WHISPER MULTI-CANDIDATE SCORING FOR LYRICS INTELLIGIBILITY PREDICTION

Yuto Kondo, Hirokazu Kameoka, Kou Tanaka, Takuhiro Kaneko

NTT, Inc.

ABSTRACT

We describe our submission to the ICASSP 2026 Cadenza Challenge task of predicting lyrics intelligibility under simulated hearing loss. Our system first derives a “Whisper multi-candidate mean score” by averaging word-correctness over multiple recognition hypotheses generated by Whisper for the left and right channels. This score is then combined with meta information (such as utterance duration, word count, and hearing-loss grade) and self-supervised acoustic features extracted from the Whisper encoder within a two-branch regression model. Experiments on the official training and evaluation sets show that the proposed system achieves a lower root mean square error (RMSE) than the official Whisper 1-best baseline, indicating the effectiveness of our feature design and model architecture for this task.

Index Terms— lyrics intelligibility, hearing loss, Whisper, self-supervised learning

1. INTRODUCTION

This paper presents our lyrics intelligibility prediction system submitted to the ICASSP 2026 Cadenza Challenge [1]. The task is to predict, for each audio excerpt processed by a hearing-loss simulator for mild and moderate hearing loss, the word-level correctness score (a continuous value between 0 and 1) that would be obtained when normal-hearing listeners attempt to recognize the lyrics.

The official baseline system is based on Whisper [2], a high-performance end-to-end automatic speech recognition (ASR) model. For each input signal, Whisper is used to obtain a single 1-best transcription, and the word-correctness between this hypothesis and the reference lyrics is directly used as the predicted intelligibility score. Thanks to the strong ASR performance of Whisper, this simple 1-best-based approach already shows a relatively high correlation with the target lyrics intelligibility, and serves as a strong baseline.

However, Whisper is trained to optimize ASR performance rather than to directly model human lyrics intelligibility under hearing-loss and noisy conditions. In particular, the 1-best output selected by Whisper does not necessarily coincide with the hypothesis that best reflects how easily a human listener can understand the lyrics in the presence of distortion and background interference. This suggests that there is a limit to the correlation that can be achieved when relying solely on the 1-best correctness score.

In this work, we seek to improve the prediction of lyrics intelligibility by exploiting richer information derived from Whisper. Specifically, we utilize multiple recognition candidates that Whisper internally generates, together with their confidence measures, and aggregate them into a single score that we refer to as the Whisper multi-candidate mean score. We further combine this score with simple meta information such as utterance duration, number of words, and hearing-loss grade, as well as self-supervised acoustic features extracted from the Whisper base encoder[1]. These features are fed into a two-branch regression model that outputs the final intelligibility score. In the following, we describe the model architecture, training procedure, and evaluation results compared with the official baseline.

2. PROPOSED MODEL

The proposed model consists of two branches: an acoustic branch based on self-supervised features from the Whisper encoder, and a meta-information branch based on statistics computed from multiple Whisper candidates and auxiliary metadata. The outputs of the two branches are fused to predict a single scalar lyrics intelligibility score.

In the acoustic branch, we use the encoder of the pre-trained Whisper base model as a self-supervised feature extractor. The input mono waveform is first converted to Whisper input features using the standard preprocessing, and is then passed through the encoder to obtain frame-wise hidden representations. For each time frame, we extract hidden states from the last four layers of the Whisper encoder and compute a fixed weighted sum with weights $(0.5, 1.0, 1.5, 1.0)$, resulting in a 512-dimensional feature vector per frame. These weights were chosen empirically to reflect differences in the information captured by the encoder layers.

The resulting 512-dimensional feature sequence is then processed by a lightweight Transformer-based encoder [3]. Concretely, we apply a linear projection to map the 512-dimensional features to a 192-dimensional space, followed by a 2-layer Transformer encoder with 4 attention heads and a feed-forward dimension of 384. After this temporal modeling, we apply mean pooling over time to obtain a single 192-dimensional acoustic embedding that summarizes the entire utterance.

In the meta-information branch, we construct a 10-dimensional feature vector from multiple Whisper candidates and auxiliary information. For each sample, we generate up to 15 recognition candidates independently for the left and right channels by combining one deterministic decoding run (temperature $T = 0.0$) with additional stochastic decoding runs using sampling at $T = 0.5$ with `best_of`=5. For each candidate, we compute the word-correctness with respect to the reference lyrics. Candidates that are clearly unnatural, such as those with extreme repetition of the same word, an excessive or extremely small number of words, or a high proportion of non-linguistic symbols, are removed using simple heuristics before aggregation.

From the remaining candidates, we compute the following statistics:

- the simple average of correctness over all candidates from both channels (Whisper multi-candidate mean score),
- the number of words in the reference lyrics,

[1] `https://github.com/openai/whisper`

- the duration of the degraded signal in seconds,
- a scalar hearing-loss value obtained by mapping *No Loss*, *Mild*, and *Moderate* to 0, 0.5, and 1, respectively,
- the mean correctness for the left and right channels,
- the maximum correctness for the left and right channels,
- the mean average log-probability for the left and right channels.

Here, the average log-probability of a candidate is defined as the average, over tokens, of the log posterior probabilities assigned by the Whisper decoder, and is used as a confidence measure indicating how plausible the candidate is according to the ASR model.

These 10 scalar values form the input to a small multilayer perceptron (MLP). The MLP has two fully connected layers with a hidden dimension of 32, each followed by a ReLU activation and dropout. Its output is a 64-dimensional meta-feature embedding.

