STEREO CHORUS OF WHISPERS: PERCEPTUALLY-AUGMENTED EAR-SPECIFIC INTELLIGIBILITY PREDICTION

Longbin Jin[1]*, Heungsoo Kim*[1]*, Youngmin Kim*[1]*, Eun Yi Kim*[2]

[1]R&D Center, Citech Co., Ltd., Seoul, Republic of Korea
[2]AI & Computer Vision Lab., Konkuk University, Seoul, Republic of Korea

ABSTRACT

This paper presents *Stereo Chorus of Whispers*, the winning system of the Cadenza Lyric Intelligibility Prediction (CLIP) Challenge. Our approach extends a Whisper-based intelligibility prediction framework to stereo listening conditions by performing ear-specific Whisper decomposition using a heterogeneous ensemble of pretrained Whisper models. This design allows diverse ASR-based perceptual cues to be captured independently for each ear. The resulting representations are further enriched with complementary perceptual metrics, including STOI, ESTOI, PESQ, and DNSMOS, together with a hearing-level indicator reflecting the simulated listening severity. These multi-cue features are fused using a lightweight transformer-based regression model to estimate the final lyric intelligibility score. Experimental results show that the proposed system achieves RMSEs of 25.89 and 26.44 on the official validation and evaluation sets, respectively, and ranks first in the CLIP Challenge, demonstrating strong robustness for lyric intelligibility prediction in complex musical environments. Our code is available at https://github.com/jinlongbin/Stereo-Chorus-of-Whispers.

Index Terms— Binaural intelligibility estimation, Whisper-based modeling, perceptual feature fusion, lyric intelligibility

1. INTRODUCTION

Understanding sung lyrics is essential for music perception and enjoyment, as listeners rely on clearly perceived words to grasp the semantic and emotional content of a song [1]. Lyric intelligibility describes how clearly and effortlessly the words in music can be heard and understood. However, predicting it directly from audio remains challenging due to large variations in vocal style, strong interference from musical accompaniment, and diverse listener hearing abilities. The Cadenza Lyric Intelligibility Prediction (CLIP) Challenge [2] provides a standardized benchmark for this problem by requiring systems to estimate lyric intelligibility from complex music mixtures. Its dataset [3] includes not only typical music recordings but also versions processed through a hearing-loss simulator, exposing models to a wide range of perceptual listening conditions. Unlike conventional speech-focused scenarios, singing introduces pronounced pitch dynamics and frequent spectral overlap with instrumentation, which often degrade the reliability of traditional speech intelligibility measures and highlight the need for robust modeling approaches that can integrate multiple cues and generalize across diverse musical and hearing-profile variations.

Our previous *A Chorus of Whispers* framework [4] demonstrated that combining multiple Whisper models enables the capture of diverse listening behaviors, as each model emphasizes different lexical and acoustic cues with varying reliability. This model diversity becomes even more critical in the CLIP task, where sung vocals interact strongly with musical accompaniment and signals processed by a hearing loss simulator introduce additional perceptual variability. These observations motivate a stereo extension of the original method that integrates heterogeneous Whisper models, ear-specific processing, and complementary perceptual cues to better reflect how listeners with diverse hearing abilities derive intelligible content from complex musical mixtures.

Building on this insight, we propose *Stereo Chorus of Whispers*, a stereo extension of Whisper-based intelligibility modeling that explicitly incorporates binaural and perceptual cues. Our method begins with ear-specific Whisper [5] decomposition, in which multiple Whisper models of different capacities independently process the left and right channels. To complement these ASR-derived cues, we extract perceptual metrics including STOI [6], ESTOI [7], PESQ [8], and DNSMOS [9]. These multi-cue features are fused using a lightweight transformer-based regression model conditioned on a hearing-level token representing the simulated listener profile, enabling effective integration of model diversity, binaural differences, and perceptual information for robust lyric intelligibility estimation. The proposed approach attains the best overall performance in the CLIP Challenge, with RMSE values of 25.89 on the validation set and 26.44 on the evaluation set.

