Predicting Intelligibility for Hearing-Impaired Listeners via Explicit Scores and Pre-trained Feature

Hanglei Zhang, Yanchen Li, Xiang Hao, Yufei Zhang, Jibin Wu, Kay Chen Tan

The Hong Kong Polytechnic University
op131@sjtu.edu.cn

Abstract

This report describes our submitted system for the Clarity Prediction Challenge 3 (CPC3). Our approach primarily draws inspiration from the high-performing systems from Clarity Prediction Challenge 2. The goal was to develop a robust speech intelligibility prediction model by carefully integrating diverse acoustic and listener-specific information, specifically leveraging both explicit evaluation scores (including intrusive and non-intrusive ones) and powerful pre-trained speech foundation models for feature extraction. Through an iterative process of feature selection and combination of these elements, our system demonstrates competitive performance in predicting speech intelligibility for hearing-impaired listeners on the CPC3 dataset.
Index Terms: speech intelligibility, hearing aid, speech quality assessment, Clarity Prediction Challenge 3

1. Introduction

Speech intelligibility prediction is crucial for evaluating hearing aid device performance and enhancing the communication experience for individuals with hearing impairment. The Clarity Prediction Challenge series provides a common platform for researchers to develop and compare various speech intelligibility prediction models. Compared to CPC2 [1], CPC3 introduced a significant change by simplifying detailed audiograms to a generalized hearing loss degree, which necessitates models to adapt to this more abstract listener information. Despite this change, the fundamental principles of intelligibility prediction remain, allowing successful approaches from previous challenges to be adapted and refined. In CPC2, the E009 [2] and E011 [3] systems achieved significant results by combining multimodal features and utilizing hierarchical features from deep learning models, respectively. Inspired by these successful approaches, we designed and implemented a comprehensive intelligibility prediction system for CPC3.

2. System Methodology

2.1. Feature Extraction

Our CPC3 system integrates the core ideas from the E009 and E011 reports, with adaptive improvements to meet the specific requirements of the CPC3 challenge.

We adopted a multi-source feature extraction strategy, primarily including:

Explicit Intelligibility Metrics: We integrated a series of explicit speech quality and intelligibility metrics as input features. These metrics can be broadly categorized as:

- *Intrusive metrics* (require clean reference signal):
  - **STOI (Short-Time Objective Intelligibility) [4]**: evaluated across frequency bands to model dynamic auditory scenarios. Similar to E009 we used two ears' 15-dimensional STOI.
  - **HASPI (Hearing Aid Speech Perception Index) [5]**: incorporates hearing-loss-informed weighting for intelligibility prediction.
  - **PESQ (Perceptual Evaluation of Speech Quality) [6]**: a full-reference metric defined in ITU-T P.862 (2001), estimating perceptual speech quality by comparing a degraded signal against a clean reference. Scores range from approximately –0.5 to 4.5, aligning with MOS.
  - **ScoreQ [7]**: a contrastive regression framework using triplet loss to train a MOS predictor that significantly improves cross-domain generalization, operating in both no-reference (NR) and non-matching reference (NMR) modes to flexibly enhance prediction performance. In our experiments we use both its NR and NMR so it worked in intrusive mode.
- *Non-intrusive metrics* (no reference needed):
  - **UTMOS [8]**: a learning-based MOS predictor from VoiceMOS Challenge 2022, based on ensembles of self-supervised models, ranking top across several test tracks.
  - **NISQA [9]**: CNN+self-attention architecture trained on crowd-sourced ratings, outputting five speech quality indicators: Overall Quality, Noisiness, Coloration, Discontinuity, and Loudness.
  - **DNSMOS-Pro [10]**: a lightweight DNN trained to predict probabilistic MOS distributions for speech quality (posterior mean and variance), significantly smaller than DNSMOS yet performing competitively across datasets.

Audiometric Data: Unlike CPC2, CPC3 simplified detailed audiograms to a generalized hearing loss degree. We accordingly incorporated the hearing loss degree information of hearing-impaired individuals as an input feature for personalized prediction.

Pre-trained Features: We primarily leveraged the noise-robust speech foundation model *Whisper* [11] to extract content-rich features from speech signals. The Whisper model excels at processing speech in complex noisy environments, effectively disentangling signal and noise information, thus providing richer representations for intelligibility prediction. We applied temporal Transformer and layer-wise Transformer structures, similar to those in E011, to aggregate these features across different time steps and model layers.

3. Experiments and Results

We conducted a comprehensive experimental evaluation of our systems on the CPC3 dataset.

For systems relying solely on explicit intelligibility scores, a Logistic Regression model was utilized. This model consists of a single dense layer followed by a sigmoid activation function, and it was optimized using Mean Squared Error (MSE) loss. It takes the concatenated explicit scores and audiometric data as input and directly predicts the intelligibility. Training was performed with a learning rate of 10^{-2} until convergence.

