Proceedings Paper:
Roa-Dabike, G., Barker, J.P., Cox, T.J. et al. (2026) Overview of the ICASSP 2026 Cadenza Challenge: predicting lyric intelligibility. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 03-08 May 2026, Barcelona, Spain. Institute of Electrical and Electronics Engineers (IEEE), pp. 21757-21759. ISBN: 9798331567026. ISSN: 1520-6149.

https://doi.org/10.1109/icassp55912.2026.11463231

OVERVIEW OF THE ICASSP 2026 CADENZA CHALLENGE: PREDICTING LYRIC INTELLIGIBILITY

Gerardo Roa-Dabike[1] *Jon P. Barker*[1] *Trevor J. Cox*[2] *Michael A. Akeroyd*[3] *Scott Bannister*[4]
Bruno Fazenda[2] *Jennifer Firth*[3] *Simone Graetzer*[2] *Alinka Greasley*[4] *Rebecca R. Vos*[2] *William M. Whitmer*[3]

[1] University of Sheffield, [2] University of Salford, [3] University of Nottingham, [4] University of Leeds

ABSTRACT

We present the first open challenge on predicting lyric intelligibility. A new dataset, CLIP1, was introduced, comprising audio samples of popular western music paired with listener intelligibility scores. To model diverse listening profiles, samples were processed with no, mild and moderate simulated hearing loss. A total of 27 systems were submitted by 22 teams. Most systems used foundation models to extract encoder embeddings as high-level acoustic representations, often complemented by signal features and perceptual metrics. Twenty-five systems outperformed the STOI baseline, and 16 outperformed a Whisper-based baseline.

***Index Terms*—** hearing loss, machine learning, intelligibility, lyrics, music

1. INTRODUCTION

About 430 million people have disabling hearing loss, and this number is forecast to reach 700 million by 2050 [1]. Listening to music benefits health and well-being [2]. However, hearing loss can harm music perception, including the ability to clearly and effortlessly hear lyrics [3]. Since lyric understanding is part of music enjoyment [4], this could make people less likely to engage with music.

In spoken communication, intelligibility metrics have driven improvements in technologies such as speech enhancement and hearing aids. In contrast, metrics for lyric intelligibility are rare [5].

Applying speech intelligibility metrics directly to music is unreliable. Spoken and sung language differ in rhythm, intonation, and production characteristics. In popular music, vocals are often processed using reverberation, double tracking, and compression. Moreover, musical accompaniment differs fundamentally from the independent noise backgrounds assumed by speech intelligibility metrics. To address this gap, the ICASSP 2026 Cadenza challenge[1] introduces a benchmark task for predicting intelligibility while also considering diverse listener acuities.

2. CHALLENGE DESCRIPTION

The task was to predict lyric intelligibility from music excerpts that contained 5-10 words. To support model development, entrants were provided with hearing-loss simulated stereo audio, corresponding unprocessed stereo audio, ground-truth lyrics and intelligibility scores to develop their models (Figure 1).

A dedicated dataset was developed for the challenge [6], comprising 11,072 processed music excerpts from unfamiliar popular Western songs derived from the FMA dataset [7]. Each excerpt was paired with ground-truth lyric and intelligibility scores obtained from listening tests. Mild and moderate hearing loss were simulated via signal processing, alongside an unprocessed normal-hearing condition.

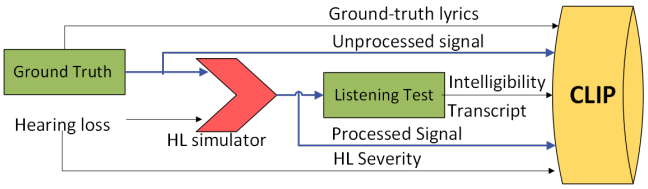

Fig. 1. Diagram of CLIP dataset.

Two baselines were defined. The first was based on the Short-Time Objective Intelligibility (STOI) model [8], a perceptual metric for speech intelligibility that compares a reference signal with a degraded one. In this baseline, vocals extracted from the unprocessed signal using the HDemucs source separation model [9] served as the reference, while the processed signal was used as the degraded input.

The second baseline used the Whisper ASR model ('base.en') [10] to transcribe the processed signal. Intelligibility was estimated as the proportion of correctly recognised words relative to the ground-truth lyrics.

For both baselines, a logistic function was used to map the raw predictions to the intelligibility scores in the training set.

3. SUBMISSIONS AND RESULTS

There were 27 submissions from 22 teams – see Table 1. Sixteen entries outperformed the Whisper baseline and 25 the STOI baseline.

