Knowledge-based Multimodal Music Similarity

Andrea Poltronieri[1 [0000−0003−3848−7574]]

Department of Computer Science and Engineering, University of Bologna, Italy
andrea.poltronieri2@unibo.it

Abstract. Music similarity is an essential aspect of music retrieval, recommendation systems, and music analysis. Moreover, similarity is of vital interest for music experts, as it allows studying analogies and influences among composers and historical periods.
Current approaches to musical similarity rely mainly on symbolic content, which can be expensive to produce and is not always readily available. Conversely, approaches using audio signals typically fail to provide any insight about the reasons behind the observed similarity.
This research addresses the limitations of current approaches by focusing on the study of musical similarity using both symbolic and audio content. The aim of this research is to develop a fully explainable and interpretable system that can provide end-users with more control and understanding of music similarity and classification systems.

Keywords: Music Similarity · Computational Musicology · Knowledge Graphs.

1 Introduction

Music similarity is a central area of research in the field of Music Information Retrieval (MIR) [11] as it enables various applications, such as music recommendation, playlist generation, music search, and classification. The ability to measure the similarity between music tracks is essential for providing personalised and relevant recommendations to users based on their listening history and preferences [26]. Music similarity also facilitates the discovery of new music that matches the user's taste [28]. Additionally, music similarity can be used for content-based music classification, such as genre classification [10]. It is also useful in musicological research, as it allows for the exploration of musical patterns and structures across different styles and genres [36].

1.1 Problem Statement

The study of musical similarity is approached from various perspectives, which can be summarised in *content-based systems* and *context-based systems* [20]. The former approach extracts information directly from the musical content (whether symbolic or audio), while the latter obtains information from non-musical data, such as metadata or information related to the song's popularity or listener

characteristics. Content-based approaches allow a quantitative measurement of similarity based on factual music data, and make it possible to investigate similarities independently of the availability and accuracy of metadata [19].

However, studying content-based similarity poses several challenges, given the multidisciplinary nature of the research, which encompasses music theory, ethnomusicology, cognitive science, and computer science [36].

In content-based music similarity, a further distinction must be made concerning the representation of music. Two types of representations have been identified: *signal representations* that are recordings of sound sources, and *symbolic representations* that represent discrete musical events [37]. Symbolic representations are context-aware and offer a structured representation from which is easy to extract information from. On the other hand, signal representations are content-unaware and not structured, which makes extracting information from them a challenging task [38]. Signal representations are by far more studied than symbolic representations, since they are more interesting from a commercial point of view (e.g. for streaming services) and the data availability is higher.

Depending on the type of musical representation, several features can be used for similarity analysis: *descriptive metadata*, *low-level features*, and *high-level features* [39]. Descriptive metadata is text-based information about the song, while low-level features are extracted from the audio signal (e.g., beat, tempo) and are efficient but difficult to interpret. High-level features, on the other hand, are content descriptors that reflect the knowledge of experienced or professional listeners, making them the most intuitive approach for music classification tasks.

Most of the available music similarity systems, especially those based on audio signals [36], rely on low-level features. Annotating high-level content descriptors is also expensive and requires the expertise of musicians and musicologists [35]. As a result, most available systems cannot explicitly recognise similarity motives, and their lack of interpretability and transparency can lead to biased recommendations.

This results in a measure of similarity that is neither interpretable nor transparent, which may result in biased results [21].

1.2 Expected Contribution

This research proposes a fully explainable and interpretable system that provides information on musical similarity based on both symbolic and audio content, with a focus on factual musical data such as melodic and harmonic patterns.

RQ1 *What is an effective method to create high-quality datasets that incorporate multimodal data that links symbolic annotations (both melodic and harmonic) and audio?*

To achieve this, the symbolic content needs to be studied first to assess similarity in a transparent and explainable way.

RQ2 *How can similarity measures be derived from this knowledge graph in order for it to be objectively measured and quantified?*

Next, an alignment of the symbolic content with the audio signal using multimodal datasets must be performed. Finally, a deep learning system is trained to analyse the audio signal informed by the symbolic content. By doing so, it is possible to provide end-users with more control and understanding of the music similarity and classification systems they use, regardless of the representation under analysis.

RQ3 *How can score-informed audio analysis be used to identify similarities and patterns in audio data, and what are the benefits of this approach for the study of music similarity?*

The current study focuses on the application of Semantic Web technologies, particularly in the representation and alignment of multimodal data. One of the key challenges is how to effectively encode knowledge graphs (KGs) to enable their use as input and mapping onto various mathematical models, such as timeseries and embedding.

