature extractor that captures the cues most predictive of intelligibility—whether mid-depth SSL/SFM representations or multi-scale CNN time–frequency features [2]. Second, we need a predictor that can exploit these cues effectively, translating them into accurate sentence-level scores by modeling how deviations in the noisy signal (relative to the clean reference, when available) lead hearing-impaired listeners to miss specific words or phonemes.

Automatic speech recognition (ASR) encoders offer rich, linguistically informed representations for speech intelligibility prediction, since they effectively transform acoustic waveforms into text-level features. Prior work suggests that intermediate hidden layers—rather than the input or final layers—often yield the most discriminative features for downstream tasks [3]. Motivated by these findings, we first seek to identify the best intermediate layers of an ASR encoder for intelligibility assessment and then investigate how incorporating clean reference signals can further improve prediction accuracy.

We select top-ranked encoder(s) and ensemble their mid-depth representations with a multi-scale CNN (MSCNN) front end, yielding a compact and accurate intelligibility predictor.

2. Model

As shown in Fig.1, each model operates on a log-Mel spectrogram of shape $[B, T, N_{\text{mels}}]$ with $N_{\text{mels}} = 128$. We first apply a three-branch 1-D temporal convolutional front end:

- **Branch F:** $\text{Conv1D}\big(N_{\text{mels}} \to H,\ k = 3,\ \text{dilation} = 1,\ \text{padding} = 1\big)$
- **Branch M:** $\text{Conv1D}\big(N_{\text{mels}} \to H,\ k = 5,\ \text{dilation} = 2,\ \text{padding} = 4\big)$
- **Branch C:** $\text{Conv1D}\big(N_{\text{mels}} \to H,\ k = 9,\ \text{dilation} = 4,\ \text{padding} = 16\big)$

The three H-channel outputs are concatenated and projected to a D-dimensional embedding, yielding per-frame features of size $[B, T, D]$.

We then extract the most informative intermediate layers from our pretrained encoders and process each layer with a `CrossAttentionBlock`, where every block includes a residual connection and LayerNorm. The SSL features are subsequently down-sampled by a factor of 8, yielding a tensor of shape $[B,\ T/8,\ D]$.

Next, the left-ear, right-ear, and reference sequences all go through a *Temporal Transformer* to model long-range dependencies. After this, each ear performs a frame-level cross-attention against the reference: so that any distortion or missing cues in the noisy signal at time t can be corrected by the clean reference. We then apply mean pooling over time to obtain a single D-dimensional vector per layer per stream.

Finally, we stack the per-layer vectors for each stream (plus a severity embedding) and feed them into a *Layer Transformer*. This is followed by layer-level reference alignment via cross-attention block, and layer-level ear fusion via cross-attention block. which integrates multi-scale information and merges the complementary strengths of the two ears. Each ear's final token is passed through a lightweight MLP with a sigmoid activation to produce a score; the higher of the two ear scores is selected as the overall intelligibility prediction.

3. Experimental setup

We reuse the same predictor as in Fig.1 and vary only which encoder layers feed it. For each model, we sweep contiguous, fixed-size layer windows across depth (e.g., 0–3, 4–7, . . .), train on the fixed train split, evaluate validation RMSE, and pick the window with the lowest RMSE as the "best layers." The selected window per model is used in all subsequent experiments.
Layers used in fusion. We take *layers 10–17 (inclusive)* from both SFMs (Canary-1B-Flash and Parakeet-TDT-0.6B-V2).
Optimization. AdamW (lr 3×10^{-5}, weight decay 10^{-2}), batch size 8, 9 epochs.
Five folds and evaluation. In each fold, the validation set contains exactly two listeners per severity class (*Mild*, *Moderate*, *Moderately severe*)—six validation listeners in total—while the remaining listeners form the training split. We create 5 listener-disjoint folds (fixed validation listeners per fold). For each fold we train to the best validation RMSE and keep that checkpoint. At test time, we predict the **dev** and **evaluation**

