

REPETITION-AWARE PHRASE PRIORS FOR DALI-ADAPTED LYRICS-TO-AUDIO ALIGNMENT

Emil Zawistowski

Department of Electrical and Photonics Engineering

Technical University of Denmark

s262793@dtu.dk

ABSTRACT

This paper describes a word-level lyrics-to-audio alignment system prepared for MIREX 2026. A common failure mode in lyrics alignment is section-selection error: when a chorus repeats, text-based anchoring may attach reference words to the wrong musical occurrence, causing large timing offsets. The central decoding modification addresses this by widening an ASR-derived phrase prior over all exact matching occurrences, allowing the acoustic decoder to choose the correct repetition based on local evidence rather than following a single text-alignment match. The system combines vocal separation, a phoneme connectionist temporal classification (CTC) model adapted on 957 English DALI songs, and global Viterbi decoding with soft timing and silence priors. At both adaptation scales, repetition-aware priors reduce macro onset absolute error. For the 957-song model, they reduce error from 0.214 to 0.174 seconds with independently sourced lyrics and from 0.153 to 0.138 seconds with reference-derived text. These gains are concentrated in a few songs. Furthermore, increasing the acoustic adaptation set alone does not consistently improve mean error. The final system places 95.05% and 95.72% of matched word onsets within 0.3 seconds, averaged across songs in the two text conditions. These are development results, not official MIREX scores. The system is packaged as an offline Linux container.

1. SYSTEM

Earlier lyrics alignment systems include end-to-end audio-to-character recognition [1] and joint phoneme/pitch modeling [2]; Huang and Benetos [3] study alignment under source separation. Here, I examine ASR-derived soft phrase priors within a singing-adapted phoneme system. Given mixed audio and written lyrics, the system returns one onset, offset, and original label per input word.

HT-Demucs separates vocals [4]. Whisper-small [5] transcribes the separated vocals. Its output provides structural timing for phrase-level anchoring instead of replacing the provided lyrics. The wav2vec 2.0 phoneme model [6,7]

(facebook/wav2vec2-lv-60-espeak-cv-ft) supplies acoustic scores for CTC decoding [8]. This separates macro-level phrase localization from micro-level phoneme alignment. English grapheme-to-phoneme conversion uses g2p_en with CMU-to-IPA mapping. Long recordings are processed in overlapping acoustic windows whose frame scores are combined before full-song Viterbi decoding. Pronunciation failures are represented explicitly; final postprocessing enforces the input word sequence and monotonic timestamps. Inference uses a fixed sampling seed and the same transcription configuration across compared models.

1.1 Soft phrase constraints

A common failure mode in lyrics alignment is *section-selection error*: when a chorus repeats multiple times, text-based anchoring may match reference words to the acoustically wrong occurrence, causing timing offsets of ten or more seconds that dominate mean absolute error. The phrase prior mechanism addresses this problem directly.

An edit-distance alignment between normalized ASR and supplied words identifies consecutive exact matches of at least three words. A matched word's first phoneme receives a soft interval constraint, with 1.5 seconds margin. The decoder adds a bounded cost of 0.5 per second of displacement outside this interval. It applies this penalty only upon token entry, not continuously over a sustained phoneme.

For repeated phrases, the system finds all exact occurrences in the ASR transcript. For each corresponding word position, the interval spans the earliest onset and latest offset across these occurrences. The acoustic decoder can therefore choose among repetitions instead of following only the occurrence selected by text alignment. This policy uses only input text and ASR output; reference timings are never available to it. Non-repeated phrases retain their original constraints.

Two complementary priors control silence. Sustained quiet vocal regions receive a blank-score boost of 8, using a threshold of -30 dB relative to a robust song-level vocal reference. An intra-word blank penalty of 1 log-score per second discourages excessive gaps between phonemes within a word. Leading, trailing and inter-word blanks remain unpenalized. No onset snapping is applied.



Configuration	ALT				Exact			
	AAE	MedAE	PCE _{.3}	PCS	AAE	MedAE	PCE _{.3}	PCS
Unadapted baseline	0.3077	0.0361	91.75	83.44	0.3088	0.0383	91.75	85.26
500 standard	0.1946	0.0249	94.85	86.46	0.1648	0.0251	95.51	89.65
957 standard	0.2138	0.0223	94.75	87.14	0.1528	0.0223	95.64	89.83
500 + repeat	0.1788	0.0249	94.93	86.50	0.1487	0.0251	95.60	89.66
957 + repeat	0.1740	0.0222	95.05	87.28	0.1381	0.0223	95.72	89.86
no quiet-region prior	0.1842	0.0221	95.12	87.56	0.1400	0.0224	95.67	89.62
no intra-word penalty	0.1659	0.0220	95.22	86.78	0.1481	0.0224	95.42	89.57

Table 1. MUSDB39 development results. AAE and MedAE are in seconds; PCE and PCS are percentages. MedAE is the average within-song median. “Standard” retains phrase, quiet-region and intra-word blank priors; “repeat” changes only the phrase-interval policy. The indented rows are single-component ablations of the 957+repeat configuration. The “Unadapted baseline” omits DALI adaptation. PCS uses proxy reference offsets.

2. ADAPTATION AND DATA

DALI supplies weak word timings and source-recording links [9]. Logged sweeps acquired available English audio, yielding 4,305 manifest entries and 4,259 entries after ID, video-identity and exact-audio deduplication. Manifests retain exclusions and provenance. This acquisition pool is distinct from the submitted model’s training set.

