

THE SCALE SYSTEM FOR MIREX 2026 MUSIC STRUCTURE ANALYSIS

Hongbing Li

Queen Mary University of London, London, UK
hongbingli96@gmail.com

Yaolong Ju

Great Bay University, Dongguan, China
juyaolong@gbu.edu.cn

ABSTRACT

This paper describes SCALE (Structural Context from Audio and Lyric Evidence), our submission to the MIREX 2026 Music Structure Analysis Task. SCALE separates boundary detection and functional labeling into two dedicated pathways, reflecting the different information required by the two tasks. In the function pathway, lyric information is further incorporated to provide complementary evidence for identifying chorus sections. This design leads to a substantial improvement in boundary detection while maintaining strong functional labeling performance.

1. INTRODUCTION

Music structure analysis (MSA) aims to divide a music recording into contiguous, non-overlapping sections and assign a functional label to each section [1]. It involves two closely related tasks: boundary detection, which identifies the start and end of structural sections, and functional labeling, which assigns labels such as intro, verse, chorus, and bridge. Recent systems such as All-In-One [2] and SongFormer [3] jointly address these tasks using attention-based architectures and pretrained music representations.

However, boundary detection and functional labeling rely on different types of information. Boundary detection depends strongly on local musical changes and precise timing, whereas functional labeling benefits more from longer-range structural context [4]. In addition, most end-to-end MSA systems remain audio-only, although lyrics provide useful structural cues: verses tend to introduce new text, while choruses often contain repeated or similar lyrics [5].

In this submission, we propose SCALE (Structural Context from Audio and Lyric Evidence), which separates boundary detection and functional labeling into two dedicated pathways so that each task can learn representations suited to its information requirements. SCALE further incorporates lyric information into the function pathway, using lyrical recurrence as complementary evidence for chorus prediction. This design substantially improves boundary detection while maintaining strong functional labeling performance.

2. METHODS

2.1 Data

HarmonixSet. We use the official HarmonixSet [6] splits released for the MIREX 2026 Music Structure Analysis task. The training and validation splits are used for model development, and the held-out test split is used for final evaluation.

SheetSage-MSA. We additionally use SheetSage-MSA, constructed from the function labels and boundary annotations released with SheetSage [7]. Following the SheetSage-A2S data acquisition procedure [8], we retrieve the corresponding recordings from the metadata and group multiple annotated excerpts belonging to the same song. Excerpt boundaries are updated using the refined beat-to-time alignments when available, after which unsupported, composite, and duplicate function annotations are removed. The retained annotations are then placed on the timeline of the corresponding full-length recording, producing partially labeled song-level examples in which unannotated regions are masked during training. The resulting dataset contains 13,880 cleaned excerpt annotations from 9,010 songs, totaling approximately 82.5 hours of annotated audio.

2.2 Model Architecture

Figure 1 shows the overall architecture of SCALE, which consists of three paths: boundary, function, and lyrics. The boundary and function paths use local and global representations extracted from a frozen MuQ encoder [9], while the lyrics path provides complementary evidence for chorus prediction.

Boundary path. The local and global MuQ representations are temporally aligned, fused with a learned gate, and downsampled. Four BiMamba blocks [10] model local temporal dependencies, after which pooled features are processed by a Transformer [11] to capture longer-range structural context. Cross-attention then combines this global context with the frame-level representation, and a boundary head predicts boundary probabilities.

Function path. The aligned local and global MuQ representations are concatenated and fed to a separate function pathway. After downsampling, a four-layer Transformer with rotary positional embeddings (RoPE) [12] predicts frame-level structural labels. This dedicated pathway allows functional labeling to focus on longer-range section ordering rather than sharing the boundary encoder.



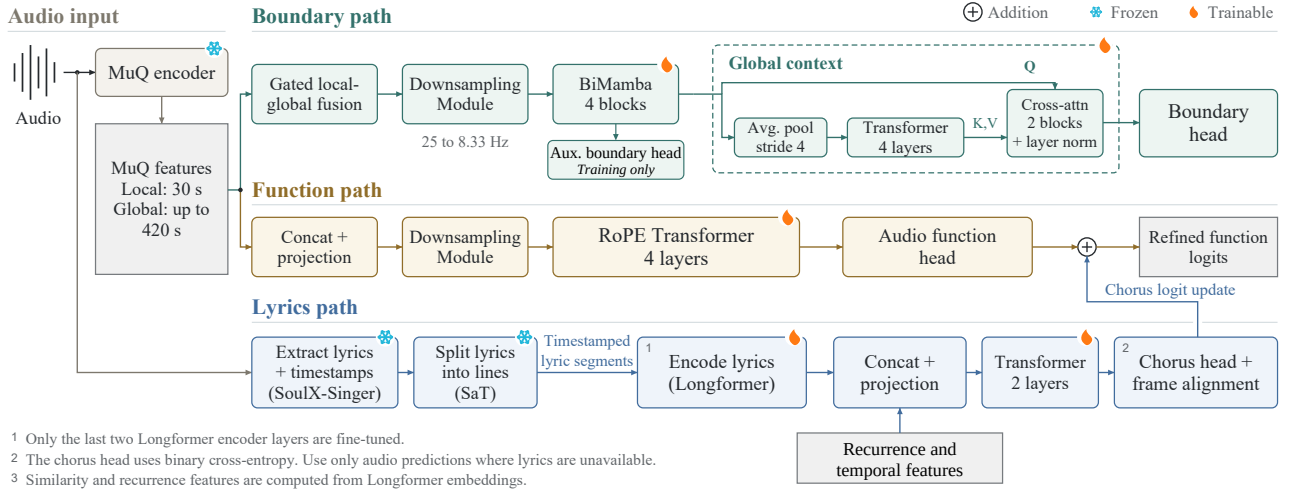


Figure 1. Overview of SCALE. The model contains three paths: a boundary path, a function path, and a lyrics path. The first two paths use MuQ audio features, while the lyrics path provides additional evidence for chorus prediction.

