Learning Audio Embeddings via Lyrics Alignment for Scalable Version Identification

Joanne Affolter[1,2], Benjamin Martin[1], Elena V. Epure[1], Gabriel Meseguer-Brocal[1], Frédéric Kaplan[2]

[1] Deezer Research, Paris, France; [2] EPFL, Lausanne, Switzerland

1. CONTEXT

Task: *Version Identification aims to identify distinct renditions of the same underlying work [1]*
→ Critical for catalog management, copyright enforcement, and music retrieval

Harmony / Melody
- Basis of most SOTA models
- Require complex pipelines to handle tempo, pitch, structural changes

→ Powerful bust costly

Lyrics
- Strong invariant across covers [2,5,7]
- Underused due to difficulty of extracting lyrics from audio and limited availability of editorial data [1]

→ Promising but either weak or overly complex

2. APPROACH

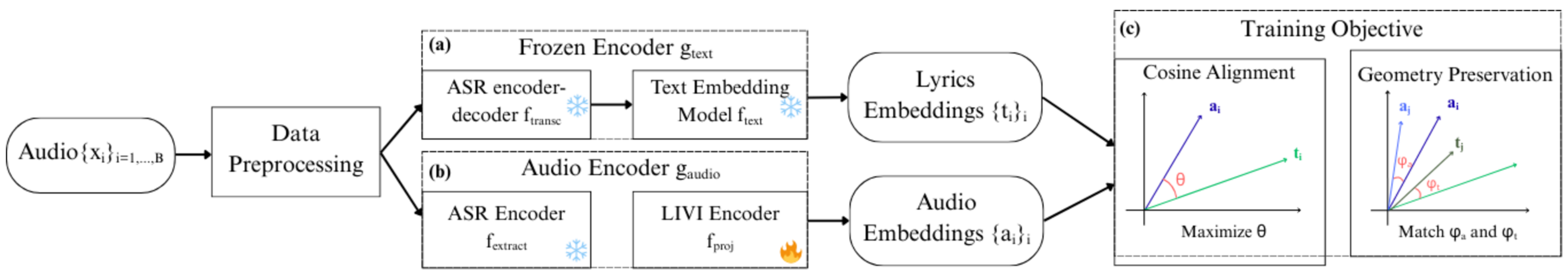
1. Build a Lyrics-Informed Embedding Space
   - Audio → ASR transcription → multilingual text encoder
   - Produces embeddings where semantically similar lyrics cluster closely, even across languages

→ Strong retrieval performance, but costly (requires full transcription at inference)

2. Train LIVI: Audio → Lyrics Space Directly
   - Use Whisper encoder latents (no decoder) + projection network
   - Learn to map audio into the lyrics-informed space, preserving its geometry

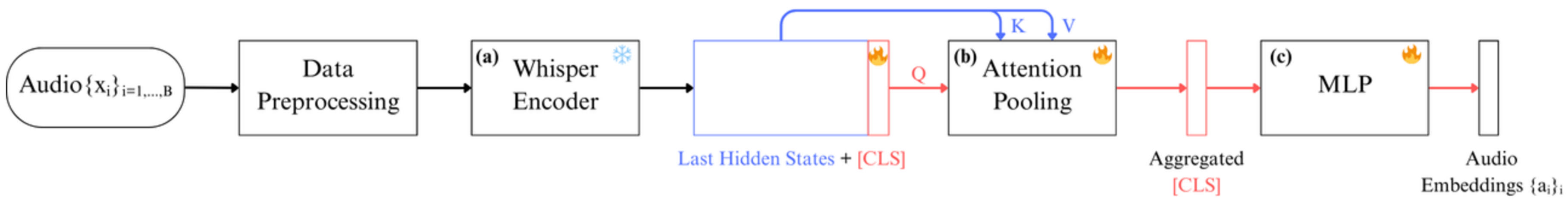
→ Removes transcription at inference, cutting cost while keeping accuracy

3. METHODOLOGY



(a) A frozen text encoder g_{text} combines an ASR model with a pre-trained text embedding model to produce lyrics embeddings t_i
(b) An audio encoder g_{audio} projects ASR encoder latent representations into the same embedding space
(c) Training optimizes a combined objective: pointwise alignment of a_i with t_i under cosine similarity, and geometry preservation ensuring that pairwise similarities between audio embeddings mirror those of their corresponding lyric embeddings

4. IMPLEMENTATION DETAILS



Data Preprocessing
- Detect and keep vocal segments only with vocal detection model → 30s segments, serving as inputs to g_{text}, g_{audio}
- Ensures input has enough lyrical content, avoids ASR hallucinations on instrumental sections

Lyrics-Informed Embedding Space
- whisper-large-v3-turbo (ASR) [8] + gte-multilingual-base (text encoder) [9]

Audio Encoder (LIVI)
(a) Raw audio is first processed by the Whisper encoder to obtain hidden representations
(b) A [CLS] token is appended to aggregate frame-level features using an attention pooling mechanism
(c) A multi-layer perceptron projects the pooled representation into the lyrics-informed embedding space, yielding the final audio embedding ai

5. RESULTS

EXPERIMENTAL SETUP

Datasets
- Covers80 (116) [6], SHS100k-TEST (890) [4], Discogs-VI (72,316 tracks) [3]
- Retain ~82–85% after vocal-content filtering

Evaluation Protocol
- Task: retrieval via cosine similarity
- Metrics: MR1, HR@1, MAP@10
- Multiple covers per query (avg. 2–12 depending on dataset)

AUDIO-LYRICS ALIGNMENT

Cosine similarity between audio and lyric embeddings, evaluated at segment (167k pairs) and track (60k segments) levels. For track-level, 30s segment embeddings are averaged into a global representation.

