ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11465056

MULTI-MODAL FEATURE FUSION AND STACKING ENSEMBLE LEARNING FOR LYRIC INTELLIGIBILITY PREDICTION

Candy Olivia Mawalim, Xiajie Zhou, Masashi Unoki

Japan Advanced Institute of Science and Technology

ABSTRACT

Predicting lyric intelligibility across varying levels of hearing loss is challenging. This study investigates a multi-modal feature set—including acoustics, speech recognition error metrics, and linguistic features—for the prediction. The proposed method achieved significant improvements over the baseline Whisper model on the Cadenza dataset. Specifically, the root mean square error on the evaluation set was reduced from 29.08% to 27.22%. This study identifies specific acoustic and linguistic features, such as prosodic cues and text embeddings, that drive the differences in how speech and lyric intelligibility are perceived.

***Index Terms*—** lyrics intelligibility, hearing loss, ensemble, vocal separation

1. INTRODUCTION

The 1st Cadenza Lyric Intelligibility Prediction (CLIP1) Challenge [1] [1] focuses on predicting how clearly listeners perceive lyrics in music, a key factor in music enjoyment for hearing-aid users. CLIP1 provides a dataset of songs, presented both in unprocessed form and after hearing loss simulation, and the ground truth word-correct-rate obtained from normal-hearing listeners in the presence of accompaniment. Lyric intelligibility is affected not only by masking from the musical accompaniment but also by the basic characteristics of the singing voice. Elements such as pitch changes, articulation, syllable rate, and vocal tone can influence lyric intelligibility. Motivated by these observations, Sharma and Wang [2] developed an approach to automatically evaluating lyric intelligibility, integrating a singing-adapted short-term objective intelligibility (STOI) measure with vocal-specific acoustic features.

While prior work like Clarity Prediction Challenge 3 (CPC3)[2] focused on predicting speech-in-noise intelligibility processed by hearing aids, CLIP1 tackles the fundamentally more complex task of sung lyric intelligibility. CLIP1 is influenced by broader musical factors (vocal style, genre, mixing) and lacks the segmental regularity of spoken sentences. Furthermore, unlike CPC3, CLIP1 does not provide a clean vocal reference. This distinction is critical for defining prediction modes: in CLIP1, an intrusive method relies on accessing the ground-truth lyric transcriptions (e.g., for automatic speech recognition (ASR) error calculation), which is not always available in real-world use, whereas a non-intrusive method must infer lyric intelligibility solely from complex audio mixtures without any textual or clean vocal references, highlighting why techniques effective for speech do not directly transfer to the musical domain.

This study aims to investigate the different factors governing speech intelligibility versus lyric intelligibility in audio degraded by simulated hearing loss. A Stacked Ensemble Regressor was built on a multi-modal feature set—including acoustic features, ASR error metrics, and linguistic features—to quantify the relative importance of these factors.

2. PROPOSED METHOD AND EXPERIMENT

Feature Extraction: Various features were extracted from input audio signals (unprocessed song (Audio 2) and processed song with a hearing loss simulation (Audio 1)). For intrusive prediction, the ground truth lyrics were also utilized.

1. **Hearing loss category**: A categorical variable of the simulated loss level (No Loss: 0, Mild: 1, Moderate: 2).
2. **Match error rate (MER)**: ASR-based metric calculated for both left and right channels by comparing ground truth lyrics with transcriptions generated by the Whisper model (`small.en`) [3] after vocal separation via a fine-tuned version of the hybrid transformer (HT) Demucs (`htdemucs_ft`) [4] (as shown in Fig. 1). The MER was calculated at both the word and phoneme levels, with phonetic transcriptions derived from the BEEP dictionary[3].
3. **Acoustic features**: Hardness, sharpness, and brightness were extracted using the AudioCommons model[4]. Vocal loudness was calculated using `pyloudnorm`[5] (ITU-R BS.1770-4 standard). Beat count features were extracted using Librosa's beat tracking functionality[6].
4. **Estimated mean opinion score (MOS)**: The predicted audio quality of Estimated Vocal 1 through P.808 MOS and MOS-OVR was obtained using DNSMOS [5].

