Non-Intrusive Speech Intelligibility Prediction Using Whisper ASR and Wavelet Scattering Embeddings for Hearing-Impaired Individuals

Rantu Buragohain[1], *Jejariya Ajaybhai*[1], *Aashish Kumar Singh*[1], *Karan Nathwani*[1], *Sunil Kumar Kopparapu*[2]

[1]Department of Electrical Engineering, Indian Institute of Technology Jammu, India
[2]TCS Research, Tata Consultancy Services Limited Mumbai, India

{2021ree1027, karan.nathwani}@iitjammu.ac.in, sunilkumar.kopparapu@tcs.com

Abstract

Hearing loss affects a significant population worldwide leading to an increase in usage of hearing aids. Ability to accurately predict intelligibility of speech, especially in noisy environments can go a long way in helping improve the performance of hearing aids. We present, as part of the 3rd Clarity Prediction Challenge (CPC3), a deep neural network framework which benefits from the contextual depth of Whisper-based embeddings and the resilience of Wavelet Scattering Transform (WST) embeddings to enable a robust speech intelligibility (SI) prediction. While the Whisper-based embeddings are the output of the final encoder (1024) and the final decoder (768) of a pre-trained encoder-decoder transformer trained on 680k hours of multilingual data, derived from the 80-channel log-Mel spectrogram of the input waveform, the second-order WST-based embeddings, with J=6 filterbanks and Q=8 wavelets per octave are extracted from the raw waveform. The WST-based embeddings provide deformation-stable time-frequency representations. We propose five systematically designed models: (Model #1) encode-only, leveraging embeddings from the final encoder layer of Whisper-medium; (Model #2) decode-only, utilizing the final decoder layer embeddings of Whisper-small; (Model #3) encode-decode, a fusion model that combines both encoder and decoder embeddings; (Model #4) hybrid, a model that uses encode-decode and WST-based embeddings; and (Model #5) ensemble, an average of (Model #1+Model #2+Model #3) with and without post-processing. Each embedding stream is independently processed using bidirectional long-short term memory (Bi-LSTM) layers and attention pooling, followed by fully connected (linear) layers to predict SI score. Our best performing ensemble with & without post processing, combining the outputs of first three models, achieves a root mean square error (RMSE) of 21.87 & 22.66 respectively, on the development set.

Index Terms: Speech recognition, intelligibility prediction, human impairment, Whisper embeddings, wavelet scattering transform (WST).

1. Introduction

Hearing loss is an emerging global public health concern. According to the World Hearing report, by 2050, around 2.5 billion individuals would experience hearing loss, with at least 700 million requiring rehabilitation [1, 2]. It predominantly impacts older persons and leads to communication difficulties, social isolation, and emotional distress, significantly impacting quality of life [3, 4]. Economically, hearing loss not only contributes to higher medical costs, especially for mental and cognitive health, but also results in premature withdrawal from the labor market, reduced income, and greater dependence on social services. Research shows that individuals with hearing loss experience a 52% higher risk of social isolation, a 47% greater probability of depression, and an unemployment rate that is double that of people with normal hearing [5]. As a result, hearing health is a significant concern in the medical domain and a pivotal element affecting social well-being, demanding urgent global initiatives.

Hearing aids, as wearable and easy-to-use devices, play a significant role in enhancing speech intelligibility (SI) for those with hearing impairment [6, 7]. Research comparing hearing aids of various designs has shown that both basic and advanced models can significantly improve speech understanding in everyday situations [8], indicating that even cost-effective devices offer considerable benefits. However, a significant gap in worldwide hearing aid services persists, especially in low-income nations, where merely 17% of those requiring hearing aids utilize them [1]. The gap arises not only from the high expense of modern hearing aids but also from a deficiency in understanding regarding hearing assessments and the public's inadequate awareness of the advantages of hearing aids, especially their contribution to improving SI. Traditionally, the evaluation of SI relies on subjective hearing tests, which are time-consuming, resource-demanding, and expensive. However, recent advances in machine learning has enabled exploration of both intrusive and non-intrusive approaches for objective prediction of SI. These approaches learn how the auditory system functions and extract key speech features to predict SI. For instance, non-intrusive SI machine learning models predicted SI by analyzing speech signals using convolutional neural networks (CNNs) [9, 10].

