A Framework for Music Similarity and Cover Song Identification

Roberto Piassi Passos Bodo*[1], Emmanouil Benetos[2], and Marcelo Queiroz[1]

[1] Institute of Mathematics and Statistics, University of São Paulo, Brazil
{rppbodo,mqz}@ime.usp.br
[2] Centre for Digital Music, Queen Mary University of London, UK
emmanouil.benetos@qmul.ac.uk

Abstract. This paper presents a framework for music information retrieval tasks which relate to music similarity. The framework is based on a pipeline consisting of audio feature extraction, feature aggregation and distance measurements, which generalizes previous work and includes hundreds of similarity models not previously considered in the literature. This general pipeline is subjected to a comprehensive benchmark of analogously defined music similarity models over the task of cover song identification. Experimental results provide scientific evidence for certain preferred combined choices of features, aggregations and distances, while pointing towards novel combinations of such elements with the potential to improve the performance of music similarity models on specific MIR tasks.

1 Introduction

Using Music Information Retrieval (MIR) techniques to deal with large sets of music files has become an increasingly common practice. Working directly with audio and musical contents has several advantages. MIR methods can provide users the ability to hum in order to retrieve a melody or to clap to fetch a rhythm, and to use an audio file as query in a search for similar tracks. The goal of MIR is to make music content more accessible and in a more intuitive way [1].

Music similarity plays a central role in several MIR tasks. It is often desirable to define and calculate similarity measures for pairs of music recordings, based on audio contents and also (derived or annotated) metadata. The use of music similarity measures on a music dataset provides a solid foundation for navigation, organization, recommendation, and search [2,3].

Since there is no universally agreed-upon formalized concept of general musical similarity, a fair solution is to look for similarity models which deal with individual aspects of music, such as pitch, rhythm, dynamics and timbre, providing tools for melodic, harmonic, rhythmic, dynamic and timbre-related retrieval tasks, among others. It is important to state explicitly that the notion of music similarity completely depends on the context of the retrieval task at hand, which is usually established by the type of dataset annotations available.

* This study was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and CNPq Grant 307389/2019-7

The literature on audio-based music similarity presents several approaches, including the use of traditional information retrieval methods, such as extracting features from audio recordings and computing their distances within a vector space [4,5,6,7,8,9], modeling the extracted feature distributions and comparing the corresponding statistical models [10,11,12,13,14,15], feature learning [16,17], metric learning [18,19], and deep neural networks [20,21].

The framework presented in this paper, which generalizes the first two approaches above, is based on a conceptual pipeline [3] that breaks down a generic music similarity model into three components (feature extraction, aggregation and distance computing), completely specified by the choices of techniques employed in each component. Its implementation allows the user to freely combine virtually any techniques within each component, thus providing a direct way of experimenting with a large number of similarity models at once.

This possibility is explored in the context of music similarity tasks, including Cover Song Identification (CSI) [22], an application which involves identifying songs[3] which are versions (covers) of each other, assuming that versions of a song should have some common music trait captured by a music similarity model. This paper presents, to the best of the authors' knowledge, the first attempt to comprehensively benchmark music similarity models in music similarity tasks, where hundreds of models not previously considered in the literature are tested.

The main goal of the experiments here presented is to identify which music similarity models lead to best results for the annotated datasets considered, which have been compiled for melodic similarity tasks, rhythmic similarity tasks, genre classification and CSI. Another contribution of this paper is a modular open-source framework[4] for music similarity offering numerous alternatives for feature extraction, aggregation and distances.

The remainder of the paper is organized as follows: Section 2 presents the music similarity models considered; Section 3 presents the metrics considered to assess the discriminating power of the music similarity models; Section 4 presents the experiments, including the selected datasets, the experimental design, the results and their discussion; Section 5 outlines the conclusions and directions for future work.

2 Music Similarity Framework

The music similarity framework considered here implements the following pipeline: 1. extract audio features; 2. aggregate local features into global features; and 3. compute the similarities of every pair of audio recordings within a dataset. A triple $\{extractor_i, aggregator_j, distance_k\}$ defines a music similarity model, and our main goal is to benchmark music similarity models, identifying which models lead to best results for each annotated dataset. Additionally, the models are also applied to datasets designed for Cover Song Identification (CSI), another similarity-based music retrieval task.

[3] in the CSI literature, *song* is often taken as a synonym of *audio recording*, regardless of containing singing voice or not.