2 Related Works

2.1 Symbolic Music Similarity

The study of similarity on symbolic content has been studied in depth in recent years. Various approaches have been proposed, ranging from harmonic similarity to melodic and rhythmic similarity.

Melodic similarity is the most extensively researched category. Algorithms that handle melodic similarity in symbolic form are typically rule-based and aim to define various types of context-dependent similarity functions, which rely on music theory [30]. However, these algorithms lack a shared definition of similarity and primarily focus on studying similarity in monophonic sequences [36].

On the other hand, algorithms for harmonic similarity has not received much attention in recent years. To the best of my knowledge, current state-of-the-art methods for this task are the *Tonal Pitch Step Distance* (TPSD) [15] and the *Chord Sequence Alignment System* (CSAS) [16]. These studies consider tracks similar only if their harmonic profiles are globally aligned, providing no information on local similarity.

Studies using a combination of harmonic and melodic content to calculate similarity are limited to a few contributions [14].

2.2 Audio Music Similarity

Music similarity in the audio signal domain has been studied for a wide range of applications, ranging from cover song identification [32] to recommendation systems [12]. These algorithms are based on the extraction of low-level features directly from the signal, such as spectrograms, MFCCs and Chorma Features [13].

One of the main limitations of these approaches is their reliance on deep learning approaches. These methods are based on end-to-end algorithms that do not provide valuable information regarding fundamental aspects of similarity, such as the explanation for why two or more tracks are similar, and the highlight of parts in common between different tracks.

2.3 Multimodal Music Similarity

Multimodality refers to the integration of multiple representation modes, such as visual, auditory, and textual.

In the realm of music, multimodality has become an increasingly popular field of research in recent years and has proven to provide better results in different tasks, if compared to approaches that consider a single modality [3, 33].

One of the primary areas of research in multimodal MIR is the integration of audio and textual data. Moreover, multimodality has been explored also for other tasks, such as audio-to-score alignment [29] and classification [22].

However, less emphasis has been placed on algorithms that combine audio and symbolic annotations, particularly in the field of classification and similarity. Some methods, like [2] and [34], aim to identify audio tracks through symbolic queries, but they rely on converting either audio into symbolic or symbolic into audio, respectively. In contrast, [24] proposes a score-informed analysis of audio. Although this approach represents a promising development, it has to be considered a preliminary study, with a small sample size of only 20 violin-only tracks.

3 Research Metodology

The primary objective of this research is to develop algorithms that can accurately measure musical similarity based on both audio and symbolic content. The proposed approach will consider factual musical data and provide an interpretable model for computing music similarity between music pieces.

Dataset creation. To achieve this goal, the first step is to create a multimodal dataset, which includes various types of data for each song in the dataset (c.f. *RQ1*). Specifically, the dataset must consist of four key elements for each track: (i) an audio track, (ii) melodic annotations, (iii) harmonic annotations, and (iv) track metadata.

The dataset will be encoded as a RDF/OWL Knowledge Graph (KG) [7], which will define semantic relationships between the various multimodal elements. The KG will also contain alignment data between different types of annotations, such as audio, melodic and harmonic data.

Similarity computation. Similarity measures based on symbolic data will then be defined (c.f. *RQ2*), focusing on both melodic and harmonic elements.

To achieve this, it is first necessary to define the concept of music similarity both musicologically and perceptually. First, repositories and datasets of known patterns will serve as a basis for the definition of similarity functions. Then, various types of matches, such as exact and fuzzy matches, will be considered between symbolic annotations at different levels, such as phrases, form, cadences, and melodies. This approach enables the investigation of musical similarity from a purely musical perspective, which would allow to the resulting similarity functions to be both explainable and transparent.

The research will also enable the definition of local similarities, allowing for the analysis of influences between different songs, as well as the detection of plagiarism in specific song sections. Moreover, the similarity analysis will be conducted by jointly analysing the harmonic and melodic data to provide more realistic and musicologically grounded similarity information.

Multimodal analysis. In the final step, the similarities extracted from the symbolic data will be used to study similarity on the audio signal. This will involve training deep learning architectures on the aligned audio and symbolic data through the application of data fusion techniques (c.f. *RQ3*). Great care will be given in selecting an architecture that is both explainable and allows for analogies to be drawn between the various components of the multimodal analysis, such as deep learning architectures and neuro-symbolic reasoning.

An architecture that will be explored is transformers [9], which in this context can be employed for the unsupervised matching between symbolic annotations and audio features. Hence, the produced unsupervised model will be fine-tuned using the similarity measures extracted from the symbolic annotations.