Lyrics path. Timestamped lyrics are extracted offline using the SoulX-Singer pipeline [13]. Each timestamped lyric segment is further divided into lines using SaT [14]. Longformer [15] encodes the lyric content, and recurrence and temporal features are used to predict a chorus score for each segment. The segment-level chorus scores are aligned to the audio timeline and added only to the chorus logit from the function path, thereby refining chorus prediction without affecting boundary detection.

2.3 Configuration

MuQ features are extracted at 25 Hz and downsampled to approximately 8.33 Hz. The boundary representation has dimension 256 and the function Transformer has hidden dimension 512; all Transformer modules use eight attention heads and 0.10 dropout. MuQ remains frozen, while the last two Longformer layers are fine-tuned. We optimize with AdamW [16] using learning rate 1.5×10^{-4} for the main network, 10^{-5} for Longformer, and weight decay 10^{-4} . A cosine schedule follows 600 warm-up steps. Gradient accumulation gives an effective batch size of eight and gradients are clipped to norm 1.0. In mixed-data training, each HarmonixSet training example is repeated four times to increase its contribution relative to the larger SheetSage-MSA set. At inference time, boundary peaks use threshold 0.35 and a minimum spacing of 2 s. Function logits are averaged within each predicted segment and the highest-scoring allowed label is emitted.

3. RESULTS

The official MSA metrics on the MIREX HarmonixSet test split are presented in Table 1. We report frame-level functional-label accuracy (ACC) and boundary F-measures with 0.5-s and 3-s tolerances (HR.5F and HR3F) [17], together with the systems reported in MIREX 2025 for comparison.

Table 1. Results on the MIREX HarmonixSet test split.

System	ACC	HR.5F	HR3F
MusicFM (MIREX 2025) [18]	0.705	0.644	0.710
KGStruct (MIREX 2025) [19]	0.720	0.590	0.762
SCALE	0.803	0.731	0.828

SCALE achieves higher scores than the previous MIREX systems across all three evaluation metrics.

4. CONCLUSION

This paper presents SCALE, our submission to the MIREX 2026 Music Structure Analysis task. SCALE separates boundary detection and functional labeling into task-specific pathways and incorporates lyrical recurrence as additional evidence for chorus prediction. Together with the additional supervision from SheetSage-MSA, the proposed system improves both functional labeling and boundary detection, with particularly strong gains on the boundary metrics.

5. REFERENCES

- [1] O. Nieto, G. J. Mysore, C.-i. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee, “Audio-based music structure analysis: Current trends, open challenges, and applications,” *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 246–263, 2020.
- [2] T. Kim and J. Nam, “All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.
- [3] C. Hao, R. Yuan, J. Yao, Q. Deng, X. Bai, W. Xue, and L. Xie, “SongFormer: Scaling music structure analysis

- with heterogeneous supervision,” arXiv:2510.02797, 2025.
- [4] J.-C. Wang, Y.-N. Hung, and J. B. L. Smith, “To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 416–420.
 - [5] R. Bao, H. Ma, S. Liu, C. Gong, C. Zhang, X.-L. Zhang, W.-Q. Zhang, and X. Li, “ALMA-Chor: Leveraging audio-lyric alignment with Mamba for chorus detection,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 16102–16106.
 - [6] O. Nieto, M. McCallum, M. E. P. Davies, A. Robertson, A. Stark, and E. Egozy, “The Harmonix Set: Beats, downbeats, and functional segment annotations of western popular music,” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 565–572.
 - [7] C. Donahue, J. Thickstun, and P. Liang, “Melody transcription via generative pre-training,” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 485–492.
 - [8] E. Cummins, Z. Huang, A. D’Hooge, Z. Mo, and Y. Ju, “Audio-to-score transcription using pre-trained features, data augmentation, and the new SheetSage-A2S dataset,” in *Proc. ACM International Conference on Multimedia (MM)*, 2026, to appear, doi: 10.1145/3767308.3835653.
 - [9] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, “MuQ: Self-supervised music representation learning with mel residual vector quantization,” arXiv:2501.01108, 2025.
 - [10] A. Gu and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces,” in *Proc. Conference on Language Modeling (COLM)*, 2024.
 - [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
 - [12] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, “RoFormer: Enhanced Transformer with rotary position embedding,” arXiv:2104.09864, 2021.
 - [13] J. Qian *et al.*, “SoulX-Singer: Towards high-quality zero-shot singing voice synthesis,” arXiv:2602.07803, 2026.
 - [14] M. Frohmann, I. Sterner, I. Vulić, B. Minixhofer, and M. Schedl, “Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation,” in *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 11908–11941.
 - [15] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document Transformer,” arXiv:2004.05150, 2020.
 - [16] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. International Conference on Learning Representations (ICLR)*, 2019.
 - [17] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “mir_eval: A transparent implementation of common MIR metrics,” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 367–372.
 - [18] M. Won, Y.-N. Hung, and D. Le, “A foundation model for music informatics,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1226–1230.
 - [19] C. Xiao, J. Wan, and C. Chen, “The 2025 KG Music Structure Analysis System,” MIREX 2025 Music Structure Analysis extended abstract, 2025.