[4] The source code can be found at `https://github.com/rppbodo/music-similarity-framework`.

Papers addressing music similarity related tasks, including CSI, often derive their similarity measurements from tonal features, such as chromagrams [23,24,7,25], tonnetz [26], and symbolic melodic sequences [27,28,29,30,31,32]. Also used in music similarity retrieval tasks are timbre features (e.g., Mel-Frequency Cepstral Coefficients (MFCC) [4,10,11,13,12,15]), spectral features (e.g., spectral centroid, bandwidth, contrast, flatness, etc) [5], rhythmic features (e.g., Rhythm Pattern) [33,6], and amplitude/energy features (e.g., Root Mean Square (RMS)) [34,35].

Among aggregation methods applied in music similarity are simple statistics (such as mean, standard deviation, skewness and kurtosis, see e.g. [15]), computed from the features themselves and their 1st order differences, Gaussian Mixture Models [11,12,13], Vector Quantization [10], Markov Chains [36], Octave and Interval Abstractions [32], and Pitch Contour (using 3-levels [27] and 5-levels [31]).

The computation of the similarity between two audio recordings is based on a chosen distance applied to the (possibly aggregated) features. Distances relevant for music similarity are Manhattan [6,8], Euclidean [4,5,7], Cosine [4,9], Longest Common Subsequence based distances [37,38], Levenshtein [39,38], Kullback-Leibler [13,15], Earth Mover [10,14], and Monte Carlo distances [11,12].

The detailed analysis of the techniques proposed in the music similarity literature allows us to observe that several papers do not explicitly argue as for why a particular extractor (or aggregator, or distance) is selected to solve a particular problem. Even less frequent are arguments about why a specific set of techniques are used in combination (instead of many other plausible alternatives). This prompted us to try to explore hundreds of combinations of extractors, aggregators, and distances that are not considered in the literature. It was thus natural to look at this problem as a benchmark, exhaustively experimenting with a large number of music similarity models.

The current list of music similarity models considered in the implemented framework started out from a large set of features, aggregators and distances appearing in the related literature, which has been modified by including and collecting techniques, but also by discarding techniques by many criteria, including the availability of open-source implementations. The rationale for this specific criterion is to avoid producing implementations that might substantially differ from their original implementations due to ambiguous or insufficiently detailed descriptions. A survey of open-source libraries (such as LibROSA[5], Essentia[6], and RP_extract[7]) led us to include techniques not previously considered in the music similarity literature. The same criteria were applied to aggregator and distance techniques, but in a softer way, since they are usually much simpler to implement.

Due to compatibility issues, not all available features, aggregators and distances can be combined. Framewise numerical features may be aggregated using any statistical aggregation methods, GMM and VQ. Symbolic melodic features use only specific aggregators (octave/interval abstractions, pitch contours and Markov chains). Numerical aggregations (single and multivariate Gaussians, GMM, vector quantization, and Markov chains) can be compared using spatial distances (Euclidean, Manhattan, Chebyshev,

[5] `http://librosa.github.io/`
[6] `http://essentia.upf.edu/`
[7] `https://github.com/tuwien-musicir/rp_extract`

and cosine). Statistical models can be compared using Kullback Leibler, Earth Mover's and Monte Carlo distances, and all symbolic global features can be compared using LCS-based and Levenshtein distances.

All the compatible combinations of features, aggregators, and distances considered result in a total of 690 music similarity models; the complete list is available at `https://rppbodo.github.io/phd/music-similarity-models.html`, along with descriptions of each function.

3 Music Similarity and Cover Song Identification metrics

The most common way to represent a particular music similarity model applied to a particular dataset is the similarity matrix. The i, j position of this matrix contains the similarity between the i-th and the j-th tracks in the dataset. It can be defined from a normalized distance measure as $\text{sim}(t_i, t_j) = 1 - \text{dist}(t_i, t_j)$.