3.1 Evaluation

The validation of the results obtained will focus on two main elements: (i) similarity measures based on symbolic content; and (ii) similarity based on audio signals.

Firstly, the similarity measures calculated on symbolic content will be evaluated to determine if the output of the defined similarity functions produces a musicologically or perceptually relevant output. Moreover, known pattern datasets [1, 27] will be used to evaluate the output of the similarity measures. Secondly, crowdsourced surveys will be conducted to gather more data on the perceptual relevance of the extracted similarities.

Regarding the similarities calculated on audio signals, global results will be evaluated on typical music information retrieval tasks, such as cover song detection. For local similarities, the audio extracted similarities will be evaluated using the symbolic-aligned data.

Similarly, we will assess the transparency and explainability of the model. While the explainability of the symbolic similarity models is inherent in their design, the explainability of the model on audio signal will be evaluated by comparing the results to the aligned symbolic annotations.

4 Current Results

As initial contributions to the development of this research project, work was conducted on several fronts, including the creation of a dataset, the study of harmonic similarity, the embedding of harmonic annotations, and the construction of ontologies for modeling musical content.

4.1 Dataset creation

As the first contribution of my research, I focused on the creation of a dataset of harmonic annotations (c.f. RQ1): ChoCo, the largest available *Chord Corpus* [4]. Choco is a large-scale dataset that semantically integrates harmonic data from 18 different sources in various representations and formats (Harte, Leadsheet, Roman numerals, ABC). The corpus leverage JAMS (JSON Annotated Music Specification) [18], a popular data structure for annotations in Music Information Retrieval, to effectively represent a variety of chord-related information (chord, key, mode, etc.) in a uniform way. ChoCo also consists of a converter module that takes care of standardising chord annotations into a single format, the Harte Notation [17]. On top of it, a novel ontology modelling music annotations and involved entities (artists, scores, etc.) has been proposed, and a 30M triple knowledge graph[1] has been built.

The proposed workflow is highly scalable and enables the seamless integration of additional data types, including melodic and structural annotations. Moreover, the Knowledge Graph utilised in ChoCo facilitates the alignment of its annotations with various metadata available on the web, such as MusicBrainz[2] and Discogs[3].

As a result, these resources provide an accurate and distinct reference point for each track, which will allow the identification of the audio recording which refers to the annotations contained in the dataset.

4.2 Studies on Harmonic Similarity

In accordance with the second research question (RQ2), a preliminary investigation into the similarity measures has been conducted.

Based on the limitations found in the state-of-the-art study of harmonic similarity, I worked on LHARP, a *Local Harmonic Agreement of Recurrent Patterns*. LHARP is a measure of harmonic similarity formulated for emphasising shared repeated patterns among two arbitrary symbolic sequences, thereby providing a general framework for the analysis of symbolic streams based on their local structures.

To evaluate the efficacy of LHARP as a method for harmonic similarity, two separate experiments were carried out – each pertaining to a case study that the

[1] ChoCo SPARQL Endpoint: https://polifonia.disi.unibo.it/choco/sparql
[2] MusicBrainz: https://musicbrainz.org/
[3] Discogs: https://www.discogs.com/

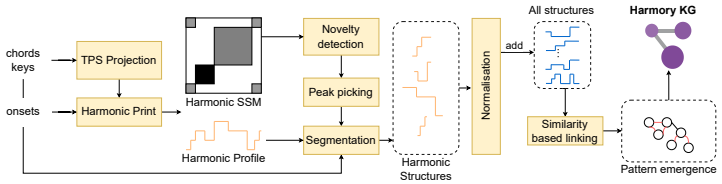

Fig. 1. Workflow used for the production of the Harmonic Memory (Harmory).

function can potentially accommodate. First, a graph analysis was performed to encode harmonic dependencies (edges) between music pieces (nodes) based on their similarity values. Second, to conform with the literature, a cover song detection experiment was conducted.

As an evolution of LHARP, I worked on the *Harmonic Memory* (Harmory) [5]. Harmory is a Knowledge Graph (KG) of harmonic patterns extracted from a large and heterogeneous musical corpus. By leveraging a cognitive model of tonal harmony, chord progressions are segmented into meaningful structures, and patterns emerge from their comparison via harmonic similarity. Akin to a music memory, the KG holds temporal connections between consecutive patterns, as well as salient similarity relationships (c.f. Figure 1).

During the creation of Harmory, I focused on the developement of both harmonic segmentation and harmonic similarity state-of-the-art algorithm.

