Zero-Shot Structure Labeling with Audio And Language Model Embeddings

Morgan Buisson, Christopher Ick, Tom Xi, Brian McFee

ZERO-SHOT STRUCTURE LABELING WITH AUDIO AND LANGUAGE MODEL EMBEDDINGS

Morgan Buisson[1*] **Christopher Ick**[2*] **Qingyang (Tom) Xi**[2] **Brian McFee**[2]
[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[2] Music and Audio Research Laboratory, New York University, USA
*Equal Contribution
morgan.buisson@telecom-paris.fr, {chris.ick, tom.xi, brian.mcfee}@nyu.edu

ABSTRACT

Recent progress on audio-based music structure analysis has closely aligned with the appearance of new deep learning paradigms, notably for the extraction of robust spectro-temporal audio features and their sequential modeling. However, most recent methods resort to supervised learning, which requires careful annotation of audio music pieces. Such annotations may sometimes operate at different temporal scales from one dataset to another or comprise inconsistent variation markers across repetitions of identical segments. This work explores language models as an alternative to manual pre-processing of the section label space, thus facilitating training and predictions across different annotated corpora. We propose a joint audio-to-text embedding space in which latent representations of audio frames and their respective section labels are close. We take inspiration from recent works on cross-modal contrastive learning and demonstrate the plausibility of this paradigm in the context of music structure analysis.

1. INTRODUCTION

Music structure analysis is the task of dividing a given piece into non-overlapping segments, with a label corresponding to semantic information about the segment [1]. Existing methods typically use labels with no semantic information (e.g. A, B, C...) [2–5], or explicitly consider the semantics of musical section labels [6, 7], where predicted categories now hold a specific meaning that is shared across tracks. These methods generate labels based on predefined taxonomies. When working with several datasets, manual pre-processing of the labels is necessary to ensure compatibility among their distinct taxonomies. As such, previous work generally restrains the label space to a handful of categories, representing common musical sections such as *Intro*, *Verse*, *Chorus*, *Bridge* and *Outro* [6]. Doing this discards valuable information that is contained in the original form of the annotations before being processed. For example, variation markers between identical segments may indicate important musical changes (*Verse A*, *Verse B*), into which several degrees of similarity between sections may be encapsulated.

In the meantime, the intrinsic relationship between text and audio has been an active area of research. Notably, multi-modal pre-training strategies based on contrastive learning have shown great promise at learning joint audio-text embedding spaces [8–11], benefiting tasks such as few-shot sound event detection, text-to-music retrieval and classification [12].

In this work, we investigate how fine-tuning via adapter networks may address annotation inconsistencies and ambiguities across datasets.

2. METHOD

The goal of the method presented in this work is to learn a shared audio-text embedding space in which latent representations of a given audio frame will be located near semantically relevant section labels. This process is summarized in Figure 1. From here, the decoded labels can be used to simultaneously define segmentation boundaries and labels associated with each predicted segment.

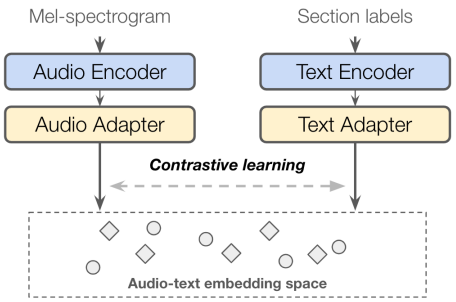

Figure 1: Audio and section labels of a given track are passed through modality encoders and adapter networks to learn the (dis)similarity of audio and section label pairs using contrastive learning. Both the audio and text encoders are kept frozen, and only the adapters are trained.

During inference, segmentation predictions are obtained through a simple frame-wise nearest-neighbor decoding, where the section label of an individual audio frame is deduced from its most similar label embedding. We use discriminative Viterbi decoding to reduce small

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

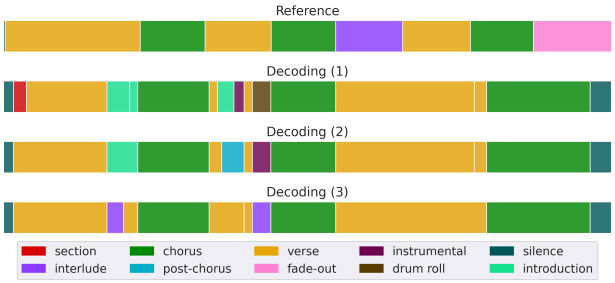

Figure 2: Segmentation example for the track n°1634 from the SALAMI dataset [15].

discontinuities in the final section assignment predictions. Here, we consider three cases where test audio frames are decoded using: (1) section labels present in the training set; (2) section labels present in the test set; (3) section labels present in a single test track. Note that in the last two cases, the model may have never seen some of the test labels during training (*i.e.* zero/few-shot setting) [13, 14]. Figure 2 illustrates an output segmentation example.

3. EXPERIMENT

Our system uses an audio encoding system from [16] and a text encoder from [17], a 2-layer MLP adapter network attached to the output of the text encoder and a 2-layer transformer encoder after the audio encoder. Both adapters produce 128-dimensional embeddings. We train the system on the Harmonix dataset [18] and evaluate it on the functional labels from the SALAMI dataset [15], which amounts to a total of 123 distinct section labels used during training and 37 for testing. We also train an equivalent audio-specific model, which only comprises the audio adapter and uses a classification loss (denoted as *Classif.*) over the training label taxonomy. We use common music segmentation metrics for evaluation [1], using a 3-second tolerance window for evaluating boundary hit rates (P3, R3, F3) and summarize results in Table 1 below:

	P3	R3	F3	PFC	NCE
Classif.	.254	**.665**	.349	.494	.511
Ours (1)	**.399**	.553	**.439**	.468	**.579**
Ours (2)	.371	.570	.425	.483	.577
Ours (3)	.339	.493	.372	**.572**	.536

Table 1: Segmentation results on the SALAMI dataset [15].

