LANGUAGE-FREE LYRICS INTELLIGIBILITY MODELING THROUGH FRACTIONAL REGRESSION AND CONTRASTIVE MULTITASK TRAINING

Kirill Abrosimov, George Tzanetakis

Music Intelligence and Sound Technology Interdisciplinary Centre (MISTIC)
Faculty of Engineering and Computer Science
University of Victoria, Canada

ABSTRACT

We present a system for predicting lyrics intelligibility in music using multitask contrastive fine-tuning of the MERT model. The system operates solely on audio signals, without relying on linguistic information. Our approach combines fractional regression for correctness prediction, classification of hearing loss, and Rank-and-Contrast loss to structure the latent space according to intelligibility. Audio recordings were segmented into overlapping 3-second fragments for training, and the model was fine-tuned with 80% of MERT parameters frozen. Evaluation on validation/test subsets achieved an RMSE of 31.44/32.30 and an NCC of 0.504/0.43, outperforming non-intrusive baselines while remaining below ASR-based methods. This work demonstrates the feasibility of predicting lyrics intelligibility using a purely signal-based approach and provides a foundation for further research on robust, language-independent music understanding.

***Index Terms*—** Lyrics Intelligibility, MERT, Fractional Regression, Multitask learning, Rank-and-Contrast loss

1. INTRODUCTION

Modern information technologies, combined with fields such as psychology and medicine, enable the creation of accessible resources for people with disabilities, including those with vision or hearing impairments. However, researchers rarely investigate how understandable song lyrics are for listeners with hearing loss. Understanding lyrics is essential for music enjoyment [1], yet people with hearing impairments often struggle to perceive them clearly [2]. Moreover, contemporary performers increasingly produce works in which it is difficult for even ordinary listeners to comprehend the lyrics.

The ICASSP 2026 Cadenza Challenge [3] focuses on predicting lyrics correctness (a numerical measure of how many words a listener can recognize) in three groups: normal hearing, mild hearing loss, and moderate hearing loss. To simulate hearing-impaired perception, organizers transformed audio fragments with a hearing loss simulator and collected human annotations on recognized words [4]. Correctness was computed as the proportion of correctly identified words, and the Challenge aims to develop systems that can automatically predict this metric.

The organizers provided two baselines: one based on signal analysis and Short-Time Objective Intelligibility (STOI), and another based on an automatic speech recognition (ASR) approach. STOI [5, 6] is a metric that measures how intelligible speech is by comparing short-time segments of the original and processed signals. In the second approach, the Whisper model [7] was used to predict words in the signal, and correctness was subsequently calculated.

In this competition, we decided to focus on signal analysis without relying on linguistic features or the language itself, as the goal of predicting correctness could serve as an effective metric for evaluating the intelligibility of new songs or songs generated by state-of-the-art models. If the metric depended on language-specific features, it would either require separate models tailored to each language (as in the ASR approach) or produce unreliable results. At the same time, the audio signal itself contains rich information, and many musical parameters (such as genre, instrumentation, or the performer's style) affect the predicted correctness, while the examples remain independent of the language used.

2. METHODOLOGY

MERT was chosen as the base model. MERT [8] is a large-scale audio representation model trained on music and speech data, which extracts rich acoustic features such as timbre, harmonic content, temporal structure, and high-level semantic embeddings from the signal. Its main advantages include strong generalization across audio domains and high performance even when fine-tuned on relatively small datasets. Due to limited computational resources, we used the 95M-parameter version of MERT, and all audio recordings were segmented into 3-second fragments with a 1-second shift. From the standpoint of correctness, this is not ideal, as only part of the lyrics appears in each fragment, meaning that correctness should technically be recalculated for each segment. However, since our work focuses on estimating the acoustic clarity of words rather than their exact alignment, we decided

