# A STUDY ON ZERO-SHOT NON-INTRUSIVE SPEECH INTELLIGIBILITY FOR HEARING AIDS USING LARGE LANGUAGE MODELS

*Ryandhimas E. Zezario*[1], *Dyah A.M.G. Wisnu*[12], *Hsin-Min Wang*[1], *Yu Tsao*[1]

[1]Academia Sinica [2]National Chengchi University

## ABSTRACT

This work focuses on zero-shot non-intrusive speech assessment for hearing aids (HA) using large language models (LLMs). Specifically, we introduce GPT-Whisper-HA, an extension of GPT-Whisper, a zero-shot non-intrusive speech assessment model based on LLMs. GPT-Whisper-HA is designed for speech assessment for HA, incorporating MSBG hearing loss and NAL-R simulations to process audio input based on each individual's audiogram, two automatic speech recognition (ASR) modules for audio-to-text representation, and GPT-4o to predict two corresponding scores, followed by score averaging for the final estimated score. Experimental results indicate that GPT-Whisper-HA achieves a 2.59% relative root mean square error (RMSE) improvement over GPT-Whisper, confirming the potential of LLMs for zero-shot speech assessment in predicting subjective intelligibility for HA users.

## 1. INTRODUCTION

Assessing speech intelligibility is a crucial factor in evaluating hearing aid (HA) applications [1]. The most direct approach is to have human listeners recognize words from an audio sample. However, despite their reliability, human-based assessments are costly and impractical for unbiased evaluations, as a sufficient number of listeners is required to minimize bias.

With two series of Clarity Prediction Challenges [1, 2], there is growing interest in developing non-intrusive speech intelligibility prediction models for HA. However, despite notable achievements, a major challenge in deploying reliable non-intrusive speech intelligibility models is the availability of sufficient training data, as data collection can be time-consuming and labor-intensive. Recently, with the growing popularity of large language models (LLM), studies have explored their potential for speech assessment. One such approach is GPT-Whisper [3], which consists of an audio-to-text module and GPT-4o for evaluating the predicted word sequence to obtain estimated scores. Experimental results confirm that GPT-Whisper demonstrates a moderate correlation with intelligibility metrics. Furthermore, considering its notable performance and the need for a reliable non-intrusive speech intelligibility model that can be deployed in a zero-shot setting, we aim to extend GPT-Whisper for zero-shot non-intrusive speech intelligibility assessment in HA.

In this study, we propose an extension of GPT-Whisper for HA, namely GPT-Whisper-HA. GPT-Whisper-HA is specifically designed for speech assessment in hearing aids by incorporating MSBG hearing loss and NAL-R simulations to mimic the audio input based on an individual's hearing profile. Furthermore, two ASR modules are selected as judges to assist in determining whether the input audio is easily recognizable. Unlike typical audio data, hearing-impaired speech may not be easily recognized by ASR systems. We assume that if both a simpler and a more advanced ASR module can easily recognize the word sequence, the audio input is likely clear. Conversely, if the two ASR modules show differing trends, the audio may contain noise or distortions that reduce intelligibility. GPT-4o is then used to generate the assessment score, leveraging naturalness as the key metric, following the original GPT-Whisper framework. Finally, score averaging is performed to obtain the final estimated score.

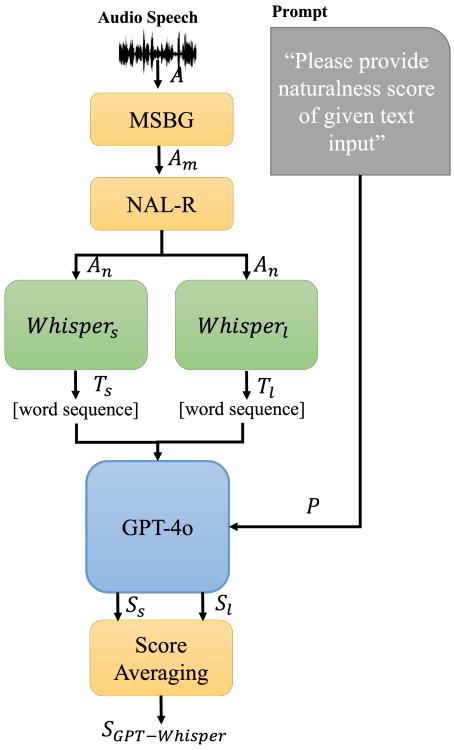



**Fig. 1**. Zero-shot speech intelligibility for HA with GPT-Whisper-HA.

## 2. GPT-WHISPER-HA

The overall framework of GPT-Whisper-HA is shown in Fig. 1, where we selected two variants of Whisper [4]—small ($Whisper_s$) and large ($Whisper_l$)—as the audio-to-text modules due to their robust performance in accurately transcribing speech across diverse acoustic conditions,

Specifically, given an input audio $A$, the audio input is first processed by MSBG and NAL-R before undergoing audio-to-text conversion using Whisper ASR, as defined below:

$$
\begin{aligned}
A_m &= MSBG(A), \\
A_n &= NALR(A_m), \\
T_s &= Whisper_s(A_n), \\
T_l &= Whisper_l(A_n).
\end{aligned} \tag{1}
$$

