Speech Intelligibility Data Improves Music Intelligibility

Adeola O. Aderemi*
Department of Electrical Engineering
Boğaziçi University
Istanbul, Türkiye
adeola.aderemi@std.bogazici.edu.tr

Abstract

Music is better enjoyed when listeners can understand the lyrics. This is often difficult and is especially hard for listeners using hearing aids. In this work, we present an approach to learn a metric to automatically quantify how intelligible the lyrics of a given piece of music is. The ability to automatically do this correctly will enable the development of algorithms to improve music understanding, especially for hearing aid users. In this work, we develop a transformer-based system to directly predict lyric intelligibility using the provided processed audio signal (with hearing loss already simulated). Our system (T017) achieved a RMSE value of 26.82% on the provided validation set and 26.67% on the evaluation set, coming fourth on the final evaluation leaderboard. Our system will appear in the challenge overview Roa-Dabike et al. [2026].

1 Introduction

Lyric intelligibility prediction is the problem of creating a metric to automatically assess how much of a given piece of music would be understandable to the average human. As in speech technology, having an automatic way to predict intelligibility can enable creating algorithms, which could make listening to music much more enjoyable to humans, especially those using hearing aids Cadenza Team.

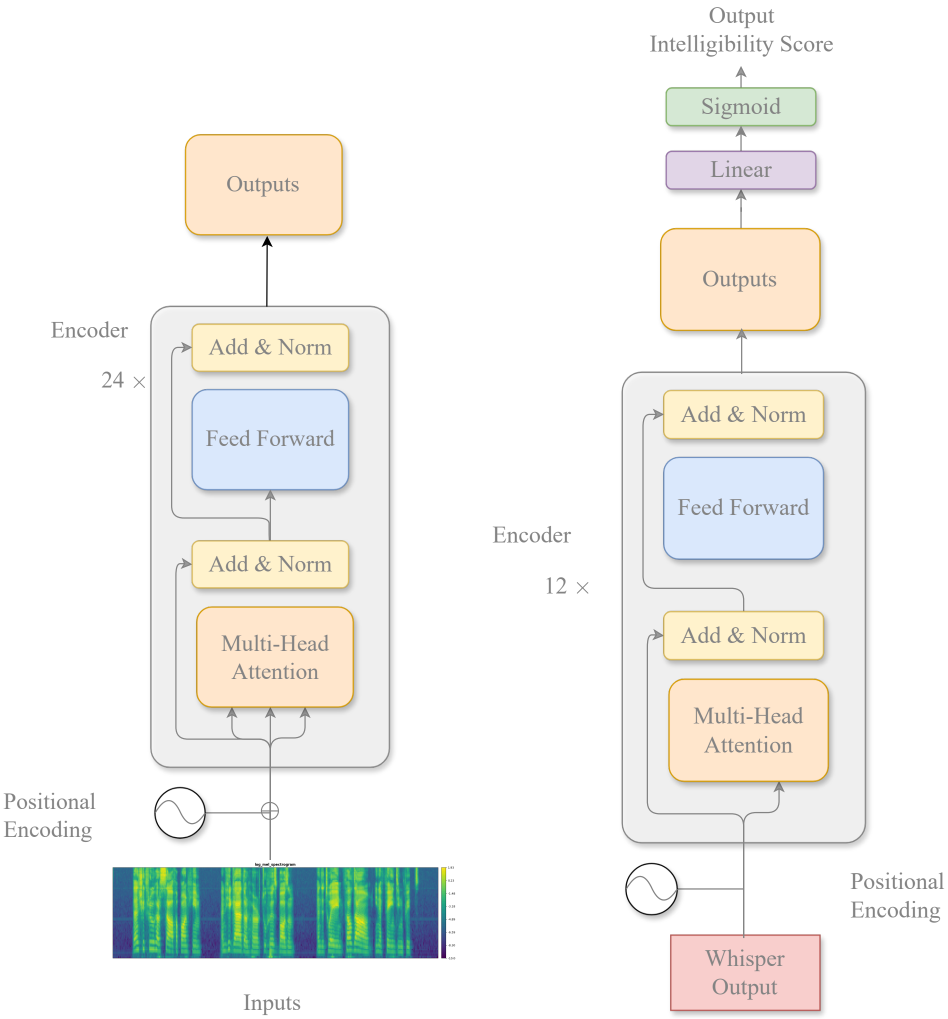
The ICASSP 2026 Cadenza challenge Cadenza Team aims to enable the creation of such a metric by providing English music data, some with hearing loss simulated and some without, and the corresponding intelligibility score obtained from native English-speaking PhD students from the Universities of Salford and Sheffield Roa-Dabike et al. [2025]. The aim of the challenge is to predict the correct intelligibility score, given the provided music data.