The Clarity Prediction Challenge 3 (CPC3) advances the work of earlier Clarity Enhancement Challenges, namely, CEC1, CEC2, and CEC3 by integrating a larger and more diverse dataset of listener data for model training and evaluation. It includes the processed audio signals produced by systems submitted by challenge participants, representing the output of diverse hearing aid algorithms, along with listener response data obtained during formal evaluation sessions. The objective of CPC3 is to predict SI scores for the hearing-impaired listeners based on the processed audio along with additional information, like the listener's hearing loss representation. CPC3 includes two tracks, namely (a) non-intrusive track, which allows the use of only the hearing aid processed noisy signal, and (b) intrusive track, which additionally permits use of clean reference audio. We propose, as part of the non-intrusive track of CPC3, a novel framework that allows for integration of two modules which extract complementary features from the waveform to predict SI. The first module leverages a pre-trained encoder-decoder openAI Whisper transformer model to extract high-level acoustic and linguistic embeddings from the final encoder (layer #24 of Whisper-medium) and final decoder (layer #12 of Whisper-

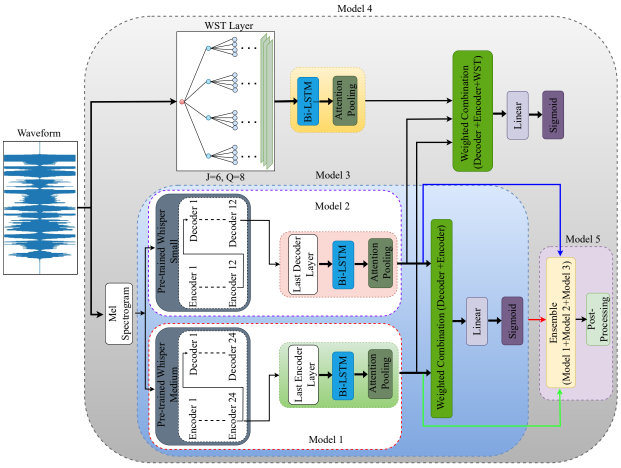

Figure 1: *Proposed model architectures.*

small) layers. These embeddings capture rich contextual information across multiple levels. These embeddings are further refined using a `Bi-LSTM` layer, followed by `attention pooling` to generate compact temporal representation of the waveform. The second module utilizes a WST layer directly on the input waveform, extracting first- and second-order coefficients. These coefficients yield low-level time-frequency (TF) translation-invariant features that are robust to signal deformations. The WST-based features are further processed using a `Bi-LSTM` layer and `attention pooling` to obtain a fixed-length embedding. Studies across different domains have reported the advantage of using WST in various applications, such as, Electroencephalogram classification [11], emotion recognition [12], unmanned aerial vehicle detection [13], infant cry scene detection and classification [14]. The embeddings obtained from both modules are then fused via a Weighted Combination Decoder, which learns to integrate the semantic richness of Whisper-based features with the robust spectral characteristics of the WST-based features. This fused representation is passed via a fully connected (linear) layers with a sigmoid activation to predict the SI score (see Figure 1). To systematically assess the contribution of each module, five progressively designed models are developed: (1) encode-only, (2) decode-only, (3) encode-decode fusion, (4) hybrid (encode-decode-WST integration), and (5) ensemble with and without post-processing. The proposed hybrid model efficiently addresses the limitations of the existing methods by integrating high-level contextual representations (Whisper-based embeddings) with noise-robust, translation-invariant spectral features (WST-based embeddings), improving resilience to background noise, capturing individual hearing profiles, and enabling robust generalization across real-world, acoustically challenging environments, which is critical for deployment in hearing aids. The rest of this paper is structured as follows: Section 2 describes the proposed systems in detail, while Section 3 outlines the experimental methodology. Section 4 presents the findings, whereas Section 5 concludes the study.

