Maximilian Eibl, Martin Gaedke. (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 139

Soundslike – Automatic content-based music annotation and recommendation for large databases

Sascha Grollmisch[1], Hanna Lukashevich[2]

Abstract: A manual indexing of large music libraries is both tedious and costly, that is why a lot of music datasets are incomplete or wrongly annotated. An automatic content-based annotation and recommendation system for music recordings is independent of originally available metadata. It allows for generating an objective metadata that can complement manual expert annotations. These metadata can be effectively used for navigation and search in large music databases of broadcasting stations, streaming services, or online music archives. Automatically determined similar music pieces can serve for user-centered playlist creation and recommendation. In this paper we propose a combined approach to automatic music annotation and similarity search based on musically relevant low-level and mid-level descriptors. First, we use machine learning to infer the high-level metadata categories like genre, emotion, and perceived tempo. These descriptors are then used for similarity search. The similarity criteria can be individually weighted and adapted specifically to specific user requirements and musical facets as rhythm or harmony. The proposed method on music annotation is evaluated on an expert-annotated dataset reaching average accuracies of 60% to 90%, depending on a metadata category. An evaluation for the music recommendation is conducted for different similarity criteria showing good results for rhythm and tempo similarity with precision of 0.51 and 0.71 respectively.

Keywords: automatic music classification, music annotation, music recommendation, music similarity search, music information retrieval

1 Introduction

The amount of digital audio files is constantly growing and retrieval of specific music recordings gets harder with every file added to the collection. Therefore extensive annotations of different categories like genre or emotion are required for efficiently searching the database. These annotations are traditionally collected from the music labels or added by experts that listen to the song. This is both tedious and costly and nearly impossible when creating huge databases from scratch. Another approach is to extract these tags automatically from the digital music recording itself. This enables the indexing of huge amounts of audio files in a comparable short amount of time. After all songs are labeled users often require to search the database for similar songs. A standard use-case is the creation of playlists according to certain criteria. Other methods compare metadata or information on listening habits from users to obtain new playlists. These methods require a large user community and multitude

[1] Fraunhofer Institute for Digital Media Technology, Ehrenbergstr. 31, 98693 Ilmenau, goh@idmt.fraunhofer.de
[2] Fraunhofer Institute for Digital Media Technology, Ehrenbergstr. 31, 98693 Ilmenau, lkh@idmt.fraunhofer.de

doi:10.18420/in2017_09

of categories for getting reasonable results. The content-based method directly analyzes the music recording for finding similar songs without having to fulfill the requirements of the data-driven approach.

In this paper, we describe the state of the art for automatic music annotation and recommendation, propose an approach and an evaluation for both tasks. Finally, we present the "Soundslike" system that combines the automatic annotation of large amounts of music recordings enriched with the possibility of recommendations based on metadata filters and musical similarity.

2 Music annotation

The following section first describes the state of the art on the field of automatic content-based music annotation. Afterwards we name the categories and classes within the proposed system. Finally, we describe and evaluate the system for the automatic music annotation.

2.1 State of the Art

Various music annotation tasks are approached with algorithms from the field of Music Information Retrieval (MIR) algorithms [Ca08]. Current algorithms combine acoustic features and apply machine learning methods for classification such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM), or deep neural networks (DNN) to automatically label a given song w.r.t. the music genre (e.g., pop, rock, jazz), the texture (e.g., hard and soft), or other categories. Automatic detection of music emotions was initially performed using categorical labels such as joyful, happy, quiet, and dark, and later performed w.r.t. the emotional dimensions valence (happy or sad) and arousal (calm or excited) [YC12]. Various feature sets have been used in the literature ranging from dynamic, rhythmic, harmonic, and tonal features. Newer studies showed that data-driven methods using Recurrent Neural Networks (RNN) outperform previous approaches for tempo detection on a wide range of musical genres [BKW15]. Until today, the main challenge is the subjectivity of ground truth annotations as well as the complexity and multi-dimensionality of the problem at hand [Fl14, Ba16]. For an overview of different computational approaches, we refer to the literature, e.g., [KS13] and [Mü15].