Similar to the Cadenza challenge is the Clarity Prediciton Challenge (CPC) CPC3 Team. The aim in that challenge is to predict intelligibility but for speech data. In addition to the speech data provided in that challenge having a hearing loss simulated, the data also has noise added to it. Deep learning approaches are the leading solutions in that challenge Barker et al. [2024].

Following a similar deep learning-based approach, we show in this work that an approach that first extracts audio features from a frozen pretrained speech foundation model using only the processed signal and then uses those features to train another transformer from the ground up leads to competitive results on this (Cadenza) challenge. Furthermore, we show that taking samples from the CPC3 data, we can finetune our model to obtain improvements in the final RMSE scores.

*

Preprint. Under review.



(a) Whisper Model

(b) Transformer Encoder Architecture

Figure 1: System illustration. With (a) showing the pretrained Whisper encoder and (b) showing the subsequent transformer encoder architecture described in this work.

2 Model

Our model consists of two stages. We first extract features by passing the processed audio signal to a speech foundation model. After the feature extraction, we then use the extracted features as inputs into a randomly initialized transformer encoder.

For the speech foundation model, we use the encoder of the medium English version of the Whisper model from OpenAI Radford et al. [2022]. This model requires the input sample to be resampled to 16kHz and converted to a Log-Mel spectrogram. We choose our features to be the output of the final layer of the Whisper encoder. Our subsequent encoder is simply a transformer encoder with 12 heads and 12 layers. From our encoder, we obtain the output as the final layer of the encoder output, ignoring the middle layers. At each time step, we project the model output using a linear layer followed by a non linearity. We average the results across time and then use a linear layer to obtain our intelligibility score. A sigmoid layer is then used to map this output to the $[0, 1]$ range. The system is shown in Figure 1.

3 Experimental setup

For our baseline model, we extracted the features from the Whisper model, as already discussed. We trained using the default training parameters using the HuggingFace Transformers library Wolf et al. [2020]. We used a learning rate of $1e-3$ for training and train for 10 epochs with a batch size of 16. We additionally used SpecAugment for data augmentation during training as a regularization method

Table 1: RMSE comparison across methods for validation and evaluation datasets (↓ indicates lower is better).

Method	Valid ↓	Eval ↓
Provided Baseline (STOI-based)	36.11	34.89
Provided Baseline (Whisper-based)	29.32	29.08
Our Baseline	26.92	26.81
Our Baseline + CPC3 data	**26.82**	**26.67**

Park et al. [2019]. For this baseline, we performed 5 fold cross-validation. Training was done using A100 GPUs available on Google Colab. In this case, training took 50 minutes for the 10 epochs.

After obtaining the baseline system, we obtained data from the CPC3 challenge. The distribution of intelligibility scores differed from that of our problem as there was a high fraction of 100% correctness scores in the CPC3 data. To make the distribution similar to that of our problem, we selected only the data samples where the score is not 100%. We used these data to augment the original per fold training set and then finetuned the baseline models using the new augmented data. For both the baseline and finetuned models, our final predictions for each input were the average of the outputs for all five models. Results on the validation and evaluation sets for the baseline and the models trained using the augmented data are shown in Table 1.