The final experiment uses 957 training songs: 946 with annotation NCC at least 0.90 and 11 retained baseline songs with NCC between 0.813 and 0.90. Four original validation songs are held out by artist. Additional reserved artists are excluded from training. Duration and annotation/video identity are checked; normalized artists and audio hashes do not rule out aliases or covers.

The feature manifest contains 48,818 usable training crops and 183 validation crops. Training-only loss filtering retains 43,936 crops for output-head training; duration and loss filtering retain 43,550 for adapter training. The head is trained for ten epochs with AdamW, learning rate 10^{-4} , batch size eight and frame distillation weight 0.25. Rank-eight query/value LoRA [10] adapters with scaling parameter 16 are then trained for three epochs, learning rate 10^{-4} , accumulation eight and no distillation. Backbone and head weights are frozen during LoRA training. Checkpoint selection uses macro validation-song CTC loss; epoch three is selected. A single seed, 20261006, is used.

Jamendo is not used for training, validation, model selection or tuning. MUSDB18 audio [11] and Huang–Benetos word annotations [3] are used for development evaluation, including system selection. Pretrained models’ full training corpora have not been independently audited for overlap.

3. DEVELOPMENT EVALUATION

I evaluate the system on the 39-song intersection of MUSDB-ALT and the Huang–Benetos word references [3]. In ALT, independently sourced lyrics are aligned and onset errors are scored on longest-common-subsequence matches (macro reference coverage 91.97%). In Exact, input text comes from the references and every word is scored. AAE is the mean absolute word-onset er-

ror; MedAE is its within-song median; PCE_{.3} is the fraction of scored onsets within 0.3 seconds. PCS measures correctly labeled duration using the next onset as a proxy offset (the references lack offset annotations). All table entries are macro averages with equal song weight.

Table 1 provides a paired ablation. DALI adaptation reduces ALT AAE from 0.308 to 0.195–0.214 s depending on scale. Increasing adaptation from 500 to 957 songs improves Exact AAE (0.165 to 0.153 s) but worsens ALT AAE (0.195 to 0.214 s); the within-song median improves on 38 of the 39 songs, but persistent large errors offset this for the mean. For the 957-song model, repetition priors improve ALT AAE on two songs and leave 37 unchanged; paired 95% bootstrap intervals for the macro AAE difference (repeat minus standard) are $[-0.103, 0.000]$ s (ALT) and $[-0.049, 0.005]$ (Exact). For the silence priors, removing the quiet-region boost raises ALT AAE by 0.010 s; removing the intra-word blank penalty lowers ALT AAE by 0.008 s but reduces PCS by 0.005. Neither is dominant. For 957+repeat, median song AAE is 0.067 s in ALT (IQR 0.031–0.138); maximum song AAEs are 1.31 and 0.83 s. Evaluation on five additional development songs yields a macro AAE of 1.01 s, which is dominated by two persistent section-selection failures (3.06 s and 1.63 s).

4. REPRODUCIBILITY AND LIMITATIONS

The container accepts three positional paths (audio, lyrics, output) and writes tab-separated ASCII onset, offset and word labels. Pretrained models, pronunciation resources and adaptation checkpoints are packaged for offline execution. Checkpoint hashes, training recipes, output hashes and per-song results are retained. CUDA execution and target-hardware throughput require separate verification; development accuracy was measured on Apple MPS. I do not claim to outperform published systems on different evaluation corpora. Limitations include weak labels, four validation songs, repeated development-set inspection, one seed, text mismatches and section-selection errors. Song-bootstrap intervals omit training-seed and selection uncertainty. Fixed epochs give larger datasets more updates; the extra five songs are also development data.

5. REFERENCES

- [1] D. Stoller, S. Durand, and S. Ewert, “End-to-end Lyrics Alignment for Polyphonic Music Using an Audio-to-Character Recognition Model,” ICASSP, 2019, pp. 181–185.
- [2] J. Huang, E. Benetos, and S. Ewert, “Improving Lyrics Alignment through Joint Pitch Detection,” ICASSP, 2022, pp. 451–455.
- [3] J. Huang and E. Benetos, “Evaluating Lyrics Alignment under Source Separated Conditions,” IS-MIR Late-Breaking Demo, 2025. Word annotations: <https://zenodo.org/records/15547046>.
- [4] S. Rouard, F. Massa, and A. Défossez, “Hybrid Transformers for Music Source Separation,” ICASSP, 2023.
- [5] A. Radford et al., “Robust Speech Recognition via Large-Scale Weak Supervision,” ICML, 2023, pp. 28492–28518.
- [6] Q. Xu, A. Baevski, and M. Auli, “Simple and Effective Zero-shot Cross-lingual Phoneme Recognition,” Interspeech, 2022, pp. 2113–2117. Checkpoint: <https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft>.
- [7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” NeurIPS, 2020.
- [8] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks,” ICML, 2006, pp. 369–376.
- [9] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, “DALI: A Large Dataset of Synchronized Audio, Lyrics and Notes, Automatically Created Using Teacher-Student Machine Learning Paradigm,” IS-MIR, 2018.
- [10] E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” ICLR, 2022.
- [11] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 Corpus for Music Separation,” 2017. <https://doi.org/10.5281/zenodo.1117372>.