2. STEREO CHORUS OF WHISPERS

2.1. Whisper & Perceptual Feature Extraction

Our method is based on the observation that different ASR models exhibit distinct perceptual behaviors when processing the same audio signal. Following our previous *A Chorus of Whispers* framework [4], we employ a heterogeneous ensemble of 12 pretrained Whisper variants [5], including *large-v3-turbo*, *large-v3*, *large-v2*, *large*, *medium*, *medium.en*, *small*, *small.en*, *base*, *base.en*, *tiny*, and *tiny.en*[1]. All models are kept frozen and provide token predictions along with posterior probabilities, which serve as ASR-derived intelligibility cues. Unlike the previous monaural formulation, the proposed stereo extension processes the left and right channels independently. Each Whisper model operates separately on both channels, enabling the system to capture ear-specific cues such as masking differences, vocal clarity asymmetries, and other binaural effects.

To enrich these ASR-derived features, we incorporate complementary perceptual metrics including *STOI* [6], *ESTOI* [7], *PESQ* [8], and *DNSMOS* [9]. DNSMOS is a non-intrusive neural predictor of perceived speech quality, and we utilize its three outputs: speech quality (*SIG*), background quality (*BAK*), and overall quality (*OVRL*). Since the ground-truth intelligibility is defined as the percentage of correctly recognized words, we additionally append the

[1] https://huggingface.co/collections/openai/whisper-release

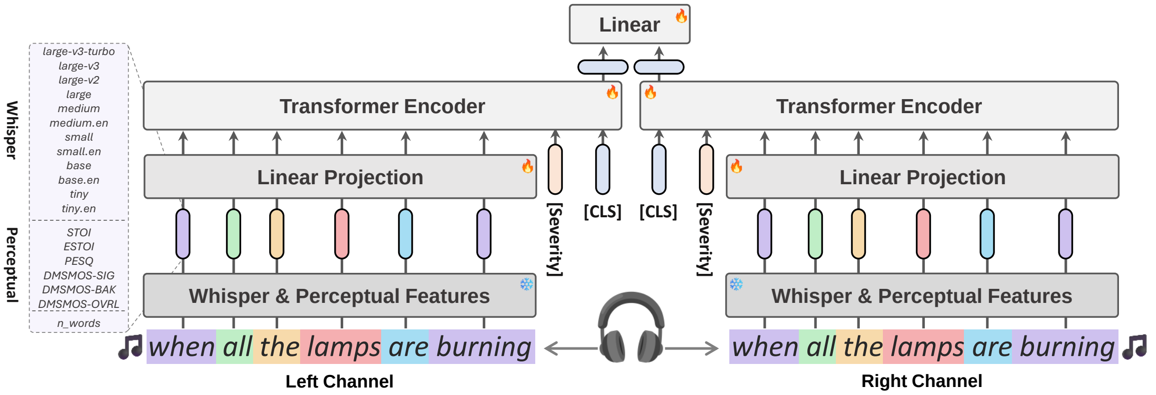

Fig. 1: Overall architecture of the Stereo Chorus of Whispers framework.

number of words (*n_words*) to account for varying lengths across lyric segment. These perceptual scores and lexical statistics are concatenated with the 12 Whisper-derived posterior features for each reference token, resulting in a token-level feature vector of 19 dimensions. Accordingly, for a lyric segment containing n words, the final input feature matrices for both the left and right channels have a dimensionality of $n \times 19$.

2.2. Lyrics Intelligibility Modeling

The Whisper-derived and perceptual features for each lyric-aligned audio segment are first projected into a 64-dimensional latent space through a linear embedding layer. To condition the model on hearing-loss severity, we prepend a learnable severity token corresponding to the simulated listener profile (no loss, mild, or moderate), which is shared across the left and right channels. In addition, a learnable class (CLS) token is introduced following a ViT-style formulation [10] to aggregate sequence-level information. A single transformer encoder layer with 4 attention heads is then applied to fuse the embedded token sequence. For stereo processing, the left and right channels are encoded independently without weight sharing, producing two channel-specific CLS embeddings. These embeddings are concatenated and passed through a single linear regression layer to predict the overall lyric intelligibility score in the range $[0, 1]$.