2. Proposed Systems

Figure 1 captures the proposed model designed to address the non-intrusive track of CPC3. The proposed system leverages the complementary strengths of (a) high-level contextual semantic embeddings derived from the Whisper model, and (b) robust, low-level TF representations obtained via the WST layer. Each module incrementally increases in complexity and representational capacity while sharing a common processing pipeline: extracted embeddings are passed through `Bi-LSTM` layers to capture temporal dependencies, followed by `attention pooling` to generate compact, informative embeddings before being combined to predict SI.

2.1. Model #1: Encode-Only

The Whisper-medium model [15] is a pretrained encoder-decoder transformer trained on 680k hours of multilingual data for tasks such as speech transcription and voice activity detection consisting of 24 encoder and 24 decoder layers [16]. The input audio signal $\hat{S}[n]$ is down-sampled to 16kHz and zero padded to make it 30 seconds long. Using a window of 25ms and a stride of 10ms, $\hat{S}[n]$ is transformed into an 80-channel log Mel spectrogram $\hat{S}$ before inputting to the Whisper model. The output of the final encoder layer (layer #24) which encodes phonetic and acoustic information, is a feature tensor of shape $1500 \times 1024 \times 1$, where 1024 representing the feature dimension per layer. The features are passed through a `Bi-LSTM` to capture temporal dynamics, subsequently followed by `attention pooling`. The aggregated embedding is then passed through a fully connected (linear) layer with a sigmoid to produce the final predicted SI (see Model #1, Fig 1).

2.2. Model #2: Decode-Only

Model #2 (see Figure 1) isolates the linguistic component of the waveform using Whisper-small (12 encoder and 12 decoder layers). For the same $\hat{S}$, embeddings from the final decoder (layer #12), capturing deep semantic and language-aware representations are used. These outputs, shaped $W \times 768 \times 1$ (where W is the number of predicted words), are processed via a `Bi-LSTM`, followed by `attention layer`, along the lines of Model #1. The resulting embedding is passed to a fully connected (linear) layer followed by sigmoid layer for SI prediction.

2.3. Model #3: Encode-Decode Weighted Embeddings

Model #3 (see Figure 1) integrated the strengths of acoustic and semantic cues. It jointly utilizes the final encoder layer (layer #24) from Whisper-medium and the final decoder layer (layer #12) from Whisper-small. Each stream is processed independently via its own `Bi-LSTM` and `attention pooling`, resulting in two compact embeddings. The embeddings are then fused through a learnable weighted combination, enabling adaptive feature balancing based on the input context. The unified representation is passed through a fully connected (linear) layer with sigmoid activation to enhance the SI prediction.

2.4. Model #4: Hybrid; Encode-Decode + WST Integration

The hybrid model integrates Model #3 the WST-based embeddings. As seen in Figure 1, the same waveform $\hat{S}[n]$ is processed through a WST layer, which provides robust, invariant, and stable TF representations crucial for SI prediction. The WST operates by applying a cascade of predefined complex wavelet filters $\phi_\lambda(t)$, and averaging via a low-pass filter $\psi(t)$. The wavelet filter $\phi(t)$ is a band pass filter with center frequency normalized to 1, and the filter bank is constructed as $\phi_\lambda(t) = \lambda\phi(\lambda t)$, where $\lambda = 2^{p/Q}$ for $p \in \mathbb{Z}$, and $Q = 8$ is the number of wavelets per octave. This setup yields a filter bank with bandwidth approximately $\frac{1}{Q}$, leading to band-pass filters centered at λ with a bandwidth of $\frac{\lambda}{Q}$ [17]. The zero-order scattering coefficient, given by $S_0\hat{S}[n] = \hat{S} * \psi(t)$, captures the global average but lacks discriminative information for prediction and is thus excluded. The first-order coefficients, computed as $S_1\hat{S}[n] = |\hat{S} * \phi_{\lambda_1}| * \psi(t)$, capture localized frequency content while preserving invariance to small temporal translations. To improve stability under time-warping and capture more complex structures, second-order coefficients are computed after the first wavelet transform and follows the Lipschitz deformation stability condition [17]. Higher-order coefficients ($m \geq 2$) are obtained as, $S_{m\hat{S}}(t, \lambda_1, ..., \lambda_m) = |||\hat{S} * \psi_{\lambda_1}| * ...|\psi_{\lambda_m}| * \zeta(t)$.