4. ANALYSIS

The decoding method (1) provides less noisy frame-wise section decoding (higher precision, P3) than the classification baseline on boundary detection, yielding better boundary detection and NCE results. On the other hand, the baseline returns a higher pairwise frame clustering score (PFC), possibly because the model is trained to classify audio frames among all possible labels in the training set. Test labels also tend to be well-separated in the learned audio-text latent space, as relatively consistent segmentation can be seen between strategies (1) and (2). This demonstrates that the text encoder somewhat generalizes across labels not observed during training. Finally, decoding only with

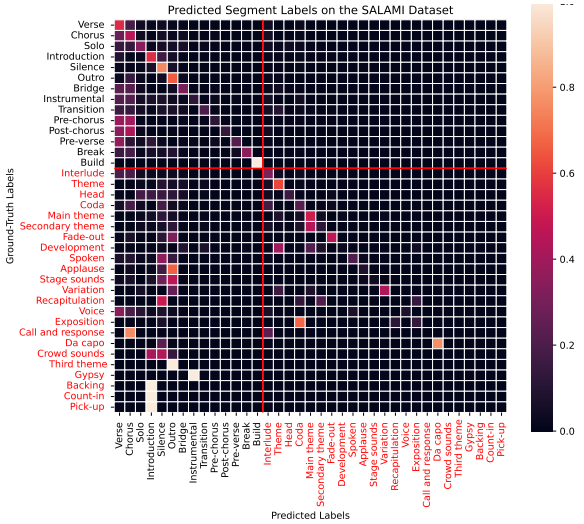

Figure 3: A heatmap showing label predictions on [15], normalized per ground-truth label. Red labels indicate that the labels were not present in the dataset [18] used for training the adapter networks.

track-specific labels retrieves a much lower proportion of segment boundaries (lower recall, R3). This also reduces the number of distinct segment labels produced during decoding, improving PFC.

When examining the cross-prediction of labels, we can see that while mislabeling occurs, a good amount of semantic information translates between different labels, allowing for relatively good *zero-shot* performance. Despite the different label spaces between the two sets, our learned joint embedding space is able to correctly decode labels outside of the original training set due to its semantic similarity to the original training dataset. Due to a large imbalance in classes, the predictions typically favor more common/well-defined labels such as *Chorus*, *Verse*, *Solo*, *Introduction*, etc. Class-balancing our training set to limit the effect of dominant labels could help regularize our results and better investigate labels' semantic proximity.

5. CONCLUSIONS

This work proposed a novel approach to music structure labeling by learning a joint audio-text embedding space. We demonstrated that this method improves upon audio-only segmentation strategies and can generalize to out-of-distribution segmentation labels leveraging the semantic information captured by the text encoder. Further extensions of this work could include experimenting with more task-appropriate text encoders, such as ones pre-trained on musically relevant semantic information. Our label space is also relatively limited due to the standardization of our datasets [15, 18]; methods in augmenting the label space could help better disperse the text embeddings. Finally, this approach could be coupled with existing music segmentation methods to retrieve labels of predicted segments at one or multiple structural levels.

This material is based upon work supported by the National Science Foundation under NSF Award 1922658.

6. 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*, Dec 2020.

[2] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 318–326, 2008.

[3] B. McFee and D. Ellis, "Analyzing song structure with spectral clustering." in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2014, pp. 405–410.

[4] J. Serra, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.

[5] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 236–240.

[6] J.-C. Wang, Y.-N. Hung, and J. B. Smith, "To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 416–420.

[7] T. Kim and J. Nam, "All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio," in *Proceedings of 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023, pp. 1–5.

[8] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2022, pp. 640–649.

[9] I. Manco, B. Weck, P. Tovstogan, M. Won, and D. Bogdanov, "Song describer: a platform for collecting textual descriptions of music recordings," in *Late-Breaking Demo Session of the 23rd Int'l Society for Music Information Retrieval Conf. India*, 2022.

[10] S. Doh, K. Choi, J. Lee, and J. Nam, "Lp-musiccaps: Llm-based pseudo music captioning," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2023, pp. 409–416.

[11] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] Y. Wang, N. J. Bryan, M. Cartwright, J. Pablo Bello, and J. Salamon, "Few-shot continual learning for audio classification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 321–325.

[14] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot Learning for Audio-based Music Classification and Tagging," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 67–74.

[15] J. Smith, J. Burgoyne, I. Fujinaga, D. De Roure, and J. Downie, "Design and creation of a large-scale database of structural annotations," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 555–560.

[16] M. Buisson, B. McFee, S. Essid, and H. C. Crayencour, "Self-supervised learning of multi-level audio representations for music segmentation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2141–2152, 2024.

[17] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[18] O. Nieto, M. C. McCallum, M. E. Davies, A. Robertson, A. M. Stark, and E. Egozy, "The harmonix set: Beats, downbeats, and functional segment annotations of western popular music." in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2019, pp. 565–572.