2.2 Categories

This section describes the categories and possible class labels which are automatically extracted by the proposed system. These definitions are also used for the annotations of the evaluation dataset and should mainly give an intuitive understanding of the categories.

Genre: Describes the style of the music. There are many possible classification approaches, e.g., by the instruments used or the country of origin. In this paper, we focus on ten popular styles of western music – Classical, Country, Electronica, Jazz, Latin, Pop, Rap, Rock, Schlager[3], and Soul.

Valence: General mood of the musical piece – High (happy) or low (sad).

Arousal: Describes the excitement of the music – High (energetic) or low (relaxing).

Emotion: Combines arousal and valence to form four possible values – Anxious (low valence with high arousal), depressed (low valence with low arousal), exuberant (high valence with high arousal) and content (high arousal with low valence).

Perceived Tempo: The perceived tempo of the music is not necessarily related to the actual tempo (BPM). Depending on the chosen rhythm, a song can be perceived from very slow to very fast in five steps.

Texture: The perceived hardness/edginess of the music – Hard or soft.

Instrumental Density: The perceived density of the music which is influenced by the quantity of used instruments, as well as its production – Full or sparse.

Distortion: The amount of distortion in the music which can be compared to the overdrive of electric guitars but can also be achieved by vocals or other instruments in four steps from clean to extreme.

Dynamic: Describes amount of dynamic changes (loud/quiet) in the musical piece – Changing or continuous.

Percussive: Describes if percussive instruments like drums are used – Non-percussive or percussive.

Synthetic: Differentiate the music by the kind of instruments being used – Acoustic, electro-acoustic or synthetic.

Key: Musical key of a song. Corresponding Major and Minor scales are combined since these share the same musical notes.

Beats per Minute (BPM): Shows the number of quarter notes (in 4/4, 3/4, etc.) or dotted quarter notes (in 6/8, 12/8, etc.) per minute and is the actually measured tempo of a song.

2.3 Proposed System

The proposed system focuses on the extraction of the described categories from the raw audio data of the music recordings. For each category, many training pieces were labeled

[3] Simple catchy music with focus on vocals and German lyrics.

and a set of low- and mid-level features was extracted. Afterwards feature selection methods and feature space transformation techniques were applied to train supervised classifiers separately for each of the categories.

We utilize a broad palette of low-level acoustic features and several mid-level representations [BP05] [Pe04]. To facilitate an overview, the audio features are subdivided into three categories covering the timbral, rhythmic, and tonal aspects of sound.

Timbral Features
Although the concept of timbre is still not clearly defined with respect to music recordings, it has proved to be very useful for automatic music classification [Le12]. To capture timbral information, we use Mel-Frequency Cepstral Coefficients, the Audio Spectrum Centroid, the Spectral Flatness Measure, the Spectral Crest Factor, and the Zero-Crossing Rate. In addition, modulation spectral features [AS03] are extracted from the aforementioned features to capture their short term dynamics. We applied a cepstral low-pass filtering to the modulation coefficients to reduce their dimensionality and decorrelate them as described in [DBG07].

Rhythmic Features
All rhythmic features used in the current setup are derived from the energy slope in excerpts of the different frequency-bands of the Audio Spectrum Envelope feature. These comprise the Percussiveness [UH03] and the Envelope Cross-Correlation (ECC). Further mid-level features are derived from the Auto-Correlation Function (ACF) [DBG07]. In the ACF, rhythmic periodicities are emphasized and phase differences are annulled. We also compute the ACF Cross-Correlation (ACFCC). In addition, the log-lag ACF and its descriptive statistics are extracted according to [GDG09].