The final embeddings extracted from the three branches, namely, embed-only, decode-only and WST-based are subsequently fused via a Weighted Combination Decoder, which learns to balance and integrate the semantic richness of Whisper embeddings with the robust, low-level spectral embeddings derived from the WST layer. This fusion module creates a unified representation that captures both abstract linguistic structure and detailed acoustic patterns. The fused embedding is passed through a series of fully connected (linear) layers, concluding with a sigmoid activation function that outputs a scalar value representing the predicted SI score for the given input waveform $\hat{S}[n]$. This integrated hybrid architecture (Model #4, Figure 1) enables robust, generalizable, and objective SI assessment across diverse acoustic conditions, particularly beneficial for hearing-impaired listeners.

2.5. Model #5: Ensemble with Post-Processing

A post-processing step was applied to the final predictions generated by ensemble (Model #1+Model #2+Model #3), particularly for instances (or cases) of mild hearing loss, which were observed to be consistently underestimated by the ensemble. This approach was designed to enhance predictions post-model training without altering the model architecture. Post-processing was applied to the Model #5 predictions ($pred$) on 900 training data samples with established intelligibility scores across various hearing loss profiles, including 266 mild samples corresponding to mild hearing loss. Instead of using a single correction factor, we adopted a band-wise correction strategy by dividing the predicted SI scores into ten 10-point bands: (0-10), (10-20),...,(90-100). For each band, a separate correction factor α_i was optimized using a grid search over 101 values from 0.00 to 1.00 to minimize the RMSE. We used the following correction ($pred_\epsilon$), namely,

$$pred_\epsilon = min\left((1 + \alpha_i) * pred, 100\right) \qquad (1)$$

The optimized α values for each band: {0-10: 0.00, 10-20: 0.00, 20-30: 0.53, 30-40: 0.33, 40-50: 0.10, 50-60: 0.35, 60-70: 0.16, 70-80: 0.19, 80-90: 0.11, 90-100: 0.04}. This band-wise adjustment slightly scaled the predictions upward, capping them at 100 to maintain valid score bounds. Extension of (1) to moderate and moderately severe hearing loss cases yielded no performance gain, indicating that Model #5 was able to handle these classes well.

3. Experimental Methodology

The CPC3 data consists of (a) a training set, (b) a development set, and (c) an evaluation set. Table 1 captures high-level statistics of the CPC3 data. The duration of the audio and the number of words spoken across the three sets are identical, while the distribution in terms of ”Hearing loss” the evaluation dataset is skewed toward `Moderate` and away from `Moderately sever`. We anticipate that the model performance, which is independent of the ”Hearing loss” might impact the performance on the evaluation dataset because of this (see Table 1).

Table 1: *CPC3 Data statistics. Hearing loss shows % of (Mild, Moderate and Sever).*

What	Training	Development	Evaluation
# Audio	15520	926	7674
Hearing Loss	(37%, 48%, 15%)	(39%, 49%, 12%)	(30%, 64%, 6%)
Length (μ, σ^2)	5.93 (0.43)	5.70 (0.40)	5.68 (0.38)
Prompt	Yes	Yes	Yes
# words (μ, σ^2)	8.29 (1.21)	8.22 (1.22)	8.26 (1.22)
SI Score	Yes	No	No

3.1. Models Information and Training Procedures

The submission for non-intrusive SI prediction comprises five systematically designed models that integrate complementary feature representations as described in earlier sections.

