

STRUCTURED LONG-FORM LYRICS TRANSCRIPTION - MIREX 2026

Maria Frangiadaki¹, Konstantinos Klapsas¹, Kosmas Kritsis¹
Charis Papaioannou¹, Georgios Paraskevopoulos¹, Vassilis Katsouros¹

¹Institute for Language and Speech Processing, Athena Research Center, Greece

{maria.frangiadaki, k.klapsas, kosmas.kritsis}@athenarc.gr

{haris.papaioannou, g.paraskevopoulos, vsk}@athenarc.gr

ABSTRACT

Automatic lyrics transcription (ALT) from polyphonic music requires not only accurate recognition of sung words, but also the recovery of the structural conventions of written lyrics. We present our submission to the MIREX 2026 Lyrics Transcription task, a long-form inference pipeline that combines automatic speech recognition, music-aware audio understanding, and refinement with a large language model (LLM). Source separation is used to support voice activity detection and long-form segmentation, while lexical transcription is performed on the original polyphonic mixture. We exploit internal segmentation and timing information produced during Whisper decoding to construct a structured full-song transcription and recover lyric formatting. Two additional full-song transcriptions, produced by Music Flamingo and Parakeet, provide complementary evidence. Finally, Kimi K3 processes the three transcriptions to produce the submitted lyric sheet. The system requires no task-specific acoustic model training and simultaneously targets lexical accuracy and the punctuation, casing, parentheses, line breaks, and section boundaries considered in the MIREX evaluation.

1. INTRODUCTION

Automatic lyrics transcription (ALT) aims to transcribe the content of sung vocals directly from music recordings. Compared with conventional automatic speech recognition, ALT must operate under instrumental interference, highly variable vocal production, and musical structures that extend over several minutes. Long-form transcription further requires the system to maintain consistency across repeated verses and choruses while avoiding errors introduced by independently processing short audio segments.

The MIREX 2026 Lyrics Transcription task additionally evaluates properties that are not captured by word error rate (WER) alone. Besides lexical accuracy, a system must reproduce the written organization of the lyrics, including casing, punctuation, parentheses, line breaks, and section boundaries. Thus the final lyric transcription should be represented as a structured lyric sheet.

Our submission treats lexical transcription and lyric-structure recovery as complementary stages rather than delegating both to a single model. The system combines Whisper large-v3 (1.55B) [1], Music Flamingo (8B) [2], Parakeet-TDT-0.6B-v3 (0.6B) [3], and Kimi K3 (2.8T total parameters; 104B activated) [4] to jointly address lexical accuracy and full-song lyric formatting.

2. SYSTEM DESCRIPTION

2.1 Source Separation and Long-form Transcription

We use the long-form inference framework released with MSS-ALT [5] as the transcription backbone of our system. Its implementation is included as a third-party component of the submission¹. Following MSS-ALT, music source separation is used to obtain a vocal estimate for voice-activity-based segmentation, while transcription itself is performed on the original polyphonic mixture.

We use the Demucs mdx_extra separator [6]. RMS-based vocal activity detection is applied to the separated vocal signal at 16 kHz, and detected regions are merged into chunks of at most 30 s. The resulting fixed temporal regions are subsequently applied to the original mixture and reused throughout the Whisper stages.

Before transcription, we determine one language label per song using Whisper. For songs containing at least three RMS-VAD chunks, we select chunks longer than 5 seconds from the beginning, middle, and end of the song. Whisper performs automatic language identification on the original polyphonic audio. The language prediction with the highest confidence is selected and fixed for all subsequent Whisper decoding of that song.

2.2 Two-pass Whisper Transcription and Lyric Formatting

Whisper large-v3 is run twice over the same fixed RMS-VAD regions. In the first pass, we decode the original polyphonic audio using beam size 10 and the fixed song-level language label. Word-level timestamps are disabled, and the concatenated output is retained as the canonical lexical transcription. This sequence is kept immutable in all subsequent formatting stages.