Digital Signal Processing (DSP) algorithms were used to perform harmonic segmentation on symbolic content. Tonal Pitch Space (TPS) [23] was used to encode the harmonic sequences and generate a Self-Similarity Matrix (SSM) [6], from which a novelty curve was extracted to identify the harmonic segment boundaries [29].

Additionally, a new algorithm for computing harmonic similarity using Dynamic Time Warping (DTW) [31] on TPS-encoded sequences was proposed, which is more efficient than the previous state-of-the-art approach [15].

4.3 Music Chord Embeddings

Another aspect of my work involved the definition of embeddings to enable the expressive encoding of harmonic annotations. To achieve this goal, I developed *pitchclass2vec*, a novel type of embedding that effectively preserves the harmonic characteristics of a chord.

The efficacy of this embedding was evaluated in a Music Structure Analysis task, where it outperformed other approaches, including those based on chord encoding [25] or textual encoding [8].

4.4 Semantic Integration of Musical Data

For the development of the aforementioned works, ontologies were created to model various types of data related to the music domain. These works respond

to RQ1, and aim to provide new methods for the representation of musical knowledge. These ontologies include the *JAMS Ontology* , which models musical notations (such as chords, patterns, and musical structures), the *Roman Chord Ontology* , which models chords expressed in Roman numeral notation, and the *Music Note Ontology* , which models musical notes and their realisation (i.e., the note played in a performance). These ontologies are part of an ontological framework named *Polifonia Ontology Network* (PON).

5 Conclusion and Next Steps

This paper presents a research project that employs a symbolic-informed architecture to study music similarities on audio signals. This allows an explainable and interpretable musically-grounded analysis of similarities in music which can be performed both on symbolic annotations and audio signal.

The use of Knowledge Graphs (KG) and Semantic Web tools is crucial to this research as they provide a foundation for data alignment and interoperability across various data types.

Moving forward, the research will focus on expanding the dataset (as described in Section 4.1) by incorporating new data types, such as melodic data and audio signals, into the knowledge graph. This will facilitate exploration of novel similarity functions that enable the study of symbolic data, integrating diverse musical elements such as melody, harmony, and structure.

Subsequently, the research will aim to align the produced data with audio signals, with the objective of training a model informed by symbolic data that is capable of analysing similarity on audio signals.

Finally, a crucial objective of this study is to extend the ontological models developed to enable multimodal analysis of other data types and in other domains.

References

1. Adegbija, T.: jazznet: A dataset of fundamental piano patterns for music audio machine learning research. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2023)
2. Balke, S., Arifi-Müller, V., Lamprecht, L., Müller, M.: Retrieving audio recordings using musical themes. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 281–285 (Mar 2016)
3. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy (2017)
4. de Berardinis, J., Meroño-Peñuela, A., Poltronieri, A., Presutti, V.: Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs. In: Manuscript under review (2022)
5. de Berardinis, J., Meroño-Peñuela, A., Poltronieri, A., Presutti, V.: The harmonic memory: a knowledge graph of harmonic patterns as a trustworthy framework for computational creativity. In: The Web Conference, to be published (2023)

6. de Berardinis, J., Vamvakaris, M., Cangelosi, A., Coutinho, E.: Unveiling the hierarchical structure of music by multi-resolution community detection. Transactions of the International Society for Music Information Retrieval **3**(1), 82–97 (2020)
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**(3), 1–22 (2009)
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguistics **5**, 135–146 (2017)
9. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization (2020)
10. Corrêa, D.C., Rodrigues, F.A.: A survey on symbolic data-based music genre classification. Expert Systems with Applications **60**, 190–210 (2016)
11. Downie, J.S.: The scientific evaluation of music information retrieval systems: Foundations and future. Comput. Music. J. **28**(2), 12–23 (2004)
12. Du, X., Chen, K., Wang, Z., Zhu, B., Ma, Z.: Bytecover2: Towards dimensionality reduction of latent embedding for efficient cover song identification. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 616–620 (2022)
13. Dörfler, M., Bammer, R., Grill, T.: Inside the spectrogram: Convolutional neural networks in audio processing. In: 2017 International Conference on Sampling Theory and Applications (SampTA). pp. 152–155 (2017)
14. Giraud, M., Groult, R., Leguy, E., Levé, F.: Computational Fugue Analysis. Computer Music Journal **39**(2), 77–96 (2015)
15. de Haas, W.B., Wiering, F., Veltkamp, R.C.: A geometrical distance measure for determining the similarity of musical harmony. International Journal of Multimedia Information Retrieval **2**(3), 189–202 (Sep 2013)
16. Hanna, P., Robine, M., Rocher, T.: An alignment based system for chord sequence retrieval. In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. pp. 101–104 (2009)
17. Harte, C., Sandler, M.B., Abdallah, S.A., Gómez, E.: Symbolic representation of musical chords: A proposed syntax for text annotations. In: ISMIR. vol. 5, pp. 66–71 (2005)
18. Humphrey, E.J., Salamon, J., Nieto, O., Forsyth, J., Bittner, R.M., Bello, J.P.: JAMS: A JSON Annotated Music Specification for Reproducible MIR Research. In: ISMIR. pp. 591–596 (2014)
19. Karydis, I., Lida Kermanidis, K., Sioutas, S., Iliadis, L.: Comparing content and context based similarity for musical data. Neurocomputing **107**, 69–76 (2013), timely Neural Networks Applications in Engineering
20. Knees, P., Schedl, M.: A survey of music similarity and recommendation from music context data. ACM Trans. Multimedia Comput. Commun. Appl. **10**(1) (dec 2013)
21. Kowald, D., Schedl, M., Lex, E.: The unfairness of popularity bias in music recommendation: A reproducibility study. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 35–42. Springer International Publishing, Cham (2020)
22. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: 2008 Seventh International Conference on Machine Learning and Applications. pp. 688–693 (2008)
23. Lerdahl, F.: Tonal pitch space. Music Perception: An Interdisciplinary Journal **5**(3), 315–349 (1988)