Twenty-four systems used one or more foundation models to compute encoder embeddings as high-level acoustic representations. These included 21 systems using one or more versions of Whisper, 2 using WavLM [16], 1 using Wav2Vec 2.0 [17], 1 using HuBERT [18], and 1 using MERT [19]. Whisper was also used to generate multiple transcription hypotheses, which were encoded using T5 [20] (1 system) or RoBERTa [21] (3 systems). Sometimes the reference text was also encoded. For some entrants, the above features were complemented with signal-processing features such as MFCCs, SNR, and pitch, along with perceptual speech metrics including STOI, ESTOI [22], and PESQ [23].

Submitted systems made use one or more of the available metadata in different ways: eleven systems (see HL in table) incorporated the hearing-loss severity category either as one input or as part of a multitask learning setup; sixteen systems (GT in table) used the ground-truth lyrics for ASR-based approaches or to inform the models with the number of words; sixteen systems (US in table) used the

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

Table 1. Results of entrants (Txx) on the validation and evaluation sets. **Info Used** indicates the information used: **HL** = hearing-loss severity, **GT** = ground-truth lyrics, **US** = unprocessed signal. **Whisper** and **STOI** are the baselines. **RMSE** = root mean square error. **r** = Pearson correlation. Technical reports for the systems will be on cadenzachallenge.org

System	Info Used			Validation		Evaluation	
	HL	GT	US	RMSE	r	RMSE	r
T45[11]	✓	✓	–	25.89	0.70	26.44	0.67
T71a[12]	–	✓	✓	26.40	0.69	26.52	0.67
T13a[13]	–	✓	–	26.61	0.69	26.54	0.67
T17	–	–	–	26.82	0.68	26.67	0.67
T72a	✓	✓	✓	26.59	0.68	26.68	0.66
T72b	✓	–	✓	26.94	0.67	26.84	0.66
T40[14]	–	–	✓	27.13	0.67	27.07	0.65
T88a[15]	✓	✓	✓	26.76	0.68	27.22	0.65
T51	–	✓	✓	26.97	0.67	27.31	0.64
T86	✓	–	✓	27.43	0.66	27.31	0.64
T91	✓	✓	–	26.96	0.67	27.43	0.64
T19	–	✓	✓	26.70	0.68	28.24	0.63
T88b[15]	✓	–	✓	28.70	0.62	28.34	0.61
T15	–	✓	✓	27.60	0.65	28.46	0.61
T24	–	✓	✓	28.68	0.63	28.69	0.61
T63	✓	–	–	28.37	0.66	28.81	0.63
Whisper	–	✓	–	29.32	0.59	29.08	0.58
T84	✓	✓	–	27.71	0.66	29.17	0.59
T13b[13]	–	–	✓	29.75	0.59	29.68	0.57
T36	–	–	–	30.98	0.53	30.97	0.50
T71b[12]	–	–	✓	31.00	0.53	31.64	0.46
T44b	–	–	–	31.45	0.51	32.20	0.45
T54	✓	✓	–	31.44	0.50	32.30	0.43
T44a	–	–	✓	34.22	0.44	32.77	0.48
T16	✓	✓	✓	28.09	0.64	32.89	0.44
T94	–	✓	–	32.75	0.58	33.25	0.56
STOI	–	–	✓	36.11	0.14	34.89	0.21
T50	–	–	–	36.54	0.48	36.08	0.45
T73	–	✓	–	36.73	0.03	36.27	-0.04

unprocessed signal either as an input or to estimate the vocals; and four systems (T17, T36, T44b, T50) relied solely on the processed signal. Four teams (T13, T71, T72, T88) submitted two systems, one using the ground-truth lyric and one not. One team (T44) submitted two systems in which one uses the unprocessed signal and the other relies solely on the processed signal.

In general, systems that used the ground-truth lyrics performed better than those that did not. For teams submitting two systems, the version using the ground-truth lyrics was consistently better than the one without. In contrast, T17 ranked highly and it just used the processed signal.

A Kruskal–Wallis test indicated a statistically significant difference among all systems, although the effect size was small (df=28, N=31,755, H=828.33, p<0.01, η^2=0.03). To identify which systems differed, we conducted pairwise paired sign tests on per-excerpt absolute errors. For each system pair, we counted the number of excerpts on which one system yielded a lower absolute error than the other (ties excluded) and evaluated this count using a binomial test with p=0.5. T45 outperformed all other systems except T72a under this criterion, indicating that it more frequently achieved lower per-excerpt error, despite only a modest advantage in overall RMSE.

