A Chorus of Whispers: Modeling Speech Intelligibility via Heterogeneous Whisper Decomposition

Longbin Jin[1], *Donghun Min*[1], *Eun Yi Kim*[1]

[1]AI&CV Lab., Department of Computer Science and Engineering, Konkuk University, South Korea

{jinlongbin, mindonghun, eykim}@konkuk.ac.kr

Abstract

This paper introduces *a Chorus of Whispers*, a simple yet effective method for modeling speech intelligibility in hearing-impaired listeners, developed for the third Clarity Prediction Challenge (CPC3). Our approach simulates a spectrum of perceptual abilities by creating a "chorus" of heterogeneous Whisper models, ranging from the powerful large version to the lightweight tiny variant. By decomposing the audio signal through the diverse outputs of this chorus, we extract robust representations that reflect listening difficulty. These representations are then fed into an ensemble of word- and sentence-level models to predict the final intelligibility score. The proposed method demonstrates strong generalization to unseen conditions, achieving a competitive RMSE of 23.62 on the CPC3 development set.

Index Terms: chorus of whisper, clarity prediction challenge

1. Introduction

Speech intelligibility prediction plays a critical role in the development and evaluation of hearing aid technologies, particularly for individuals with hearing impairments. The third Clarity Prediction Challenge (CPC3) addresses this problem by providing a benchmark task for predicting the intelligibility of speech processed by hearing aids in complex acoustic environments. In previous challenges, most approaches relied on a single model or a fixed metric, which limited their ability to capture the diversity of human hearing. Such models often overfit to specific speech patterns, acoustic conditions, or listener types, and struggle to generalize across a broad range of hearing profiles.

In nature, many phenomena are composed of smaller components working together. For example, a chord is not a single note but a combination of multiple notes played simultaneously. Similarly, any signal can be decomposed into a mixture of frequency components. Extending this analogy to human hearing, particularly for individuals with impaired auditory systems, perception can be viewed as the interplay of multiple perceptual channels, each with varying sensitivities and strengths. In other words, listening is like a chorus where some channels hear clearly while others struggle, and the overall perception emerges from this complex blend.

Motivated by this insight, we construct a *Chorus of Whisper* models, spanning from the powerful large to the lightweight tiny (Figure 1). Each model processes the same audio input but "listens" differently, capturing distinct cues related to intelligibility. The diversity in model behavior mirrors the variability in human auditory perception, yielding a comprehensive representation of the challenges involved in understanding speech across different listener profiles. By passing these heterogeneously Whisper-decomposed features to word- and sentence-level models, our method offers generalizable predictions of speech clarity, even under unseen acoustic conditions.

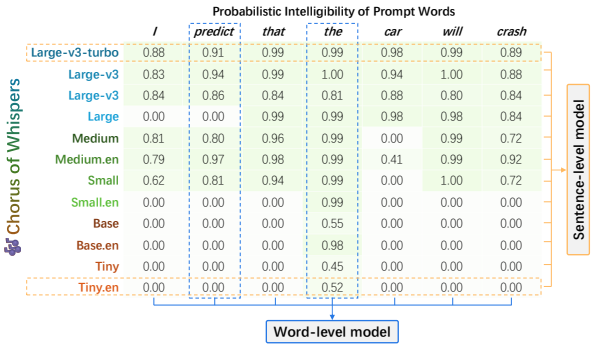

Figure 1: ***Overview of the Chorus of Whispers.*** *Multiple frozen Whisper models of varying sizes process the same input and produce distinct listening probabilities for each prompt word. From the* ***model perspective****, larger models have better intelligibility, while smaller models may miss content. From the* ***word perspective****, some words (e.g., "the") are detected by all models, while others (e.g., "predict") are recognized only by the stronger ones. These heterogeneous responses form the basis of our Whisper decomposition, which feeds word- and sentence-level models to predict speech intelligibility.*

2. Chorus of Whispers

2.1. Extracting Heterogeneous Representations

At the core of our approach is the idea of combining the diverse perceptual abilities of multiple Automatic Speech Recognition (ASR) models. To this end, we adopt 12 pretrained Whisper model variants [1] (large-v3-turbo, large-v3, large-v2, large, medium, medium.en, small, small.en, base, base.en, tiny, and tiny.en) to construct a fine-grained measure of intelligibility. The ".en" models are trained specifically for English, while the others are multilingual. All models are kept frozen and used solely for inference, predicting token sequences along with their associated probabilities.

To compute the intelligibility score for each prompt word, we align the predicted tokens with the corresponding words. If a word corresponds to a single token, its probability is used directly; for multi-token words, we average the token probabilities. This process is repeated across all 12 models, yielding a $n \times 12$ matrix of intelligibility scores for each input with n prompt words. As illustrated in Figure 1, the Whisper-derived features capture diverse perspectives on speech intelligibility. Models of different sizes yield distinct predictions due to differences in capacity, while individual words may vary in recognition difficulty depending on their acoustic and linguistic properties. These heterogeneous responses form a multi-dimensional representation that serves as input to our downstream models.