Intra-Inter Class Similarity Ratios (IICSR) When the dataset partitions its recordings into labeled classes (e.g. genres, composers, melodic or rhythmic patterns), we may define the quality of a music similarity model using the intra-inter-class similarity ratio, computed from the similarity matrix according to the following formula:

$$IICSR(c) = \frac{\frac{\sum_{t_1 \in T_c} \sum_{t_2 \in T_c, t_1 \neq t_2} sim(t_1,t_2)}{(|T_c|^2 + |T_c|)}}{\frac{\sum_{t_1 \in T_c} \sum_{t_2 \in T_{C \setminus c}} sim(t_1,t_2)}{(|T_{C \setminus c}||T_c|)}}, \tag{1}$$

where T_c is the set of all recordings in class c and $T_{C \setminus c}$ is the complement of T_c. This measure compares the average similarity within the class c (weighted by the number of recordings in this class) with the average similarity for pairs of recordings in different classes (with one member of the pair in class c). If these ratios are greater than 1, the similarity model may be used to classify pairs of recordings as belonging to the same class or to different classes. Intra-inter-class similarity ratios may be summarized by their weighted average:

$$weighted\_mean\_IICSR = \frac{1}{\sum_{c \in C} |T_c|} \sum_{c \in C} IICSR(c) \times |T_c|, \tag{2}$$

where each class is weighted by its size (number of recordings).

Mean Rank (MR) The Mean Rank is broadly used in the CSI literature [9,40], where queries return ranked lists of cover candidates. MR corresponds to the average position (rank) where the first cover appears in the resulting list.

Mean Reciprocal Rank (MRR) The reciprocal rank (inverse of a rank) [41] converts index positions to the $[0, 1]$ range, where higher values represent covers higher up in the list (topmost ranks). MRR corresponds to the average of the reciprocal ranks, and its inverse may be viewed as the harmonic mean of the original ranks.

dataset	n_{tracks}	$n_{classes}$	annotations
Ballroom	698	10	dance styles names
GTZAN	1000	10	musical genres
IOACAS-QBH	1057	298	ground-truth melody id
Panteli's melody dataset	3000	30	original melody id
Panteli's rhythm dataset	3000	30	original rhythm id
MAST	3104	40	ground-truth rhythm id
1517-Artists	3180	19	musical genres
MIR-QBSH	4479	48	ground-truth melody id
FMA-Small	8000	8	musical genres
Covers80	160	80	original song
YouTubeCovers	350	50	original song
Covers1000	1000	395	original song
Mazurkas	2741	49	mazurka id
SHS9K	9286	143	original song

Table 1. Datasets selected to experiment Music Similarity models.

Median Rank (MDR) The MDR is a robust statistic based on the positions of the first retrieved cover, obtained as the median of the ranks for all queries.

Mean Average Precision (MAP) Kim Falk [42] defines Mean Average Precision in the context of recommender systems, in which users perform queries, and each query returns a list of ranked items. Precision at K (P@k) is the number of relevant items found in the first k items; Average Precision (AP) = $\frac{1}{m}\sum_{k=1}^{m} P@k(u)$, where m is the length of the ranked list, and u is the user performing the query; Mean Average Precision (MAP) = $\frac{1}{|U|}\sum_{u\in U} AP(u)$, where U is the set of all users.

4 Experiments and Results

4.1 Datasets

In this Section we present the datasets used to benchmark models within our music similarity framework. The first part of Table 1 presents datasets designed for various music similarity tasks, and the second part presents the datasets designed specifically for Cover Song Identification.

Three datasets are designed for melodic similarity tasks: MIQ-QBSH[8] and IOACAS-QBH[9] are designed for the query-by-humming task (classes are composed of a reference melody and a set of recordings of people trying to hum it), and Maria Panteli's melody dataset[10] uses synthesis to test similarity models against several melodic transformations.

[8] http://mirlab.org/dataSet/public/MIR-QBSH-corpus.rar
[9] http://mirlab.org/dataSet/public/IOACAS_QBH.rar
[10] https://archive.org/details/panteli_maria_melody_dataset

Three other datasets – Ballroom[11], MAST[12], and Maria Panteli's rhythm dataset[13] — are designed for tasks related to rhythm similarity. The Ballroom dataset is composed of recordings from distinct dance styles; MAST has recordings of students successfully reproducing rhythmic patterns; Maria Panteli's rhythm dataset is composed of different synthesized rhythms subjected to several transformations.

The three remaining datasets in the first part of Table 1 — GTZAN[14], 1517-Artists[15], and FMA-Small[16] -– are annotated with music genres assigned to each recording. Several papers in the literature claim that there is a relationship between genre and timbre [43,44,3], and under this assumption, these datasets could be used to test timbre similarity models.

The second part of Table 1 presents datasets designed for CSI: Covers80[17], YouTube-Covers[18], Covers1000[19], Mazurkas[20], and SHS9K[21]. The latter is a sub-set of the SHS100K[22] dataset crafted by the authors by selecting the original songs that have from 50 to 100 covers.