Based on the predicted word sequences $T_s$ and $T_l$, we use GPT-4o to estimate two assessment scores, $S_s$ and $S_l$. For prompt $P$ engineering, we assess the text representation based on the naturalness score, which measures how similar the predicted text is to human-generated text in terms of fluency, coherence, and context. The final GPT-Whisper score is obtained through score averaging, with the detailed process defined as follows.

$$
\begin{aligned}
S_s &= GPT4o(T_s, P), \\
S_l &= GPT4o(T_l, P), \\
S_{GPT-whisper} &= ScoreAve(S_s, S_l),
\end{aligned} \tag{2}
$$

## 3. EXPERIMENTS

### 3.1. Experimental Setup

The Clarity Prediction Challenge (CPC) 2023 dataset [2] consists of recordings from six talkers and ten enhancement methods, each representing a different HA system, with corresponding subjective intelligibility scores for the output of each HA. Specifically, we select Track 1 of the test set, which contains 305 utterances, to evaluate our system. The evaluation metrics include root mean square error (RMSE), linear correlation coefficient (LCC), and Spearman's rank correlation coefficient (SRCC). A lower RMSE signifies better alignment with ground-truth scores, while higher LCC and SRCC values indicate stronger correlations between predictions and actual scores. All supervised models in this study were trained on the CPC 2023 training set for fair comparison.

### 3.2. Experimental Results

To evaluate the performance of GPT-Whisper-HA, we prepared three system comparisons: MBI-Net [5], MBI-Net+ [6], and GPT-Whisper [3]. MBI-Net and MBI-Net+ achieved top-three performances among the best non-intrusive systems in the first and second Clarity Challenges. Both models were trained using the Track 1 training set of the CPC 2023 dataset. For model architecture, both MBI-Net and MBI-Net+ employ convolutional layers with channel sizes of 16, 32, 64, and 128, a one-layer Bidirectional Long Short-Term Memory (BLSTM) network with 128 nodes, a fully connected layer with 128 neurons, and an attention mechanism. The key difference between the models is that MBI-Net+ utilizes Whisper to extract cross-domain features, while MBI-Net uses WavLM. Additionally, MBI-Net+ incorporates additional modules for objective-based assessment metrics and a classifier module to distinguish different HA model inputs. For GPT-Whisper, we follow the original setup to predict the estimated score.

Table 1 presents the evaluation results of different models in terms of LCC, SRCC, and RMSE. Among the models, MBI-Net+ achieves the highest performance, with an LCC of 0.721, an SRCC of 0.714, and the lowest RMSE of 28.370. MBI-Net, while slightly behind MBI-Net+, also demonstrates competitive performance with an LCC of 0.669, an SRCC of 0.665, and an RMSE of 30.260. These results align with our assumption, as both models benefit from supervised training, allowing them to optimize their predictions using labeled data. Interestingly, GPT-Whisper-based models, which operate in an unsupervised manner, exhibit moderate correlation scores for LCC (0.541) and SRCC (0.501). Despite their lower prediction performance, the predicted scores from GPT-Whisper still show some moderate correlation with subjective human ratings. Notably, our proposed GPT-Whisper-HA, which incorporates HA-related adaptations into a zero-shot modeling strategy, enhances overall prediction performance—improving LCC from 0.541 to 0.570, SRCC from 0.501 to 0.558, and reducing RMSE from 37.019 to 34.767. Furthermore, employing two ASR models as judges for the final assessment score demonstrates consistent improvement in prediction performance in the zero-shot scenario.

**Table 1**. LCC, SRCC, and RMSE results of GPT-Whipser-HAS and other methods.

| Model | Unsupervised | LCC | SRCC | RMSE |
|---|---|---|---|---|
| MBI-Net [5] | No | 0.669 | 0.665 | 30.260 |
| MBI-Net+ [6] | No | **0.721** | **0.714** | **28.370** |
| GPT-Whisper [3] | Yes | 0.541 | 0.501 | 37.019 |
| GPT-Whisper-HA | Yes | 0.570 | 0.558 | 34.767 |

## 4. CONCLUSIONS

In this paper, we proposed GPT-Whisper-HA, a zero-shot model that incorporates HA-related adaptations for speech intelligibility prediction. Our results demonstrate that GPT-Whisper-HA outperforms the baseline GPT-Whisper model, improving LCC from 0.541 to 0.570, SRCC from 0.501 to 0.558, and reducing RMSE from 37.019 to 34.767. Additionally, employing two ASR models as judges for the final assessment further enhances prediction performance. These improvements highlight the effectiveness of integrating hearing aid-specific features into a zero-shot speech intelligibility model.

## 5. REFERENCES


[1] J. Barker, M. Akeroyd, J. Trevor, J. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. Interspeech*, 2022, pp. 3508–3512.

[2] J. Barker, M. A. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. ICASSP*, 2024, pp. 11551–11555.

[3] R. E. Zezario, S. M. Siniscalchi, H.-M. Wang, and Y. Tsao, "A study on zero-shot non-intrusive speech assessment using large language models," *To appear in IEEE ICASSP*, 2025.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28492–28518.

[5] R. E. Zezario, F. Chen, C. S. Fuh, H.-M. Wang, and Y. Tsao, "MBI-Net: A non-intrusive multi-branched speech intelligibility prediction model for hearing aids," in *Proc. INTERSPEECH*, 2022, pp. 3944–3948.

[6] R. E. Zezario, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Non-intrusive speech intelligibility prediction for hearing aids using whisper and metadata," in *Proc. INTERSPEECH*, 2024, pp. 3844–3848.