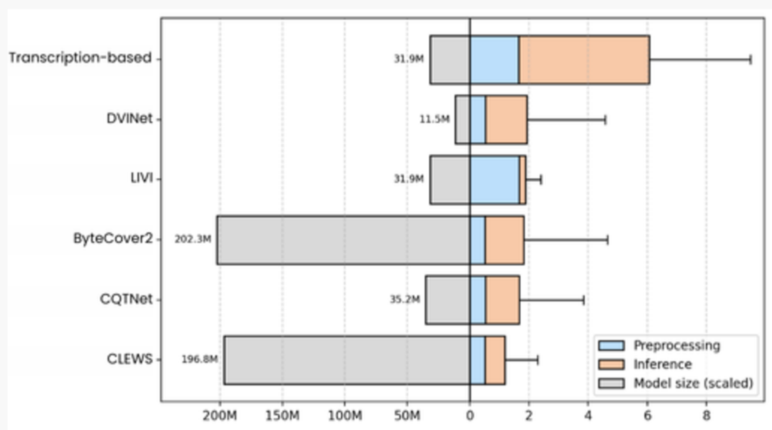
S: 0.857
std: 0.075

T: 0.910
std: 0.037

- Higher mean + lower variance at track level = stable global representations
- Confirms tight alignment of audio and lyric embeddings

MODEL SIZE AND INFERENCE

Runtime and model size comparison. Average preprocessing and inference times are shown alongside model sizes for LIVI and baseline models. Error bars denote std across runs.



- End-to-end: 1.9s / track
  - 3.2× faster than transcription pipeline (6.07s, Whisper = 4.41s)
- Inference: 0.22s
  - 20× faster than Whisper
  - 3–6× than audio baselines
- Accuracy vs. Complexity trade-off
  - Accuracy comparable to SOTA at a fraction of size
  - Surpasses models of similar size

APPLICATION TO VERSION IDENTIFICATION

Comparison of LIVI audio encoder against transcription-, Whisper-, and audio-based baselines. t_{global} denotes lyrics embeddings from the full transcription, while t_{local} and LIVI correspond to the mean of 30s segment-level embeddings (lyrics and audio). Bold numbers indicate the best result and underlined numbers the second-best within each row.

	Metric		LIVI	t_{global}	t_{local}	Whisper	Bytecover2	CLEWS	CQTNet	DViNet
C80	MR1	↓	<u>1.51</u>	**1.10**	1.92	7.67	1.57	2.24	3.43	3.05
	HR1	↑	<u>0.949</u>	**0.975**	0.937	0.632	0.865	0.835	0.848	0.861
	MAP	↑	<u>0.966</u>	**0.979**	0.945	0.691	0.877	0.880	0.856	0.886
SHS	MR1	↓	**3.25**	6.05	5.52	6.56	4.66	<u>3.97</u>	5.59	7.63
	HR1	↑	0.935	**0.954**	0.925	0.777	<u>0.953</u>	0.931	0.900	0.931
	MAP	↑	<u>0.875</u>	**0.910**	0.870	0.558	0.884	0.847	0.789	0.859
D-VI	MR1	↓	**232.21**	<u>275.77</u>	360.21	1051.36	312.32	410.39	810.89	507.04
	HR1	↑	<u>0.853</u>	**0.856**	0.843	0.524	0.843	0.816	0.641	0.751
	MAP	↑	**0.923**	<u>0.832</u>	0.817	0.406	0.812	0.790	0.568	0.719

- LIVI nearly matches lyric upper bounds
- Outperforms raw Whisper (avg pooling over encoder hidden states)
- Competes with or surpasses SOTA audio baselines

TAKEWAY

LIVI: a compact audio encoder aligned with lyrics
- Balances accuracy and efficiency
- Competes with or surpasses SOTA audio baselines using a simpler, reproducible design

Limitations
- Relies on vocal detection → adds preprocessing cost, excludes instrumental tracks
- Uses an off-the-shelf text encoder

Future Work
- Fine-tune text encoder to improve discriminability
- Reduce cost of vocal detection
- Extend to multimodal systems to handle non-vocal music

RESOURCES

Deezer
https://research.deezer.com

Contact for any questions
+41 78 250 08 59 (Whatsapp)
joanne.affolter@gmail.com

REFERENCES

[1] Yesiler, F., Doras, G., Bittner, R.M., Tralie, C.J., Serra, J.: Audio-based musical version identification: Elements and challenges. IEEE Signal Processing Magazine 38(6), 115–136 (Nov 2021). https://doi.org/10.1109/msp.2021.3105941
[2]Abrassart, M., Doras, G.: And what if two musical versions don't share melody, harmony, rhythm, or lyrics? In: Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR). pp. 677–684. Bengaluru, India (2022)
[3] Araz, R.O., Serra, X., Bogdanov, D.: Discogs-vi: A musical version identification dataset based on public editorial metadata. In: Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR). San Francisco, USA (2024)
[4] Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR). pp. 591–596. Miami, Florida, USA (2011)
[5] Du, X.: X-cover: Better music version identification system by integrating pretrained asr model. In: Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR). San Francisco, USA (2024)
[6] Ellis, D.P.W.: The "covers80" cover song data set (2007), http://labrosa.ee.columbia.edu/projects/coversongs/covers80
[7] Vaglio, A., Hennequin, R., Moussallam, M., Richard, G.: The words remain the same: Cover detection with lyrics transcription. In: 22nd International Society for Music Information Retrieval Conference ISMIR 2021 (2021)
[8] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23,JMLR.org (2023)
[9] Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., et al.: mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 1393–1412 (2024)