Correspondence: linkaili@stanford.edu, sxwang@stanford.edu

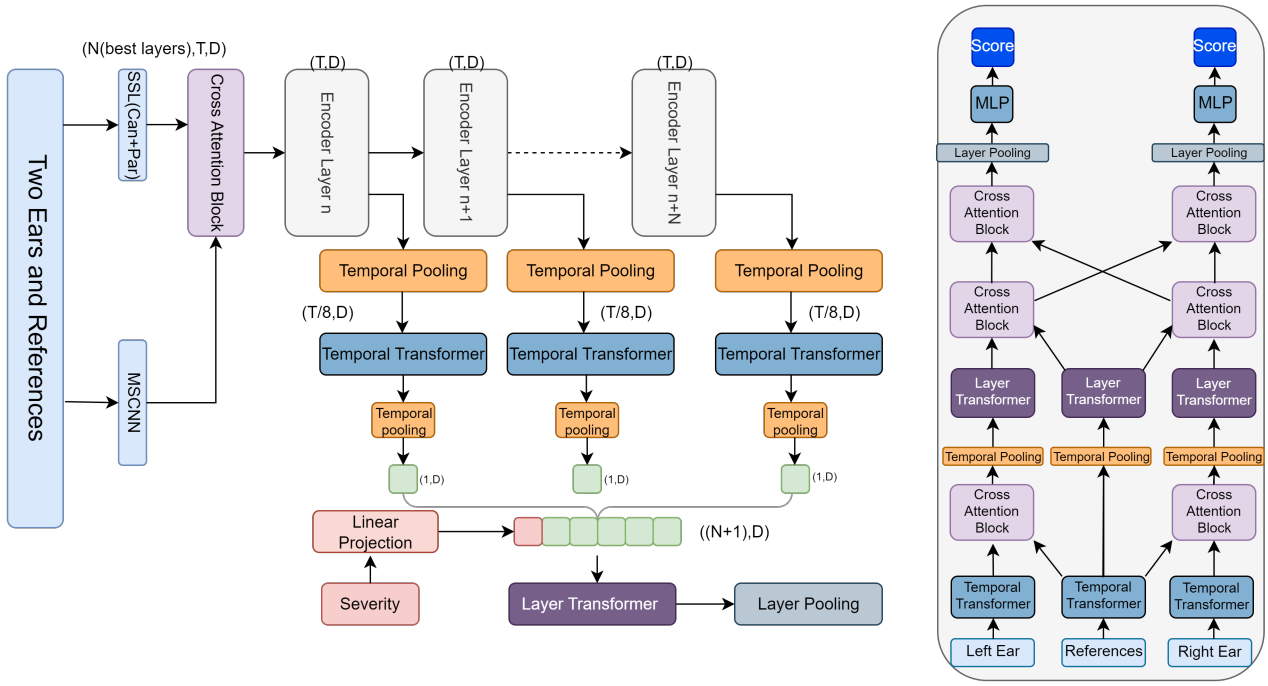

Figure 1: ***Model overview.*** *(a) Shared per-stream encoder for left, right, and reference signals. (b) Fusion stage: each ear attends to the reference (L→Ref, R→Ref) and to the other ear (L↔R). Panels (a) + (b) form the full model.*

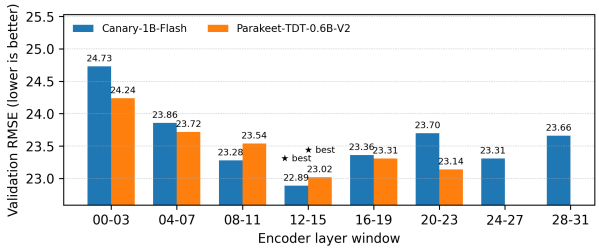

Figure 2: *Validation RMSE by encoder layer window.*

Table 1: *Ablation on the dev set (RMSE ↓).*

Configuration	RMSE
Single ear (L + R averaged), no cross-attn	25.00
+ Dual-ear with cross-ear attention	23.60
+ Reference cross-attention	23.40
+ Swap to Canary-1B-Flash & add Parakeet	22.40
+ Add MSCNN features	**22.36**

sets with all five checkpoints and *average the five fold scores* to obtain the final submission.

Compute and runtime. All experiments were run on Google Colab with a single NVIDIA L4 GPU (10 GB GPU RAM) and 30 GB host RAM.

4. Results and Analysis

Starting from a single–ear setup (left/right averaged) we obtain an RMSE of 25.00. Modeling the two ears explicitly with cross–ear attention lowers RMSE to 23.60. Adding reference–guided cross–attention yields a small further gain to 23.40. Replacing *Canary-1B* with the *Canary-1B-Flash+Parakeet-TDT-0.6B-V2* ensemble improves RMSE to 22.40, and adding MSCNN features brings the final score to **22.36**.

Replacing the temporal/layer transformers with `Conexibimamba` (a Conformer variant where multi-head self-attention is replaced by Mamba-style blocks) did not help [4], likely because the SSL encoders already provide strong sequence representations. The reference signal provides a modest dev-set gain (23.60→23.40); future work will evaluate whether it improves robustness on unseen data. Overall, the full model reduces RMSE from the baseline 28.00 (Table **??**) to 22.36 (∼20% relative).

5. Conclusions

We presented a reference–aware intelligibility predictor that selects the most informative mid–depth layers from two speech foundation models (Canary-1B-Flash and Parakeet-TDT-0.6B-V2; layers 10–17), fuses them with a multi-scale CNN front end, and employs cross-reference and cross-ear attention at both temporal and layer levels with a severity token. Experiment showed no gains from replacing transformers with Mamba-style blocks or from score-level fusion.

With the clean reference signal available, we will further exploit TextGrid alignments to move beyond sentence-level scores and attempt true word-level intelligibility prediction.

6. References

[1] "Cpc2_e011 report on non-intrusive speech intelligibility prediction," https://claritychallenge.org/clarity2023-workshop/papers/CPC2_E011_report.pdf, 2023, clarity 2023 Workshop.

[2] W. Ren, Y.-C. Lin, W.-C. Huang, R. E. Zezario, S.-W. Fu, S.-F. Huang, E. Cooper, H. Wu, H.-Y. Wei, H.-M. Wang, H.-y. Lee, and Y. Tsao, "Highratemos: Sampling-rate aware modeling for speech quality assessment," *arXiv preprint arXiv:2506.21951*, 2025. [Online]. Available: https://arxiv.org/abs/2506.21951

[3] H. Zhou, B. Cao, C. Mo, L. Li, and S. X. Wang, "Unveiling the best practices for applying speech foundation models to speech intelligibility prediction for hearing-impaired people," *arXiv preprint arXiv:2505.08215*, 2025. [Online]. Available: https://arxiv.org/abs/2505.08215

[4] X. Zhang, Q. Zhang, H. Liu, T. Xiao, X. Qian, B. Ahmed, E. Ambikairajah, H. Li, and J. Epps, "Mamba in speech: Towards an alternative to self-attention," *arXiv preprint arXiv:2405.12609*, 2025. [Online]. Available: https://arxiv.org/abs/2405.12609