The second pass uses the same chunk boundaries, language label, and decoding settings, but enables word-level



¹ <https://github.com/jaza-syed/mss-alt>

timestamps. In addition to word timings, we retain the internal segment boundaries generated by Whisper. In our experiments, timestamp-aware decoding slightly degraded the lexical accuracy compared with the first pass. Thus, we use the second pass only to obtain timing and segmentation cues. The first- and second-pass word sequences are globally aligned using Levenshtein alignment and word timings, Whisper internal-segment boundaries, and RMS-VAD chunk boundaries are projected on the canonical first-pass sequence.

Lyric-line boundaries are reconstructed from the union of three cues; projected Whisper internal-segment boundaries, RMS-VAD chunk boundaries, and capitalization cues. For English, Spanish, and French, an uppercase initial token is additionally treated as a potential line-break cue; the standalone English pronoun “I” is excluded from this heuristic. The capitalization heuristic is disabled for German. Boundaries immediately preceding attached punctuation are shifted after the punctuation token. Formatting is applied exclusively by replacing inter-token whitespace with line breaks, ensuring that the canonical lexical sequence is unchanged.

2.3 Complementary Full-Song Transcriptions

We additionally process each original full-song waveform with Music Flamingo [2]. The model receives the complete polyphonic song and is prompted to generate a structured lyric sheet. Its output serves both as an independent full-song transcription candidate and as a source of higher-level structural information.

We observed that Music Flamingo is able to generate plausible candidates for the structure of the song, but suffers from frequent hallucinations. We thus transfer the structural cues from Music Flamingo to the formatted Whisper transcription. The two transcriptions are token-aligned using SequenceMatcher-based alignment, after which Music Flamingo’s parentheses predictions spans are projected on the corresponding Whisper word spans. Additionally, blank-line section boundaries from Music Flamingo are mapped to the nearest existing Whisper line boundary as a section break. Thus, Music Flamingo may introduce parentheses and section divisions, but cannot alter the Whisper lexical sequence or create new lyric-line positions.

Finally, we use Parakeet-TDT-0.6b-v3 [3] as a lexical transcription candidate to resolve disagreements between the two other transcriptions, ie as a tie-breaker. All three transcriptions are subsequently retained in full-song form for the final fusion stage.

2.4 LLM-Based Full-Song Fusion

For the final stage, we follow the general idea of LLM-assisted hypothesis refinement used in LyricWhiz [7], but applying it to three heterogeneous full-song candidates. We provide Kimi-K3 with the structured Whisper transcription, the Music Flamingo transcription, and the Parakeet transcription. The LLM is instructed to compare the candidates line by line, favor lexical content supported

Lang.	N	WER	WER _{case}	Punc. F_1	Paren. F_1	Line F_1	Sect. F_1
EN	160	20.92	36.32	26.01	3.70	77.97	62.74
DE	19	34.06	52.80	33.95	–	70.23	50.23
ES	9	18.92	28.51	9.86	–	63.72	62.81
FR	7	28.82	39.76	35.20	–	73.38	42.74
All	195	22.37	37.68	27.17	2.95	76.55	60.84

Table 1. Final Kimi K3 fusion results (%) on our 195-song held-out DALI test set.

System	WER	WER _{case}	Punc. F_1	Paren. F_1	Line F_1	Sect. F_1
Whisper (first pass)	27.36	40.99	26.64	–	–	–
Structured Whisper	27.36	40.99	26.25	0.00	74.19	34.43
Music Flamingo	60.15	71.37	18.49	1.65	49.85	38.07
Parakeet	44.68	53.79	15.40	–	–	–
Kimi K3 fusion	22.37	37.68	27.17	2.95	76.55	60.84

Table 2. Intermediate transcription outputs and final fusion results (%) on our 195-song held-out DALI test set.

by multiple systems while also considering grammatical and semantic coherence and lyric rhyme or meter. It is explicitly discouraged from introducing unsupported content and is asked to output only the final lyrics, preserving blank lines for section boundaries and parentheses for likely backing vocals. Fusion is performed once per complete song, allowing long-range repetitions and song-level context to contribute to the final decision.