4.2 Experiment Design

Two experiments are proposed. The goal of the first experiment is to check which music similarity models lead to best results for the selected datasets. In order to accomplish this we run each one of the 9 datasets considered through our music similarity framework, compute the Intra-Inter Class Similarity Ratio (IICSR) for every annotated class within the dataset, and finally compute the weighted mean IICSR for each one of the 690 considered models.

The second experiment has a similar goal to the previous one – to check which music similarity models lead to the best results – but now with CSI datasets considering the specific metrics used in this task. We compute similarity matrices using all 690 models for the 5 CSI datasets, and then calculate the Mean Rank (MR), Mean Reciprocal Rank (MRR), Median Rank (MDR), and Mean Average Precision (MAP).

4.3 Results

The results of the first experiment are organized as follows: the best weighted mean IICSR values for each dataset are presented in Table 2, and the entire list of IICSR values computed in this experiment is published in `https://rppbodo.github.io/phd/experiment_1.html`.

[11] `http://mtg.upf.edu/ismir2004/contest/tempoContest/node5.html`
[12] `https://zenodo.org/record/2620357`
[13] `https://archive.org/details/panteli_maria_rhythm_dataset`
[14] `http://marsyas.info/downloads/datasets.html`
[15] `http://www.seyerlehner.info/index.php?p=1_3_Download`
[16] `https://github.com/mdeff/fma/`
[17] `https://labrosa.ee.columbia.edu/projects/coversongs/covers80/`
[18] `https://sites.google.com/site/ismir2015shapelets/data`
[19] `http://www.covers1000.net/`
[20] `http://www.mazurka.org.uk/`
[21] `https://rppbodo.github.io/phd/shs9k.html`
[22] `https://github.com/NovaFrost/SHS100K`

dataset	mean IICSR	extractor	aggregator	distance
Ballroom	1.33824	spectral_bandwidth	vector_quant.	cosine
GTZAN	1.74945	spectral_contrast	vector_quant.	cosine
IOACAS-QBH	1.13599	pitch_cont._seg.	octave_abst.	lcs_circular_min
Panteli's melody	3.12167	pitch_cont._seg.	interval_abst.	levenshtein_max
Panteli's rhythm	2.82936	chroma_cens	vector_quant.	manhattan
MAST	1.47572	mfcc	vector_quant.	cosine
1517-Artists	1.21584	mfcc	vector_quant.	cosine
MIR-QBSH	1.29421	pitch_cont._seg.	octave_abst.	levenshtein_circular_max
FMA-Small	1.21128	mfcc	vector_quant._default	cosine

Table 2. Results obtained with Music Similarity datasets.

dataset	MR	MRR	MDR	MAP	extractor	aggregator	distance
Covers80	41.575	0.19359	31.0	0.19359	chroma_stft	diff_stats_1	cosine_oti
YouTubeCovers	7.97143	0.6942	1.0	0.36114	pitch_cont._seg.	octave_abst.	lcs_circular_mean
Covers1000	144.041	0.25731	35.0	0.19159	pitch_cont._seg.	octave_abst.	lcs_circular_mean
Mazurkas	4.15724	0.95774	1.0	0.82286	pitch_cont._seg.	octave_abst.	levenshtein_circular_max
SHS9K	47.57883	0.40387	6.0	0.05102	pitch_cont._seg.	octave_abst.	lcs_circular_mean

Table 3. Results obtained with Cover Song Identification datasets.

The results of the second experiment are displayed as follows: the best models for each dataset are presented in Table 3, and the entire list of metrics computed in this experiment is published in `https://rppbodo.github.io/phd/experiment_2.html`.

4.4 Discussion

Analysing the models that achieved the best weighted mean IICSR for MIQ-QBSH, IOACAS-QBH, and Maria Panteli's melody dataset, it is possible to verify that all of them have Pitch Contour Segmentation as their feature, which matches the hypothesis that melodic features lead to better results for melodic datasets. Regarding the aggregators, two models have Octave Abstraction and one has Interval Abstraction. This is somehow expected, since the alternative abstractions (3-level and 5-level Pitch Contours) are relatively weaker due to their simplistic representations of the original pitch sequences.

The models that performed best for the Ballroom, MAST, and Maria Panteli's rhythm dataset are relatively surprising, not only because none of the features are specifically designed for rhythmic similarity tasks, but also because they are very different from each other: Spectral Bandwidth is related to the spectrum spread, MFCC is usually associated with timbre, and Chroma Energy Normalized Statistics (CENS) is a tonal feature.