3. EXPERIMENTS

3.1. Implementation Details

The provided training set is divided into training and validation subsets with an 8:2 ratio. To enhance generalizability, the split is constructed such that reference texts do not overlap between the two subsets, ensuring evaluation on entirely unseen lyric content and preventing prompt-level memorization.

All models are trained for 100 epochs using a combined loss consisting of mean squared error (MSE) and negative correlation loss with a weighting factor of 0.4. The Adam optimizer is adopted with a learning rate of 1×10^{-4}, and training is performed on a single NVIDIA GeForce RTX 5080 GPU. We report performance on the official validation and evaluation sets.

Table 1: Progressive ablation study on the CLIP validation and evaluation datasets with incrementally added model components.

Component	**Validation set**		**Evaluation set**	
	Corr. ↑	**RMSE ↓**	**Corr. ↑**	**RMSE ↓**
Chorus of Whispers [4]	0.6697	27.2311	0.6556	27.6874
+ Stereo	0.6879	26.4877	0.6605	26.8566
+ Perceptual Metrics	0.6975	26.1780	0.6684	26.7244
+ Number of Words	0.6967	26.1443	0.6689	26.6562
+ Correlation Loss	**0.7014**	**25.8867**	**0.6712**	**26.4404**

3.2. Ablation Results on the CLIP Benchmark

The ablation results in Table 1 show that each component contributes complementarily to lyric intelligibility prediction. Stereo processing provides a clear benefit by enabling the model to exploit ear-specific cues such as channel-dependent masking and vocal clarity. The addition of perceptual metrics further improves performance by supplementing ASR-derived features with signal-level indicators of speech quality and distortion. Moreover, incorporating the number of words stabilizes prediction across lyric segments with varying lengths by normalizing differences in the intelligibility denominator. Finally, correlation-based loss optimization enhances the alignment between predicted scores and perceptual targets, indicating that explicitly modeling relative intelligibility trends improves robustness beyond MSE-based training alone.

4. CONCLUSION

This paper presents Stereo Chorus of Whispers, a stereo lyric intelligibility prediction framework that integrates heterogeneous Whisper ensembles, perceptual speech quality metrics, and hearing-level conditioning. By explicitly modeling binaural cues and fusing ASR-derived and perceptual features, the proposed system effectively captures both linguistic and signal-level factors that govern intelligibility in complex musical mixtures. Experiments on the CLIP benchmark demonstrate that the proposed approach achieves state-of-the-art performance and ranks first in the challenge. These results highlight the effectiveness of stereo modeling and multi-cue fusion for robust lyric intelligibility prediction in intrusive evaluation settings. We hope that the proposed framework will generalize to broader scenarios, including multilingual lyrics and diverse musical conditions, a direction we leave for future exploration.

5. REFERENCES

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

[2] 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.

[3] 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.

[4] Longbin Jin, Donghun Min, and Eun Yi Kim, “A chorus of whispers: Modeling speech intelligibility via heterogeneous whisper decomposition,” in *Proceedings of the 6th Clarity Workshop on Improving Speech-in-Noise for Hearing Devices (Clarity-2025)*, 2025, pp. 34–35.

[5] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” *arXiv preprint arXiv:2212.04356*, 2022.

[6] 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.

[7] 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.

[8] Antony W Rix, John G Beerends, Michael P Hollier, and Andries 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)*. IEEE, 2001, vol. 2, pp. 749–752.

[9] Chandan K A Reddy, Vishak Gopal, and Ross Cutler, “Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” in *ICASSP 2022*, 2022, pp. 886–890.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ICLR*, 2021.