Tonal Features
Tonality descriptors are computed from a Chromagram based on Enhanced Pitch Class Profiles (EPCP) [Le06]. The EPCP undergoes a statistical tuning estimation and correction to account for tuning deviations. Pitch-space representations as described in [Ga07] are derived from the Chromagram as mid-level features. Their usefulness for audio description has been shown in [GD09].

Machine Learning
The described features are extracted from the music recordings and collected in feature vectors. Each music recording is now represented by a feature matrix, containing a set of high dimensional feature vectors per time frame. While temporal changes in one feature often correspond to temporal changes in the other feature (for instance, timbre is changing along with loudness), the individual dimensions of feature vectors can often be strongly correlated or/and cause information redundancy. Such raw feature vectors might cause various problems on classification stage and need to be treated properly. Therefore, we first apply the feature selection (FS) algorithm Inertia Ratio Maximization using Feature

Space Projection (IRMFSP) as proposed by Peeters and Rodet [PR03]. In addition, we use feature transformation techniques (FST) as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [Fu90]. Finally, we train either a Support Vector Machine (SVM) or a Gaussian Mixture Model (GMM) supervised classifier. The optimal choice of the FS, the FST, and the classifier is made for each of the categories individually by choosing the combination with the highest accuracy during the training and cross-validation.

2.4 Evaluation

For evaluation, an additional data set was created and only used for testing the proposed system. In other words the performance of the classification was measured on unseen data. The test set includes a total of 100 songs from all genres and musical eras. The selection was balanced by genres and no artist was selected more than once. Since some categories can change during a song, only snippets of 30 seconds were taken. As already mentioned, most of the categories are subjective and will be labeled differently by each annotator. To lower the influence of subjectivity, each song had been annotated by 3 expert listeners and the majority vote was chosen if there were differences between the annotators. These differences already between the expert listeners have to be kept in mind when inspecting the evaluation results. Therefore a perfect accuracy of 100% can hardly be achieved for most of the categories. Table 1 shows the accuracy (percentage of correct labeled songs) and the baseline that a random classifier would achieve on each category.

Category	Accuracy in %	Baseline in %
Genre	80	10
Valence	70	50
Arousal	75	50
Emotion	40	25
Perceived Tempo	47	20
Texture	73	50
Instrumental Density	71	50
Distortion	62	25
Dynamic	80	50
Percussive	75	50
Synthetic	65	33.33
Key	63	8.33
BPM	63 / 87	– / 1.43

Tab. 1: Evaluation results for proposed metadata categories.

The first result for BPM was for the tempo in the full range over all possible octaves. During annotation it was observed that some songs might be perceived in half or double time. Therefore a second evaluation was conducted where all tempos have been mapped to one

octave from 70 to 140 BPM. This led to a huge improvement since it eliminated ambiguity in the annotation.

All classifiers performed better than a random baseline. Very subjective categories like Emotion and Perceived Tempo were harder to classify than well defined categories like Key. The difficulties expert annotators had are also reflected in the automatic extracted labels and the results show that the extracted categories add additional information to existing metadata or databases without any metadata.

3 Music recommendation

In this section we overview the state of the art in content-based music similarity analysis. Furthermore, we propose a usage of musically-motivated similarity profiles. Finally, the proposed system is described and evaluated with respect to those similarity profiles.

3.1 State of the art

Music similarity can refer to different attributes of a music recording such as timbre, tempo, rhythm, melody, or harmony. Most content-based musical similarity algorithms represent audio recordings using several audio feature representations. Traditionally, these features relate to low-level spectral properties, such as the spectral centroid or spectral flux [Pe04]. By incorporating additional knowledge about human auditory perception mechanisms, mid-level features, such as Mel-frequency Cepstral Coefficients (MFCC), which originated from speech recognition, were used to represent the timbre of audio recordings. These features relate to smaller time scales from 10 ms to several seconds. Also, the temporal and rhythmic structure of music pieces is analyzed using Fluctuation Patterns [Po09]. In order to measure the similarity between pairs of songs, these frame-level features are grouped to a song-level time-scale using factor analysis methods such as i-vector analysis [Eg15] or supervector representations based on a Universal Background Model (UBM) using Gaussian Mixture Models (GMM) [Ch11]. Nowadays, automatic feature learning methods based on deep neural networks (DNN) and convolutional neural networks (CNN) have shown to outperform hand-crafted audio features in many content-based audio analysis tasks [HBL13]. Measuring the song-wise similarity across large databases allows to automatically generate playlists [BJ15] while song transitions appear smoother between similar tracks.