3. EVALUATION

Table 1 reports the final post-fusion results overall and by language. We evaluate the final post-fusion system on our held-out DALI test set, comprising 195 songs. This split was originally created at the song level during data preparation for planned task-specific fine-tuning experiments, and was retained as the held-out evaluation partition for the final inference-only system. The resulting test set is imbalanced across the four target languages, containing 160 English, 19 German, 9 Spanish, and 7 French songs.

The final system achieves an overall WER of 22.37%, while recovering lyric-line and section structure with F1 scores of 76.55% and 60.84%, respectively. Because the test split is strongly English-dominated, aggregate results should be interpreted together with the language-specific scores.

As shown in Table 2, structured Whisper preserves the first-pass lexical WER, while introducing explicit lyric-line and section structure. Music Flamingo and Parakeet score substantially higher standalone WERs and are therefore used as complementary sources rather than primary lexical hypotheses. The final Kimi K3 fusion reduces WER and further improves line and section F1.

4. RESOURCES DECLARATION

No task-specific model training was performed. End-to-end inference on the task’s known test sets (7.49 h of audio) on an NVIDIA GeForce RTX 3090 (24 GB) required approximately 9 h, corresponding to 1.20 h of inference time per hour of audio.

5. REFERENCES

- 175 [1] A. Radford, J. W. Kim, T. Xu, G. Brockman,
 176 C. McLeavey, and I. Sutskever, “Robust speech
 177 recognition via large-scale weak supervision,” in
 178 *International Conference on Machine Learning, ICML*
 179 *2023, 23-29 July 2023, Honolulu, Hawaii, USA*,
 180 ser. Proceedings of Machine Learning Research,
 181 A. Krause, E. Brunskill, K. Cho, B. Engelhardt,
 182 S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR,
 183 2023, pp. 28 492–28 518. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
 184
- 185 [2] S. Ghosh, A. Goel, L. Koroshinadze, S. Lee, Z. Kong,
 186 J. F. Santos, R. Duraiswami, D. Manocha, W. Ping,
 187 M. Shoenybi, and B. Catanzaro, “Music flamingo:
 188 Scaling music understanding in audio language
 189 models,” in *Proc. Int. Conf. Learning Represen-*
 190 *tations (ICLR)*, 2026. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2026/hash/7f64034009f4a5fa417a57e1a987c5cd-Abstract-Conference.html
 191
 192
 193
- 194 [3] M. Sekoyan, N. R. Koluguri, N. Tadevosyan,
 195 P. Zelasko, T. Bartley, N. Karpov, J. Balam, and
 196 B. Ginsburg, “Canary-1b-v2 and parakeet-tdt-0.6b-
 197 v3: Efficient and high-performance models for
 198 multilingual asr and ast,” 2025. [Online]. Available:
 199 <https://arxiv.org/abs/2509.14128>
- 200 [4] Kimi Team, “Kimi K3: open frontier intelligence,”
 201 2026. [Online]. Available: <https://arxiv.org/abs/2607.24653>
 202
- 203 [5] J. Syed, I. M. Higgs, O. Cifka, and M. Sandler,
 204 “Exploiting music source separation for automatic
 205 lyrics transcription with Whisper,” in *Proc. IEEE Int.*
 206 *Conf. Multimedia and Expo Workshops (ICMEW)*.
 207 IEEE, 2025, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/ICMEW68306.2025.11152264>
 208
- 209 [6] A. Défossez, “Hybrid spectrogram and waveform
 210 source separation,” *arXiv preprint arXiv:2111.03600*,
 211 2021. [Online]. Available: <https://arxiv.org/abs/2111.03600>
 212
- 213 [7] L. Zhuo, R. Yuan, J. Pan, Y. Ma, Y. Li, G. Zhang,
 214 S. Liu, R. B. Dannenberg, J. Fu, C. Lin, E. Benetos,
 215 W. Chen, W. Xue, and Y. Guo, “Lyricwhiz:
 216 Robust multilingual zero-shot lyrics transcription by
 217 whispering to chatgpt,” in *Proceedings of the 24th*
 218 *International Society for Music Information Retrieval*
 219 *Conference, ISMIR 2023, Milan, Italy, November*
 220 *5-9, 2023, 2023*, pp. 343–351. [Online]. Available:
 221 <https://doi.org/10.5281/zenodo.10265295>