Model #1 employs only the final encoder layer of the Whisper-medium model, extracting 1024-dimensional high-level acoustic features processed through a 2-layer `Bi-LSTM` (384 units per direction, dropout=0.3), followed by `attention pooling` to obtain a compact 768-dimensional embedding used for SI score prediction via a sigmoid-activated linear layer. Model #2 focuses on semantic representation by using the final decoder layer of the Whisper-small model, extracting 768-dimensional linguistic features. These are also processed through a single-layer `Bi-LSTM` (384 units per direction) and `attention pooling` to yield a 768-dimensional embedding, followed by a sigmoid layer for SI estimation.

Model #3 fuses both encoder and decoder embeddings, each transformed into 768-dimensional representations through independent `Bi-LSTM` and `attention pooling` blocks. These are combined using a learnable softmax-based fusion mechanism, followed by a fully connected (linear) layer (768→1) with a dropout of 0.5, and a final sigmoid-activated output layer for SI prediction.

Model #4 extends the Model #3 architecture by integrating low-level, stable, deformation-invariant TF features derived from a second-order WST. The input waveform $\hat{S}[n]$ is processed through a WST layer configured with a Morlet wavelet filter-bank ($Q = 8$ wavelets per octave, average scale = 2^J, where $J = 6$). The WST computation of order two ($m = 2$), implemented via the `kymatio` 1-D wavelets [18], produces 126-D coefficients that encode localized spectral structures while maintaining invariance to small deformations. These coefficients are passed through a 2-layer `Bi-LSTM` (384 units in each direction, dropout=0.3), followed by `attention pooling` to generate a 768-D feature embedding. The three 768-D embeddings (from the decoder, encoder, and WST branches) are combined using a learnable softmax-based fusion mechanism, which assigns dynamic weights to each branch during training. The fused representation is then passed through a prediction block comprising a fully connected (linear) layer (768→1), with a dropout of 0.3, followed by a sigmoid activation outputs a normalized SI score. Model #5a, an ensemble of the first three models, delivers the best baseline performance by enhancing generalization. Model #5b further improves accuracy through post-processing, reducing RMSE while preserving robustness.

The number of epochs, learning rate, weight decay, and batch size have been set to 10, 4e-5, 3e-5, and 8, respectively. Adam and mean square error (MSE) were chosen as the optimizer and loss function. The train and validation sets are split in the ratio of 90:10 and the model's performance is assessed based on root mean square error (RMSE). All experiments were conducted on an AMD Ryzen Threadripper PRO 3975WX (32-cores), 128GB RAM, Nividia GeForce RTX 3090 GPU. The software environment had Python 3.9 running on Ubuntu 20.04.

4. Results and Discussion

We discuss the performance of all the proposed models for the SI prediction. Table 2 presents the RMSE results for different model configurations evaluated on the CPC3 development set for SI prediction. The encode-only Model #1, using embeddings from the final encoder layer of the Whisper-medium model, achieves an RMSE of 23.45 on the development set, highlighting the utility of high-level acoustic representations. The decoder-only Model #2, leveraging features from the final decoder layer of the Whisper-small model, yields a slightly higher RMSE of 23.63, indicating that semantic features alone are less effective than the acoustic features for the non-intrusive CPC3 task. The encode-decode Model #3 integrating both encoder and decoder features through a weighted mechanism shows improved performance, achieving an RMSE of 23.37. However, when WST-based TF features are integrated as in Model #4 along with encoder and decoder embeddings in the tri-modal fusion model, the RMSE improves to 23.61. This suggests that while the inclusion of invariant spectral features enhances the diversity of the representation, it may also introduce complexity that requires careful balancing. Significantly, the ensemble model (Model #5a) surpasses the others, achieving the lowest RMSE of 22.66 on the development set, demonstrating that the integration of multiple model predictions improves generalization and yields more reliable SI estimates across diverse acoustic scenarios. While using post-processing (Model #5b), the optimized α reduced RMSE on the training data from 20.14 to 19.60. When applied to the development set, the model 5b led to the RMSE reduction from 22.66 to 21.87, demonstrating its effectiveness and generalization capability. Similar performance is anticipated on the evaluation dataset.