This work was supported by JSPS KAKENHI Grant Numbers (20KK0233, 21H03463, 25H01139 and 25K21245).

[1] https://cadenzachallenge.org/docs/clip1/intro

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

[3] https://www.openslr.org/14/

[4] https://github.com/AudioCommons/timbral_models

[5] https://github.com/csteinmetz1/pyloudnorm

[6] https://librosa.org/doc/0.11.0/generated/librosa.beat.beat_track.html

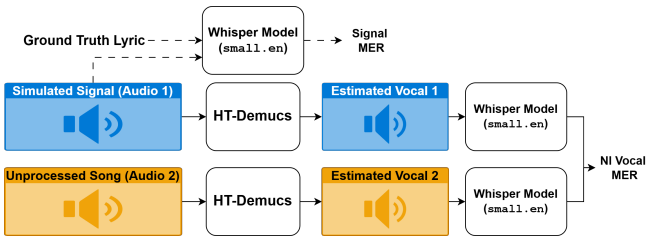

Fig. 1. MER calculation using Whisper.

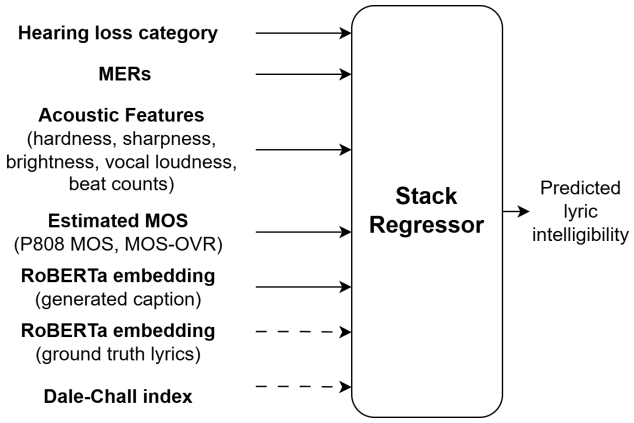

Fig. 2. Overview of proposed lyric intelligibility prediction methods. Solid lines represent feature pathways shared by both **Proposed #1** (non-intrusive) and **Proposed #2** (intrusive). Dashed lines indicate features or components absent in non-intrusive approach

5. **Text embeddings and index**: Semantic features were extracted using `RoBERTa-base` [6] from ground truth lyrics and captions generated by [7]. The Dale-Chall index, a readability score, was also calculated quantifying the inherent difficulty of ground truth lyrics [8].

Prediction Model: The overall prediction model, shown in Fig. 2, is a Stacked Ensemble Regressor[7] that optimally combines the predictions of three diverse base models. The Level 0 estimators include the Gradient Boosting Regressor (200 estimators, 0.05 learning rate, depth 4, 0.7 subsample), which corrects errors iteratively, the Random Forest Regressor (300 estimators, depth 10), which uses bagging to reduce variance, and Linear Regression, which provides a simple baseline for linear relationships. The meta-model then learns to blend these initial predictions for the final output.

The Stacking Regressor ($\mathbf{ereg}$) is defined as $\mathbf{ereg} = \text{StackingRegressor}(\text{estimators}, \text{final\_estimator}, \text{cv})$. Base estimator predictions were generated via 5-fold cross-validation ($\text{cv} = 5$) to ensure unbiased Level 1 features. A second Gradient Boosting Regressor was chosen as the Final Estimator (50 estimators, 0.1 learning rate, depth 3) and was trained on these cross-validated predictions to learn the optimal non-linear blending function. The resulting trained ensemble model, $\mathbf{ereg}$, produced the final prediction.

[7]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingRegressor.html

Table 1. Performance comparison on CLIP1 dataset. **Proposed #1**: non-intrusive (NI) metric uses Audio 1 (Simulated signal) and Audio 2 (Unprocessed song). **Proposed #2**: intrusive metric uses reference text (ground truth lyrics), Audio 1, and Audio 2.