Regarding the highest weighted mean IICSR obtained for GTZAN, 1517-Artists, and FMA-Small, two out of three best performing models have MFCC as their feature, and the other one has Spectral Contrast. MFCC is a feature usually related to timbre, so this matches the initial hypothesis. According to Jiang et al. [45], Spectral Contrast is

reported to have a better discriminating power for different music types than MFCC, so it is noteworthy that this feature has also emerged here.

The best models that lead to the lowest Mean Rank values for the five CSI datasets are shown in Table 3. Four models out of five share the same feature (Pitch Contour Segmentation) and the same aggregator (Octave Abstraction), which is a very good indication of the relevance of these methods, while the remaining model uses a Chromagram as feature. All features from the best models encode tonal information, which matches the observation in the literature that tonal differences are the less frequent between versions [22,46,47].

5 Conclusions

In this paper we introduced a modular music similarity framework designed to benchmark 690 music similarity models applied to specific music information retrieval tasks. Our experiments compared these models under several datasets compiled for tasks requiring different music similarity perspectives, showing that the choices of features, aggregators and distances not only have a significant impact on the performance of the corresponding models, but also that many useful techniques and combinations have been largely overlooked by the music similarity literature, corroborating the importance of comparative studies such as the present one.

As future work, we consider expanding the lists of features (HPCP, crema-PCP, Onset Patterns, Scale Transform, Pitch Bihistogram, Intervalgram, etc), aggregators (Dynamic Time Warping (DTW), Self-Organizing Map (SOM), vector quantization using tree-based clustering, n-grams, etc), and distances (Mahalanobis, Jensen-Shannon, Smith- Waterman, Mongeau-Sankoff, etc) in the music similarity framework, as well as incorporating alternative approaches to music similarity that not necessarily follow the current pipeline, such as feature learning [16,17], metric learning [18,19], and deep neural networks [20,21].

References

1. J Stephen Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.
2. Meinard Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.
3. Peter Knees and Markus Schedl. *Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies*. Springer, 2016.
4. Jonathan T Foote. Content-based retrieval of music and audio. In *Multimedia Storage and Archiving Systems II*, volume 3229, pages 138–147. International Society for Optics and Photonics, 1997.
5. Tao Li and Mitsunori Ogihara. Content-based music similarity search and emotion detection. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–705. IEEE, 2004.
6. Klaus Seyerlehner, Markus Schedl, Tim Pohle, and Peter Knees. Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX*, 2010, 2010.

7. Xiaoqing Yu, Jing Zhang, Junwei Liu, Wanggen Wan, and Wei Yang. An audio retrieval method based on chromagram and distance metrics. In *2010 International Conference on Audio, Language and Image Processing*, pages 425–428. IEEE, 2010.
8. Peter Knees and Markus Schedl. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):1–21, 2013.
9. Christopher J Tralie. Early mfcc and hpcp fusion for robust cover song identification. *arXiv preprint arXiv:1707.04680*, 2017.
10. Beth Logan and Ariel Salomon. A music similarity function based on signal analysis. In *ICME*, pages 22–25, 2001.
11. Jean-Julien Aucouturier, Francois Pachet, et al. Music similarity measures: What's the use? In *ISMIR*, pages 13–17, 2002.
12. Francois Pachet and Jean-Julien Aucouturier. Improving timbre similarity: How high is the sky. *Journal of negative results in speech and audio sciences*, 1(1):1–13, 2004.
13. Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.
14. Rainer Typke, Frans Wiering, and Remco C Veltkamp. Evaluating the earth mover's distance for measuring symbolic melodic similarity. In *MIREX-ISMIR 2005: 6th International Conference on Music Information Retrieval*. Citeseer, 2005.
15. Dominik Schnitzer, Arthur Flexer, and Gerhard Widmer. A fast audio similarity retrieval method for millions of music tracks. *Multimedia Tools and Applications*, 58(1):23–40, 2012.
16. Maria Panteli, Emmanouil Benetos, Simon Dixon, et al. Learning a feature space for similarity in world music. ISMIR, 2016.
17. Zhesong Yu, Xiaoshuo Xu, Xiaoou Chen, and Deshun Yang. Learning a representation for cover song identification using convolutional neural network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 541–545. IEEE, 2020.
18. Malcolm Slaney, Kilian Weinberger, and William White. Learning a metric for music similarity. In *International Symposium on Music Information Retrieval (ISMIR)*, volume 148, 2008.
19. Hoon Heo, Hyunwoo J Kim, Wan Soo Kim, and Kyogu Lee. Cover song identification with metric learning using distance as a feature. In *ISMIR*, pages 628–634, 2017.
20. Manan Mehta, Anmol Sajnani, and Radhika Chapaneri. Cover song identification with pairwise cross-similarity matrix using deep learning. In *2019 IEEE Bombay Section Signature Conference (IBSSC)*, pages 1–5. IEEE, 2019.
21. Mohamadreza Sheikh Fathollahi and Farbod Razzazi. Music similarity measurement and recommendation system using convolutional neural networks. *International Journal of Multimedia Information Retrieval*, 10(1):43–53, 2021.
22. Joan Serra, Emilia Gomez, and Perfecto Herrera. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. pages 307–332, 2010.
23. Jesper Hojvang Jensen, Mads G Christensen, Daniel PW Ellis, and Soren Holdt Jensen. A tempo-insensitive distance measure for cover song identification based on chroma features. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2209–2212. IEEE, 2008.
24. Joan Serrà, Emilia Gómez, Perfecto Herrera, and Xavier Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1138–1151, 2008.
25. Suman Ravuri and Daniel PW Ellis. Cover song detection: from high scores to general classification. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 65–68. IEEE, 2010.