4 Results and analysis

As can be seen in the table, our baseline system achieved a RMSE value of 26.92% on the validation set and 26.81% on the evaluation set, outperforming the best baseline provided by the organizers 29.32% and 29.08% on the validation and evaluation sets respectively. Additionally, using the CPC3 data improved the performance of the baseline model on both the validation and final evaluation sets.

4.1 Post-competition experiments

In the preceding experiments, the input to the downstream architecture for intelligibility prediction has been the final output layer of the pretrained Whisper encoder. It has been found elsewhere that this final set of representations may not necessarily be the best for a given downstream task, hence authors have often experimented with the different encoder layer outputs to obtain the layer which can lead to the best downstream performance. Given a limited compute budget, running the same experiment for up to 25 different encoder layer output may not be the best use of time. The authors in Yang et al. [2024] instead chose to use a weighted sum of all the encoder output layers, with the idea that by learning the weights in this weighted sum using a neural network, it should be possible to automatically find the best performing layer without iterating over each layer output. They reported competitive results with SOTA methods in their work. We followed a similar approach to this method, using the implementation provided in the HuggingFace library, and report the results in Table 2. The result using this method is worse than that obtained using only the final layer, suggesting that adding the representations from the lower layers may not be useful in this case. It was observed that, although the final layer outputs from the HuggingFace Whisper encoder output is normalized, the outputs from the other encoder layers obtained from the pretrained model are not normalized. To alleviate this discrepancy, we proposed to normalize the outputs of every encoder layer and retrain the model with the weighted sum. The result with this normalization step is also shown in Table 2. We see that the normalization improved the results, although the final result is still worse than simply using the final layer as reported in Table 1. We additionally experimented with feeding the normalized encoder outputs individually into the downstream model, but did not see any meaningful improvement on the final layer by the lower layers, so we decided to ignore this direction of research.

5 Conclusions

We have been able to show that transformer-based models can be used to directly estimate intelligibility scores without the need to first obtain a transcript. Additionally, we have been able to show that speech intelligibility data can be used to improve the music intelligibility model. This implies

Table 2: RMSE comparison for weighted sum of encoding layers (↓ indicates lower is better).

Method	**Valid ↓**	**Eval ↓**
No normalization	28.26	28.41
Normalized	27.73	27.69

that available data can be easily used to improve systems without the need to collect extra expensive data. In future work, we would like to explore longer training and how the downstream transformer is initialized. Also, it would be interesting to explore incorporating the reference text into the model and exploring how much improvement can be gained by doing so.

References

Jon Barker, Michael A. Akeroyd, Will Bailey, Trevor J. Cox, John F. Culling, Jennifer Firth, Simone Graetzer, and Graham Naylor. The 2nd Clarity Prediction Challenge: A Machine Learning Challenge for Hearing Aid Intelligibility Prediction. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11551–11555, 2024. doi: 10.1109/ICASSP48485.2024.10446441.

Cadenza Team. ICASSP 2026 Cadenza Challenge: Predicting Lyric Intelligibility. `https://cadenzachallenge.org/docs/clip1/intro`. Accessed: 2025-12-01.

CPC3 Team. The 3rd Clarity Prediction Challenge. `https://claritychallenge.org/docs/cpc3/cpc3_intro`. Accessed: 2025-12-01.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, page 2613–2617. ISCA, September 2019. doi: 10.21437/interspeech.2019-2680. URL `http://dx.doi.org/10.21437/Interspeech.2019-2680`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL `https://arxiv.org/abs/2212.04356`.

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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu-hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Shang-Wen Li, Abdelrahman Mohamed, Shinji Watanabe, and Hung-yi Lee. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899, 2024. doi: 10.1109/TASLP.2024.3389631.