Method	NI	Validation		Evaluation	
		RMSE (%)	ρ	RMSE (%)	ρ
STOI-based Baseline	Yes	36.11	0.14	34.89	0.21
Whisper-based Baseline	No	29.32	0.59	29.08	0.58
Proposed #1	Yes	28.70	0.62	28.34	0.61
Proposed #2	No	**26.76**	**0.68**	**27.22**	**0.65**
w/o acoustic features		27.11	0.67	27.31	0.65
w/o text emb. and index		27.27	0.66	27.51	0.64
w/o MOS		26.89	0.68	27.26	0.65

Evaluation The CLIP1 dataset [9] was divided into sets featuring 8,802 clips for training, 1,174 clips for validation, and 1,095 clips for evaluation. All subsets maintained a uniform balance of simulated signals across all defined hearing loss categories. For model development, the training set was initially split 90:10 for internal parameter optimization. Subsequently, model performance was evaluated using the Root Mean Squared Error (RMSE) and the Pearson correlation coefficient (ρ).

We evaluated two variations of the proposed method: **Proposed #1**, a non-intrusive metric that excludes ground-truth lyrics (utilizing only cues indicated by the solid arrows in Fig. 2), and **Proposed #2**, an intrusive metric leveraging all extracted cues. For comparison, we provide two baselines from CLIP1: the STOI-based Baseline and the Whisper-based Baseline.

Table 1 demonstrates that the proposed methods offer significant performance improvements over the Whisper-based baseline. To evaluate the contribution of each feature set, an ablation study was conducted on **Proposed #2**. This involved systematically removing feature groups, specifically, acoustic features (w/o acoustic features), text embeddings and the Dale-Chall index (w/o text emb. and index), and Mean Opinion Score metrics (w/o MOS). The results of this study confirm that all feature groups contribute significantly to the final predictive accuracy.

3. CONCLUSION

This study proposed a Stacked Ensemble Regressor for predicting lyric intelligibility, integrating a diverse feature set that includes acoustic, linguistic, and speech recognition error metrics. By utilizing a second-level Gradient Boosting Regressor to strategically combine three base models, the ensemble significantly outperformed the baseline Whisper model across both validation and evaluation sets. On the evaluation data, the model achieved a substantial reduction in RMSE from 29.08% to 27.22% and increased the ρ from 0.58 to 0.65. Ablation analysis confirmed that each feature group is necessary for optimal performance.

4. REFERENCES

[1] G. Roa-Dabike, J. P. Barker, et al., "Overview of the ICASSP 2026 Cadenza challenge: Predicting lyric intelligibility," in *Proc. IEEE ICASSP*, 2026. (To appear)

[2] B. Sharma and Y. Wang, "Automatic evaluation of song intelligibility using singing adapted STOI and vocal-specific features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 319–331, 2020, doi: 10.1109/TASLP.2019.2955253.

[3] A. Radford, J. W. Kim, et al., "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, vol. 202, pp. 28492–28518, Honolulu, Hawaii, USA, 2023, doi: 10.5555/3618408.3619590

[4] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proc. IEEE ICASSP*, pp. 1–5, Rhodes Island, Greece, 2023, doi: 10.1109/ICASSP49357.2023.10096956.

[5] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE ICASSP*, pp. 886–890, Singapore, 2022, doi: 10.1109/ICASSP43922.2022.9746108.

[6] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019, doi: 10.48550/arXiv.1907.11692.

[7] M. Kadlčík, A. Hájek, et al., "A Whisper transformer for audio captioning trained with synthetic captions and transfer learning," *arXiv:2305.09690*, 2023, doi: 10.48550/arXiv.2305.09690.

[8] E. Dale and J. S. Chall, "A formula for predicting readability," *Educ. Res. Bulletin, 27*, pp. 1–20, 1948.

[9] G. Roa-Dabike, et al., "The Cadenza Lyric Intelligibility Prediction (CLIP) dataset," *Data in Brief*, vol. 65:112466, 2026, doi: 10.1016/j.dib.2026.112466.