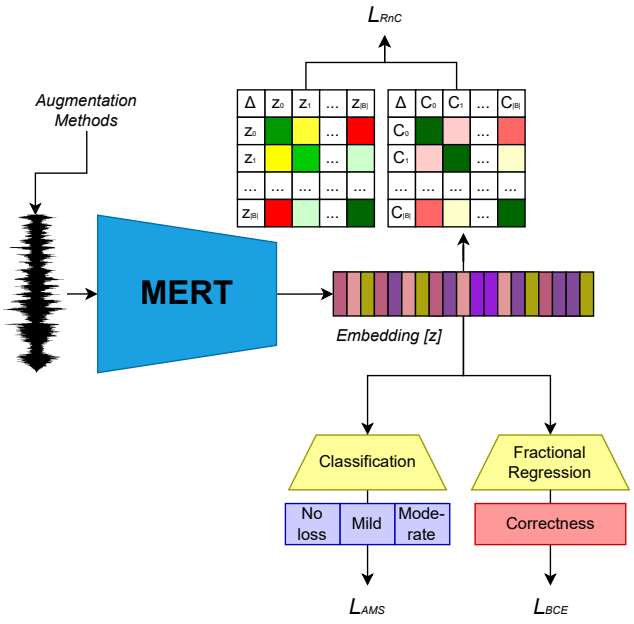

Fig. 1. Overall architecture of the proposed approach based on multitask fine-tuning of MERT.

to omit this aspect. For the final prediction, we averaged the correctness values across all fragments belonging to the same $signal_{id}$.

To increase the amount of data and improve system robustness, we applied several augmentation techniques (pitch shifting, time stretching, adding noise). However, augmentations must be used carefully, as changes in the signal directly affect the underlying correctness value.

The core component of our training pipeline is multitask learning (Figure 1): predicting correctness using fractional regression, predicting the hearing-loss class, and contrastive training using Rank-and-Contrast loss. Fractional regression models a continuous target bounded between 0 and 1 by applying a sigmoid activation and optimizing cross-entropy loss, making it suitable for correctness scores interpreted as probabilities [9, 10]. For classification, we used Additive Margin Softmax loss (AM-Softmax or AMS loss) [11] to increase class separability, which, according to our hypothesis, should improve system performance, since there is a clear relationship between hearing-loss class and correctness, as demonstrated by the organizers in their EDA of the dataset.

$$\mathcal{L}_{AMS} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos\theta_{y_i}-m)}}{e^{s(\cos\theta_{y_i}-m)}+\sum_{j\neq y_i}e^{s\cos\theta_j}}, \quad (1)$$

where N is the batch size, y_i is the ground-truth class of the i-th sample, and $\cos\theta_{y_i}$ denotes the cosine similarity between the normalized embedding and the weight vector of the correct class. The parameter m is the additive angular margin that increases the separation between classes, while s is a scaling factor that stabilizes optimization.

The Rank-and-Contrast loss [12] organizes the latent space by combining embedding similarity with the regression objective. It encourages embeddings of similar samples to be closer while pushing apart those corresponding to different correctness values. The loss operates on positive and negative pairs of embeddings, using a temperature-scaled similarity function and contrastive normalization over the set of negatives.

$$\begin{aligned}\mathcal{L}_{\text{RNC}} &= \frac{1}{2N}\sum_{i=1}^{2N}\frac{\ell_i}{2N-1},\\ \ell_i &= \sum_{\substack{j=1\\ j\neq i}}^{2N} -\log\frac{\exp(\text{sim}(v_i,v_j)/\tau)}{\sum_{v_k\in S_{i,j}}\exp(\text{sim}(v_i,v_k)/\tau)},\end{aligned} \quad (2)$$

where v_i and v_j are embedding vectors of two augmented views of the same audio sample, and $\text{sim}(v_i, v_j)$ denotes their cosine similarity. The temperature τ adjusts the sharpness of the similarity distribution, and $S_{i,j}$ is the set of negative samples for the anchor v_i. The batch contains $2N$ embeddings, and each anchor is contrasted with all other elements except itself. This design encourages embeddings corresponding to similar correctness values to cluster together, while separating those with different values in the latent space.

The final loss used for fine-tuning was defined as:

$$\mathcal{L} = \gamma_{\text{class}}\mathcal{L}_{\text{AMS}} + \gamma_{\text{reg}}\mathcal{L}_{\text{BCE}} + \gamma_{\text{rnc}}\mathcal{L}_{\text{RNC}}. \quad (3)$$

3. RESULTS AND DISCUSSION

Each $signal_{id}$ was segmented into 3-second fragments with a 1-second shift. Fine-tuning was performed on MERT with 80% of the parameters frozen. The hyperparameters were: $\gamma_{\text{class}} = 0.5$, $m = 0.3$, $s = 30$ for classification; $\gamma_{\text{reg}} = 5.0$ for regression; $\gamma_{\text{rnc}} = 2.0$, $\tau = 1.0$ for RnC. Fine-tuning was done on an RTX 4080 laptop GPU for 30 epochs with a batch size of 16 (32 embeddings for RnC). We used the AdamW optimization algorithm with a learning rate ranging from 1×10^{-5} to 1×10^{-6}, together with a `CosineAnnealingLR` scheduler.