For systems combining explicit scores with pre-trained features, the model was built upon the Transformer-based architecture proposed in E011. This architecture processes each audio channel from the binaural signal through a noise-robust foundation model (Whisper in our primary configuration), yielding sequences of representations at each layer. These representations undergo temporal pooling and linear projection. A key adaptation in our system is the integration of the explicit intelligibility metrics (as described in Section 2.1) into this architecture. Specifically, instead of concatenating the listener's audiogram information (as done in CPC2) before the layer-wise transformer, we concatenate the combined explicit intelligibility scores with the layer representations across the layer axis. This allows our model to leverage both the rich, learned features from the pre-trained model and the interpretable, engineered features from established metrics. The resulting sequence is then fed into a layer-wise transformer, yielding multi-level features. These features are subsequently compressed via global average pooling. The final representations from both channels are averaged and linearly projected to output the predicted speech intelligibility score.

Table 1 summarizes the RMSE results for various explicit intelligibility metrics and their combinations. Unless otherwise specified, all metrics incorporate the hearing loss label of each hearing-impaired listener.

Metric	Type	Dim	RMSE
HASPI (Baseline)	Intrusive	1×2	28.03
STOI	Intrusive	15×2	27.41
PESQ	Intrusive	1×2	31.23
ScoreQ	Intrusive	2×2	29.42
UTMOS	Non-intrusive	1×2	Worse
DNSMOS-Pro	Intrusive	2×2	Worse
NISQA	Non-intrusive	5×2	Worse
STOI + UTMOS	Intrusive	16×2	26.77
STOI + HASPI	Intrusive	16×2	25.45
STOI + HASPI + UTMOS	Intrusive	17×2	26.54
STOI + HASPI + PESQ	Intrusive	17×2	25.80
STOI + HASPI + ScoreQ	Intrusive	18×2	25.60
STOI + HASPI + ScoreQ + PESQ	Intrusive	19×2	**24.94**
All Metrics (w/o NISQA and DNSMOS-Pro)	Intrusive	20×2	25.02
All Metrics (w/o NISQA)	Intrusive	21×2	25.04
All Metrics	Intrusive	26×2	25.80
Whisper	Non-intrusive	–	24.11
Whisper + STOI + HASPI + ScoreQ + PESQ	Intrusive	–	**24.05**

Table 1: *RMSE performance and input dimensions of individual and combined metrics.*

From the results, it is evident that several non-intrusive metrics (UTMOS, DNSMOS-Pro, NISQA) individually showed high loss values, indicating their limited utility in isolation for this task, and were not further evaluated for submission due to their poor performance during training. The combination of STOI, HASPI, ScoreQ, and PESQ yielded a significantly improved RMSE of 24.94, suggesting that ScoreQ, in particular, contributed positively to the overall performance when combined with other effective metrics.

While Whisper provides strong non-intrusive performance, adding explicit intrusive metrics (STOI, HASPI, ScoreQ, PESQ) yields only a minor improvement. This suggests that pre-trained representations already encode much of the perceptual information.

Our final submission used the output from Whisper + STOI + HASPI + ScoreQ + PESQ system (intrusive).

4. Conclusion

In this report, we presented a speech intelligibility prediction system for hearing-impaired listeners in the Clarity Prediction Challenge 3. By integrating both explicit intelligibility metrics—ranging from traditional intrusive measures like STOI, HASPI, and PESQ to more recent learning-based indicators like ScoreQ and UTMOS—and robust pre-trained features from the Whisper model, our system achieved competitive performance. While explicit metrics alone showed meaningful improvements when combined, we observed that the Whisper model alone already delivered strong results, with additional explicit features providing only marginal gains. This suggests that powerful pre-trained representations implicitly capture many aspects of perceptual intelligibility.

5. References

[1] "Clarity prediction challenge 2." [Online]. Available: https://claritychallenge.org/CPC2_announcement_page/

[2] M. Huckvale and G. Hilkhuysen, "Combining acoustic, phonetic, linguistic and audiometric data in an intrusive intelligibility metric for hearing-impaired listeners," Speech, Hearing and Phonetic Sciences, University College London, UK, Tech. Rep., 2024.

[3] S. Cuervo and R. Marxer, "Speech foundation models on intelligibility prediction for hearing-impaired listeners," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Apr. 2024, p. 1421–1425. [Online]. Available: http://dx.doi.org/10.1109/ICASSP48485.2024.10447907

[4] "C. taal, "stoi – short-time objective intelligibility measure". matlab implementation:." [Online]. Available: https://ceestaal.nl/code/

[5] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (haspi)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[6] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[7] A. Ragano, J. Skoglund, and A. Hines, "Scoreq: Speech quality assessment with contrastive regression," *Advances in Neural Information Processing Systems*, vol. 37, pp. 105 702–105 729, 2024.

[8] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[9] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.

[10] F. Cumlin, X. Liang, V. Ungureanu, C. K. Reddy, C. Schüldt, and S. Chatterjee, "Dnsmos pro: A reduced-size dnn for probabilistic mos of speech," in *Proc. Interspeech*, vol. 2024, 2024, pp. 4818–4822.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.