2.2. Modeling Speech Intelligibility

We employ two models to estimate intelligibility: a sentence-level model and a word-level model. The sentence-level model can capture differences in overall recognition performance across Whisper variants, while the word-level model further accounts for variations at the individual word level.

Sentence-level Model. Each sentence contains a varying number of words, and thus produces an $n \times 12$ matrix of intelligibility scores. To obtain a unified representation of sentence-level intelligibility, we average the matrix along the word dimension, resulting in a 12-dimensional vector. This vector is passed through a multi-layer perceptron (MLP) regression model with a layer configuration of 12–64–32–16–1. We apply Layer Normalization, the SiLU activation function, and a dropout rate of 0.01. Due to differences in intelligibility distributions across severity levels (Mild, Moderate, Moderately Severe), we train separate sentence-level models for each group. This severity-specific training helps each model capture distinct intelligibility patterns, improving regression accuracy and perceptual modeling.

Word-level Model. Each word is represented as a fixed 12-dimensional vector and used as an input token to a two-layer transformer encoder with a hidden size of 64 and 4 attention heads. A 64-dimensional learnable class token, following the Vision Transformer (ViT) [2] design, is prepended to the sequence. Unlike the sentence-level MLP models, we train a single transformer model across all severity levels by prepending a severity-specific embedding token to the input. The output corresponding to the class token is passed through a linear head to predict the final sentence-level intelligibility score.

3. Experiment results

3.1. Implementation Details

For training and validation data separation, we employ a prompt-aware strategy to ensure that the same prompt sentence does not appear in both sets. This prevents data leakage and improves the model's robustness to unseen prompts. The data is split into an 8:2 ratio. Both models are trained using a combined loss function consisting of root mean square error (RMSE) and a weighted Pearson correlation loss (weight = 0.1). We use a learning rate of 0.001 and train for 1000 epochs with full-batch training. The best model checkpoint is selected based on the lowest RMSE on the validation set.

3.2. Performance of Speech Intelligibility

Table 1 presents the performance of the proposed Chorus of Whispers. For sentence-level modeling, we train separate models for each severity group. Among them, the model trained on listeners with mild hearing loss achieves the best performance, as their responses are generally more stable and predictable. In contrast, the models for moderate and moderately severe groups perform worse due to higher variability in individual hearing conditions. Some individuals may better perceive high-frequency sounds, while others may be more sensitive to low frequencies. The overall sentence-level model achieves RMSE scores of 23.69 on the validation set and 23.88 on the development set.

The word-level model benefits from a larger parameter capacity and a transformer-based architecture, enabling it to capture fine-grained acoustic patterns more effectively. To further improve performance, we ensemble the sentence- and word-level predictions by averaging their outputs. This strategy leverages the complementary strengths of both models, achieving the lowest RMSE on both validation and development sets. As a result, the ensemble model was selected for final test set submission. Notably, our method significantly outperforms the HASPI baseline [3], underscoring the effectiveness of our approach in modeling speech intelligibility.

Table 1: *RMSE of Speech Intelligibility Prediction Models on the Validation and Development Sets.*

Model	Validation	Development
HASPI (Baseline) [3]	-	28.00
Sentence (Mild)	19.93	-
Sentence (Moderate)	25.78	-
Sentence (Moderately Severe)	24.94	-
Sentence (All)	23.69	23.88
Word (All)	23.36	23.84
Ensemble (All)	**22.95**	**23.62**

4. Discussions and Conclusions

We also experimented with alternative ASR models such as wav2vec2 and HuBERT, but they failed to recognize meaningful content under noisy conditions, making them unsuitable for this task. We further explored transcribing clean reference signals to extract additional information for intrusive systems. However, Whisper models transcribe clean speech with near-perfect accuracy, resulting in uniformly high intelligibility scores that failed to reflect meaningful variation. Since real listeners also do not have access to clean references, the Chorus of Whispers excludes them to better approximate real-world listening conditions.

While our method provides stable and accurate intelligibility predictions, it has limitations. It requires more computational resources than single-model approaches due to the use of multiple ASR systems and predictors. However, efficiency was not a key constraint in this challenge. The method also relies on prompt information, similar to how human raters assess correctness, and thus remains consistent with non-intrusive evaluation.

In summary, we proposed Chorus of Whispers, which simulates speech intelligibility by leveraging a diverse set of ASR systems with varying capabilities. By integrating sentence- and word-level predictors through ensembling, our approach captures a broad range of intelligibility cues and achieves strong performance. It significantly outperforms the HASPI baseline, offering a robust and practical solution for intelligibility assessment in noisy environments.

5. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

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