Finally, we concatenate the 192-dimensional acoustic embedding and the 64-dimensional meta-feature embedding, and feed the resulting 256-dimensional vector into a fusion network. The fusion network consists of a fully connected layer with hidden dimension 192, a ReLU activation and dropout, followed by a final linear layer that outputs a single scalar. This scalar represents the predicted lyrics intelligibility score in the range $[0, 1]$. Overall, the model can be viewed as a regression network that refines the information contained in the Whisper-based score by incorporating additional acoustic and meta-information.

The proposed method uses the reference text and the processed signal, while not making use of the unprocessed signal.

3. TRAINING PROCEDURE

We use the 8,802 training samples provided in the Clarity Cadenza Challenge [4]. These samples are randomly split into 7,802 pseudo-training samples and 1,000 pseudo-validation samples. All models are trained only on the pseudo-training split, and hyperparameter selection and model selection are based on the pseudo-validation split.

We optimize the proposed model using the Adam optimizer with a learning rate of 5×10^{-5}, a batch size of 4, and 30 training epochs. The objective is to minimize the root mean square error (RMSE) between the predicted scores and the ground-truth correctness scores.

In addition to training on the Cadenza data alone, we also consider a variant of the model that is pretrained on data from the Clarity CPC3 Challenge[2]. The CPC3 dataset consists of noisy speech signals with intelligibility ratings obtained from listeners with hearing loss. We treat the CPC3 task as the same type of regression problem and perform 5 epochs of pretraining using 8,202 samples with a hearing-loss grade corresponding to mild loss. The pretrained model is then fine-tuned on the Cadenza pseudo-training split using the same optimization settings as above.

For each random seed, we train the model and record the RMSE on the pseudo-validation split at every epoch. The epoch with the smallest pseudo-validation RMSE for that seed is selected as the representative model for that run. Among all runs, the model with the best pseudo-validation RMSE is finally chosen as our submitted system for the Cadenza evaluation.

4. RESULTS AND DISCUSSION

We first examine the effect of using multiple Whisper candidates. When we follow the official baseline approach and use only the Whisper 1-best transcription to compute the correctness score, the RMSE on the training data is 0.327. In contrast, when we compute the simple mean correctness over all available candidates from both channels (after removing obviously unnatural candidates), the RMSE is reduced to 0.313. This indicates that aggregating information from multiple plausible hypotheses provides a more robust indicator of lyrics intelligibility than relying solely on the 1-best output.

We then evaluate the full two-branch regression model that takes the multi-candidate mean score, meta features, and acoustic features as input. When trained only on the Cadenza data, the proposed model achieves a pseudo-validation RMSE of 0.271, which is clearly lower than the RMSE obtained by using the multi-candidate mean score alone (0.313). This suggests that self-supervised acoustic features from the Whisper encoder, together with simple meta information such as utterance duration, word count, and hearing-loss grade, are effective in capturing residual variability in lyrics intelligibility that is not explained by the multi-candidate score.

When we further apply pretraining on the CPC3 data before fine-tuning on Cadenza, the pseudo-validation RMSE is reduced to 0.269. Although the numerical improvement from 0.271 to 0.269 is modest, it indicates that exposure to additional data involving speech under hearing-loss-related conditions can slightly enhance the representation learned by the model and benefit the downstream Cadenza task.

On the official validation and evaluation sets, our final submitted model achieves RMSEs of 27.0 and 27.4, respectively, whereas the official Whisper 1-best baseline reports RMSEs of 29.3 and 29.1 on the same scales. These results confirm that combining the Whisper multi-candidate mean score with meta information and self-supervised acoustic features within a regression framework is effective for predicting lyrics intelligibility under simulated hearing loss.

5. CONCLUSION

We have presented our system for the Clarity Cadenza Challenge lyrics intelligibility prediction task. The system defines a Whisper multi-candidate mean score by averaging word-correctness over multiple recognition hypotheses, and combines this score with meta information and self-supervised acoustic features extracted from the Whisper base encoder in a two-branch regression model. Experiments on the challenge data show that the proposed approach achieves lower RMSE than the official Whisper 1-best baseline on both pseudo-validation and official evaluation sets, indicating that the proposed feature design and model architecture are effective for predicting lyrics intelligibility under hearing-loss simulation.

Future work includes exploring word-level models that directly predict the correctness of individual words, and incorporating word-alignment information derived from clean audio to better model the temporal structure of misperceptions.

6. REFERENCES

[1] Gerardo Roa-Dabike, Jon P. Barker, Trevor J. Cox, Michael A. Akeroyd, Scott Bannister, Bruno Fazenda, Jennifer Firth, Simone Graetzer, Alinka Greasley, Rebecca R. Vos, and William M. Whitmer, "Overview of the icassp 2026 cadenza challenge: Predicting lyric intelligibility," in *Proc. IEEE ICASSP*, 2026, To appear.

[2] `https://claritychallenge.org/docs/cpc3/cpc3_intro`

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

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez N, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[4] Gerardo Roa-Dabike, Trevor J. Cox, Jon P. Barker, Bruno M. Fazenda, Simone Graetzer, Rebecca R. Vos, Michael A. Akeroyd, Jennifer Firth, William M. Whitmer, Scott Bannister, and Alinka Greasley, "The cadenza lyric intelligibility prediction (clip) dataset," *Data in Brief*, 2025.