Table 2: *RMSE results of development and evaluation sets for the final models.*

Model	**RMSE**	
	Development	**Evaluation**
Model #1 Encode-Only	23.45	-
Model #2 Decode-Only	23.63	-
Model #3 Weighted Combination Encode-Decode	23.37	-
Model #4 Hybrid Weighted Combination Encode-Decode+WST	23.61	-
Model #5a Ensemble	22.66	-
Model #5b Ensemble with Post-processing	**21.87**	-

5. Conclusion

This study introduces a novel speech intelligibility (SI) score prediction framework by combining Whisper-based contextual embeddings with resilient time-frequency (TF) representations extracted from the WST layer. The Whisper model captures rich linguistic and acoustic information through its encoder and decoder layers, while the WST provides translation-invariant and deformation-stable features directly from the raw waveform. The five systematically designed models—encode-only, decode-only, encode-decode weighted embeddings, encode-decode+WST, and ensemble/average model (with & without post processing)—demonstrate the progressive value of combining complementary features within a unified `Bi-LSTM`-based architecture with `attention pooling` and fully connected (linear) layers. Among them, the ensemble model, which aggregates predictions from individual models, achieved the best performance on the CPC3 data. The findings emphasize the efficacy of combining high-level semantic features with low-level invariant features for robust and generalizable SI prediction, particularly for hearing-impaired individuals.

6. References

[1] W. H. Organization *et al.*, *World report on hearing*. World Health Organization, 2021.

[2] Loughrey et al., "Association of age-related hearing loss with cognitive function, cognitive impairment, and dementia: a systematic review and meta-analysis," *JAMA otolaryngology–head & neck surgery*, vol. 144, no. 2, pp. 115–126, 2018.

[3] Ciorba et al., "The impact of hearing loss on the quality of life of elderly adults," *Clinical interv. in aging*, pp. 159–163, 2012.

[4] Shukla et al., "Hearing loss, loneliness, and social isolation: a systematic review," *Otolaryngology–Head and Neck Surgery*, vol. 162, no. 5, pp. 622–633, 2020.

[5] F. R. Lin, R. Thorpe, S. Gordon-Salant, and L. Ferrucci, "Hearing loss prevalence and risk factors among older adults in the united states," *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, vol. 66, no. 5, pp. 582–590, 2011.

[6] Ferguson et al., "Hearing aids for mild to moderate hearing loss in adults," *Cochrane Database of Systematic Reviews*, no. 9, 2017.

[7] Wu et al., "Factors associated with the efficiency of hearing aids for patients with age-related hearing loss," *Clinical interventions in aging*, pp. 485–492, 2019.

[8] R. M. Cox, J. A. Johnson, and J. Xu, "Impact of advanced hearing aid technology on speech understanding for older listeners with mild to moderate, adult-onset, sensorineural hearing loss," *Gerontology*, vol. 60, no. 6, pp. 557–568, 2014.

[9] Andersen et al., "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 24, no. 11, pp. 1908–1920, 2016.

[10] ——, "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proc.*, vol. 26, no. 10, pp. 1925–1939, 2018.

[11] Buragohain et al., "Classification of Motor Imagery Tasks Using EEG Based on Wavelet Scattering Transform and Convolutional Neural Network," *IEEE Sensors Letters*, 2024.

[12] Liu et al., "Exploiting Wavelet Scattering Transform & Squeeze-Excitation Blocks with Cross-Modal Attention for Multi-modal Emotion Recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[13] Ali et al., "Exploiting wavelet scattering transform & 1d-cnn for unmanned aerial vehicle detection," *IEEE Sig. Proc. Lett.*, 2024.

[14] Bali et al., "Infant Cry Scene Detection and Classification Using High-Frequency Wavelet Scattering Transform and Convolutional Recurrent Neural Network," *IEEE Sensors Letters*, 2025.

[15] Radford et al., "Robust speech recognition via large-scale weak supervision," in *Int. Conf. on Machine Lear.* PMLR, 2023, pp. 28 492–28 518.

[16] Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[18] Andreux, Mathieu, et al., "Kymatio: Scattering transforms in Python," *Journal of ML Research*, vol. 21, no. 60, pp. 1–6, 2020.