STRUCTURED LONG-FORM LYRICS TRANSCRIPTION - MIREX 2026

Maria Frangiadaki [1], Konstantinos Klapsas [1], Kosmas Kritsis [1]
Charilaos Papaioannou [1], Georgios Paraskevopoulos [1], Vassilis Katsouros [1]
[1] Institute for Language and Speech Processing, Athena Research Center, Greece

1. Task & Motivation

Automatic Lyrics Transcription (ALT) aims to automatically recover sung lyrics from polyphonic music recordings. The MIREX 2026 challenge focuses on multilingual transcription, with the goal of producing accurate lyrics directly from full-song audio.

Unlike conventional speech recognition, singing introduces additional challenges, including instrumental interference, prolonged vowels, variable vocal styles, and overlapping sounds. Moreover, accurate word recognition alone is not sufficient, since the output should preserve meaningful line breaks and song structure to produce a readable lyric sheet.

2. Our Approach

We propose a modular, multi-model pipeline combining segment-level transcription, full-song recognition, and LLM-based candidate fusion. The system integrates complementary transcription hypotheses to improve lexical accuracy while preserving the structure and readability of the resulting lyrics.

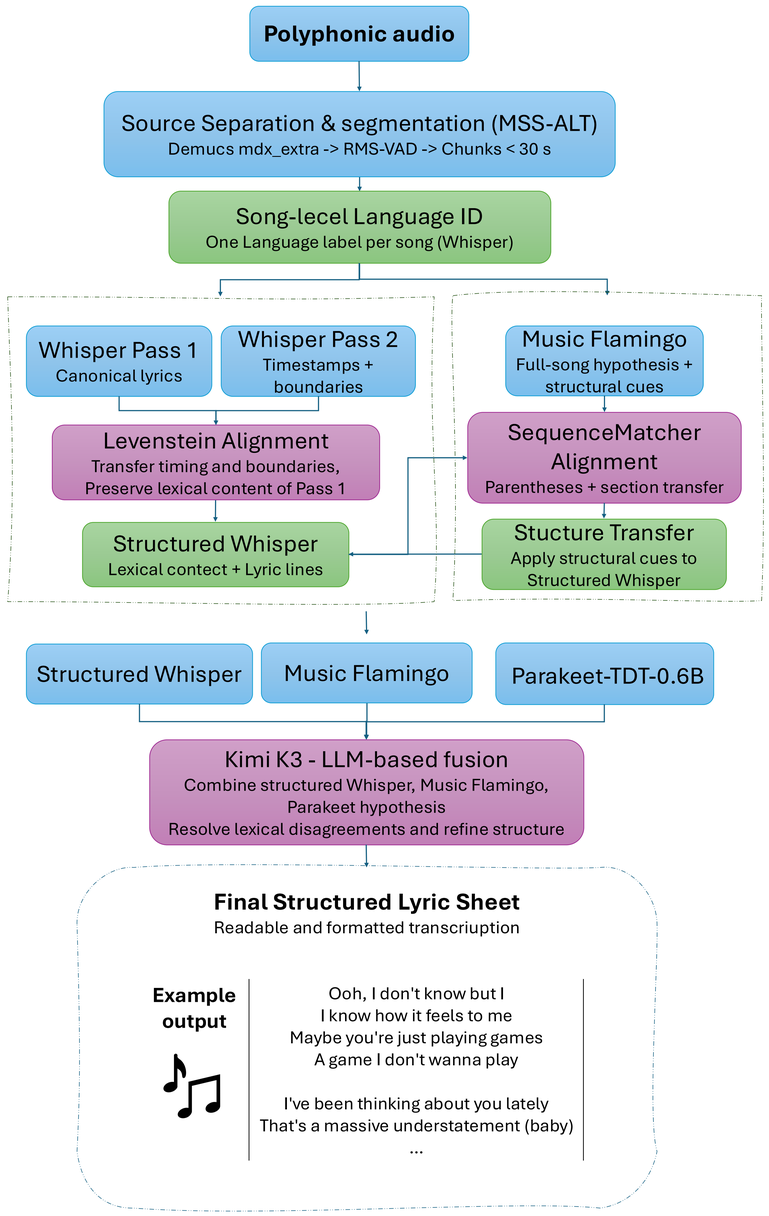
Key contributions:

- Integration of complementary music and speech recognition models.
- Combination of segment-level and full-song transcription hypotheses.
- LLM-based fusion for lexical correction and lyric-sheet reconstruction.
- Evaluation across three multilingual ALT benchmarks.

Key Findings

- **Lexical accuracy:** Kimi K3 fusion reduces WER from 27.36% (Whisper) to 22.37% on DALI, an 18.2% relative reduction.
- **Whisper baseline:** Whisper, combined with long-form segmentation and structure-aware post-processing, provides a strong lexical foundation while recovering lyric-line and section boundaries.
- **Complementary evidence:** Music Flamingo provides structural cues, while Parakeet contributes additional lexical hypotheses despite their higher standalone WERs.

3. System Overview



4. Results

Dataset	CER ↓	WER ↓	WER' ↓	F-score (P) ↑	F-score (B) ↑	F-score (L) ↑	F-score (S) ↑
JAM-ALT	8.81%	14.70%	19.13%	51.34%	49.32%	80.26%	75.87%
MUSDB-ALT	11.63%	16.17%	19.70%	29.41%	31.90%	83.78%	74.85%
JamendoLyrics++	12.63%	17.68%	—	—	—	82.50%	76.55%

Table 1. Official MIREX 2026 challenge results (%) for our submitted system.

Lang.	N	WER ↓	WERcase ↓	Punc. F1 ↑	Paren. F1 ↑	Line F1 ↑	Sect. F1 ↑
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 2. Final Kimi K3 fusion results (%) on our 195-song held-out DALI test set.

System	WER ↓	WERcase ↓	Punc. F1 ↑	Paren. F1 ↑	Line F1 ↑	Sect. F1 ↑
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 3. Intermediate transcription outputs and final fusion results (%) on our 195-song held-out DALI test set.

Acknowledgements

The authors acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC super-computer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium, through a EuroHPC Development Access call (Project No. EUHPC-D34-036).Partial funding from the European High-Performance Computing Joint Undertaking (JU) under Grant Agreement No. 101234269 for the Pharos AI Factory project, as well as from the Greek Ministry of Digital Governance and Artificial Intelligence.