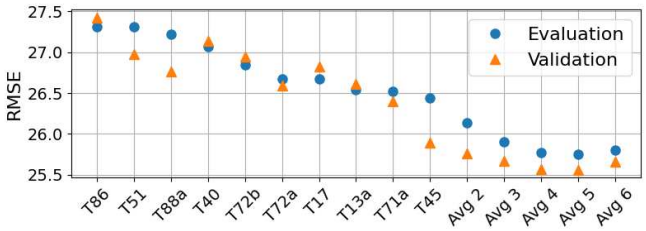

Fig. 2. Performance of the top-10 systems and progressively averaged systems (Avg 2 –Avg 6)

Most of the systems used just the CLIP dataset. However, four systems employed data augmentation. T24 and T54 used techniques such as pitch shifting, additive noise, or channel swapping. These may not be ideal for this task, however, as any deviation in the processed signal could alter the underlying intelligibility. In contrast, T17 and T91 utilized the CPC3 dataset [24] – a dataset for speech-in-noise intelligibility with ratings from people with hearing loss listening through hearing aids – for augmentation or pretraining. This approach may be more suitable, as it enriches the learning space (T17 showed it enabled a reduction of 0.1 RMSE). More ablation studies would be useful to assess the effectiveness of data augmentation.

Figure 2 shows the evaluation and validation RMSE performance of the top-10 systems. Most show similar performance for the evaluation and validation sets, with exceptions being T51, T88a, and T45. While this could indicate overfitting, confirming this would require more in-depth analysis. The figure also shows the results for progressively averaged systems from the top-2 to the top-6, (e.g. for each sample, the Avg2 prediction is the average of the predictions from T45 and T71a). Ensemble averaging yields modest improvements in predictive accuracy: for example, the Avg2 system yields an improvement of 0.31 RMSE, despite there only being a 0.08 difference in RMSE between T45 and T71a. However, this improvement is not statistically significant. The improvement peaks when averaging the top 5 systems, resulting in a statistically significant RMSE reduction of 0.69 compared with T45 alone.

4. DISCUSSION AND CONCLUSIONS

The high RMSE score for the best individual system and the ensemble average system indicates room for improvement. Significant sources of variance in the data include singing style and 'interfering' accompaniment. Larger datasets could help systems to model this variance better. Another source of variance is the scoring of intelligibility by listeners, further motivating the need for more data.

This challenge was the first large scale exploration of lyric intelligibility prediction methods. Consequently, there should be potential for improving system architectures and training processes. The large number of entrants shows there is an interest in the machine learning community for the area, and consequently a second lyric intelligibility prediction challenge is planned for 2026 by extending CLIP1 with AI generated songs.

5. ACKNOWLEDGEMENTS

Funding came from the Engineering and Physical Sciences Research Council EP/W019434/1. Our partners are: BBC, Google, Logitech, RNID, Sonova, Universität Oldenburg.

6. REFERENCES

[1] World Health Organization (WHO), "Deafness and hearing loss," 2023.

[2] Raymond MacDonald, Gunter Kreutz, and Laura Mitchell, *Music, health, and wellbeing*, Oxford University Press, 2013.

[3] 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.

[4] 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.

[5] 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, 2020.

[6] Gerardo Roa-Dabike, Trevor J. Cox, Bruno M. Fazenda, Simone Graetzer, Rebecca R. Vos, Michael A. Akeroyd, Jennifer Firth, William M. Whitmer, Scott Bannister, Alinka Greasley, and Jon P. Barker, "The Cadenza Lyric Intelligibility Prediction (CLIP) Dataset," *Data in Brief*, Under review.

[7] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "Fma: A dataset for music analysis," in *ISMIR 2017*, 2017.

[8] 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.

[9] Alexandre Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[10] 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*, 2023.

[11] Longbin Jin, "Stereo Chorus of Whisper: Perceptually-augmented ear-specific intelligibility prediction," in *ICASSP*, 2026.

[12] Rahul Peter, Vivek Mohan, Seralathan Subramanian, and Lauri Juvela, "Cadenza Challenge: Intrusive and non-intrusive lyric intelligibility via hybrid attention over ASR embeddings and feature blocks," in *ICASSP*, 2026.

[13] Jing Yang, Shenghao Liao, Shuqing Zhang, Yongyi Deng, and Pan Li, "Music lyric clarity assesment via multi-encoder fusion: from invasive to non-invasive Scenarios," in *ICASSP*, 2026.

[14] Ram C.M.C. Shekar and Ivan Lopez-Espejo, "Liwhiz: a non-intrusive lyric intelligibility prediction system for the Cadenza Challenge," in *ICASSP*, 2026.

[15] Candy Olivia Mawalim, Xiajie Zhou, and Masashi Unoki, "Multi-modal feature fusion and atacking ensemble learning for lyric intelligibility prediction," in *ICASSP*, 2026.

[16] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Takuya Yoshioka, Long Zhou, Shuo-Wei Li, Yao Qian, and Furu Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2022.

[17] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[19] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu, "Mert: Acoustic music understanding model with large-scale self-supervised training," in *12th International Conference on Learning Representations (ICLR)*, 2024.

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[22] Jesper Jensen and Cees H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[23] 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)*, 2001.

[24] J. Barker, M. A. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 3rd clarity prediction challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. of the 6th Clarity Workshop on Improving Speech-in-Noise for Hearing Devices (Clarity-2025)*, 2025.