26. Dmitri Tymoczko. Three conceptions of musical distance. In *International Conference on Mathematics and Computation in Music*, pages 258–272. Springer, 2009.
27. Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of the third ACM international conference on Multimedia*, pages 231–236, 1995.
28. Wei-Ho Tsai, Hung-Ming Yu, Hsin-Min Wang, et al. Query-by-example technique for retrieving cover versions of popular songs with similar melodies. In *ISMIR*, volume 5, pages 183–190, 2005.
29. Klaus Frieler and Daniel Müllensiefen. The simile algorithm for melodic similarity. *Proceedings of the Annual Music Information Retrieval Evaluation exchange*, 2005.
30. Matija Marolt. A mid-level melody-based representation for calculating audio similarity. In *ISMIR*, pages 280–285, 2006.
31. Seungmin Rho and Eenjun Hwang. Fmf: Query adaptive melody retrieval system. *Journal of Systems and Software*, 79(1):43–56, 2006.
32. Justin Salamon, Joan Serra, and Emilia Gómez. Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, 2(1):45–58, 2013.
33. Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organization and visualization of music archives. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 570–579, 2002.
34. Costas Panagiotakis and Georgios Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on multimedia*, 7(1):155–166, 2005.
35. Cory McKay and I Fujinaga. Automatic music classification and similarity analysis. In *International Conference on Music Information Retrieval*. Citeseer, 2005.
36. Holger H Hoos, Kai Renz, and Marko Görg. Guido/mir-an experimental musical information retrieval system based on guido music notation. In *ISMIR*, pages 41–50, 2001.
37. Alexandra Uitdenbogerd and Justin Zobel. Melodic matching techniques for large music databases. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 57–66, 1999.
38. Matthew Kelly. *Evaluation of melody similarity measures*. PhD thesis, 2012.
39. Kjell Lemström and Esko Ukkonen. Including interval encoding into edit distance based music comparison and retrieval. In *Proc. AISB*, pages 53–60. Citeseer, 2000.
40. Marc Sarfati, Anthony Hu, and Jonathan Donier. Ensemble-based cover song detection. *arXiv preprint arXiv:1905.11700*, 2019.
41. J Stephen Downie, Mert Bay, Andreas F Ehmann, and M Cameron Jones. Audio cover song identification: Mirex 2006-2007 results and analyses. In *ISMIR*, pages 468–474, 2008.
42. Kim Falk. *Practical recommender systems*. Manning Publications, 2019.
43. George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals. In *Proceedings of the 2nd international symposium on music information retrieval, Indiana*, 2001.
44. Jean-Julien Aucouturier and Francois Pachet. Representing musical genre: A state of the art. *Journal of new music research*, 32(1):83–93, 2003.
45. Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, pages 113–116. IEEE, 2002.
46. Justin Salamon, Joan Serrá, and Emilia Gómez. Melody, bass line, and harmony representations for music version identification. In *Proceedings of the 21st International Conference on World Wide Web*, pages 887–894, 2012.
47. Ning Chen, Mingyu Li, and Haidong Xiao. Two-layer similarity fusion model for cover song identification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1):1–15, 2017.