24. Li, P.C., Su, L., Yang, Y.H., Su, A.W.Y.: Analysis of expressive musical terms in violin using score-informed and expression-based audio features. In: International Society for Music Information Retrieval Conference (2015)
25. Madjiheurem, S., Qu, L., Walder, C.: Chord2vec: Learning musical chord embeddings. In: Proceedings of the constructive machine learning workshop at 30th conference on neural information processing systems (NIPS2016), Barcelona, Spain (2016)
26. McFee, B., Barrington, L., Lanckriet, G.: Learning content similarity for music recommendation. IEEE Transactions on Audio, Speech, and Language Processing **20**(8), 2207–2218 (2012)
27. Medina, R., Smith, L., Wagner, D.: Content-based indexing of musical scores. In: 2003 Joint Conference on Digital Libraries, 2003. Proceedings. pp. 18–26 (2003)
28. Mehrotra, R.: Algorithmic balancing of familiarity, similarity, & discovery in music recommendations. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. p. 3996–4005. CIKM '21, Association for Computing Machinery, New York, NY, USA (2021)
29. Müller, M.: Fundamentals of music processing: Audio, analysis, algorithms, applications, vol. 5. Springer (2015)
30. Orio, N., Rodà, A.: A measure of melodic similarity based on a graph representation of the music structure. In: Hirata, K., Tzanetakis, G., Yoshii, K. (eds.) Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009. pp. 543–548. International Society for Music Information Retrieval (2009)
31. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing **26**(1), 43–49 (1978)
32. Sheikh Fathollahi, M., Razzazi, F.: Music similarity measurement and recommendation system using convolutional neural networks. International Journal of Multimedia Information Retrieval **10**(1), 43–53 (Mar 2021)
33. Simonetta, F., Ntalampiras, S., Avanzini, F.: Multimodal music information processing and retrieval: Survey and future challenges. In: 2019 International Workshop on Multilayer Music Representation and Processing (MMRP). pp. 10–18 (2019)
34. Suyoto, I.S.H., Uitdenbogerd, A.L., Scholer, F.: Searching musical audio using symbolic queries. IEEE Trans. Audio Speech Lang. Processing **16**(2), 372–381 (Feb 2008)
35. Tan, H.H., Herremans, D.: Music fadernets: Controllable music generation based on high-level features via low-level feature modelling. In: Cumming, J., Lee, J.H., McFee, B., Schedl, M., Devaney, J., McKay, C., Zangerle, E., de Reuse, T. (eds.) Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020. pp. 109–116 (2020)
36. Velardo, V., Vallati, M., Jan, S.: Symbolic Melodic Similarity: State of the Art and Future Challenges. Computer Music Journal **40**(2), 70–83 (06 2016)
37. Vinet, H.: The representation levels of music information. In: Wiil, U.K. (ed.) Computer Music Modeling and Retrieval. pp. 193–209. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
38. Wiggins, G., Miranda, E., Smaill, A., Harris, M.: A framework for the evaluation of music representation systems. Computer Music Journal **17**(3), 31–42 (1993)
39. Zheng, E., Moh, M., Moh, T.S.: Music genre classification: A n-gram based musicological approach. In: 2017 IEEE 7th International Advance Computing Conference (IACC). pp. 671–677 (2017)