For evaluation, signals were similarly segmented, predictions averaged, and multiplied by 100. This yielded RMSE = 31.44 and NCC = 0.5043 in validation subset and RMSE=32.30, NCC = 0.43 in evaluation subset. As shown in Table 1, performance exceeds the non-intrusive baseline, though it does not surpass the ASR approach. Figure 2 shows a t-SNE visualization of the latent space, displaying a smooth gradient of correctness values across the audio samples.

While the results are moderate, solving the task without linguistic information is inherently more challenging. Language-free models offer more generalizable predictions

Table 1. Comparison of methods on lyrics intelligibility prediction for validation and evaluation subsets.

Method	Validation		Evaluation	
	RMSE ↓	NCC ↑	RMSE ↓	NCC ↑
Baseline STOI	36.11	0.14	34.89	0.21
Baseline Whisper (ASR)	29.32	0.59	29.08	0.58
Our approach	31.44	0.504	32.30	0.43

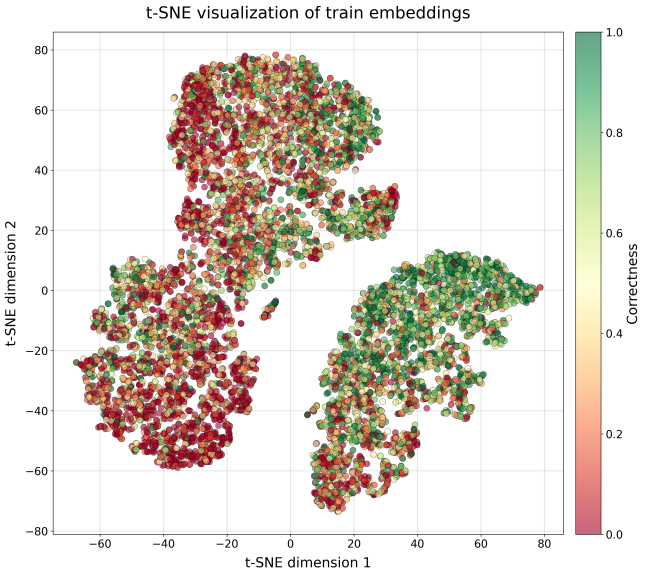

Fig. 2. t-SNE visualization of the embedding space produced by the fine-tuned MERT model.

across different languages, unlike ASR-based approaches that rely on language-specific features. Note that each word contributes equally to correctness, though some phrases are easier to understand yet less informative. For instance, the line *"it's a, it's a, it's a* `sin`*"* may have a high correctness score even if a missing word renders the phrase meaningless. Other challenges include differences between native and non-native speakers and the interpretability of intermediate correctness scores (e.g., 0.6 vs. 0.8), which may correspond to minor function-word errors.

4. CONCLUSION

In this study, we proposed a system based on multitask contrastive fine-tuning of the MERT model for predicting lyrics intelligibility. While the system does not achieve state-of-the-art performance, it has notable advantages: it operates solely on the audio signal without relying on linguistic information, which enables the development of a robust model applicable to multilingual datasets.

5. REFERENCES

[1] Philip A Fine and Jane Ginsborg, "Making myself understood: perceived factors affecting the intelligibility of sung text," *Frontiers in psychology*, vol. 5, pp. 809, 2014.

[2] Alinka Greasley, Harriet Crook, and Robert Fulford, "Music listening and hearing aids: perspectives from audiologists and their patients," *International Journal of Audiology*, vol. 59, no. 9, pp. 694–706, 2020.

[3] 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.

[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.

[5] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[6] Bidisha Sharma and Ye Wang, "Automatic evaluation of song intelligibility using singing adapted stoi and vocal-specific features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 319–331, 2019.

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[8] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, et al., "Mert: Acoustic music understanding model with large-scale

self-supervised training," in *International Conference on Learning Representations*, 2024.

[9] Joao A Bastos, "Predicting bank loan recovery rates with neural networks," *CEMAPRE, ISEG, Technical University of Lisbon*, 2010.

[10] Leslie E Papke and Jeffrey M Wooldridge, "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates," *Journal of applied econometrics*, vol. 11, no. 6, pp. 619–632, 1996.

[11] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[12] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi, "Rank-n-contrast: learning continuous representations for regression," *Advances in Neural Information Processing Systems*, vol. 36, pp. 17882–17903, 2023.