One of the main remaining challenges is a proper evaluation of the music similarity systems. Defining musical similarity directly is extremely challenging as myriad features play some role (e.g., cultural, emotional, timbral, rhythmic) [Mc12].

3.2 Proposed system

To obtain the similarity between two music pieces, we first extract acoustic features in the similar manner as for the automatic music annotation, see Section 2.3. For each of the features we calculate the similarity between music pieces based on the chosen similarity measure, e.g., Manhattan distance, Euclidean distance, Kullback-Leibler divergence [KL51], or other [LDB08]. The choice of the similarity distance is set in the similarity profile. The similarity lists obtained for distinct features are aggregated implementing the Borda's method [Dw01] [Bo84].

In our system we predefine the following four similarity profiles:*Timbre* is the most general similarity profile based on the low-level timbral features. *Harmony* similarity profile is based on the chromagram and EPCP tonal features. *Rhythm* similarity profile uses rhythmical feature derived from the ACF. *Tempo* similarity profile directly uses the extracted tempo in beats per minute.

3.3 Evaluation

Evaluating music similarity is an extremely challenging task known to be highly subjective. In this paper, we aim for an objective evaluation of proposed similarity profiles and thus do not perform a user study.

The evaluation dataset is compiled with 7109 audio snippets originating from the MAGIX[4] Soundpool collections. These recordings are professionally produced sounds and music tracks destined for professional and hobby music production. The selected audio snippets are organized in 9 collections within the following three genres: Electro, HipHop, and RockPop with 3 collections per genre. All audio snippets are annotated with the information on the type of sound (Bass, Backbeats, Brass, Drums, Audio Effect, Guitars, Keys, Mallets, Pads, Percussion, Pianos, Sequence, Scratches, Strings, Synths, Vibes, Vocals, Winds). For all harmonic sounds, there are 6–7 versions in several musically related keys available (mostly C major, D minor, E minor, F major, G major, A minor, and optionally B minor). All audio snippets within one collection have the same BPM (90, 100, 125 or 160 bpm). Summarized, for each audio sample (usually 7-15 seconds long) we have the information about its genre, key, tempo, and type of sound. Note, that several of this annotations (e.g., a genre of an audio effect sound) could only be vaguely defined, as the same sound could also match another other genre or key.

We obtain similarity lists for all audio samples in the dataset according to the four similarity profiles and evaluate the percentage (precision@10) of items in the lists sharing the same tempo, key, or genre. The results of the evaluation are presented in Table 2. Here, the last column in this table shows the result for a randomized similarity list for a particular evaluation criterion.

[4] http://www.magix.com

Similarity profile	Evaluation criteria	Precision@10	Baseline	Labels
Timbre	Genre	0.47	0.33	3
Rhythm	Genre	0.51	0.33	3
Tempo	Genre	0.71	0.33	3
Harmony	Genre	0.39	0.33	3
Tempo	Tempo	0.67	0.25	4
Harmony	Key	0.46	0.15	8

Tab. 2: Results of the similarity evaluation

While using genre information as an evaluation criterion for music similarity we observe the following results. The Rhythm and Tempo similarity profiles are having 51% and 71% samples sharing the same genre as the query sample within the similarity results lists. As Timbre similarity profile is mostly sensitive to the type of sound (i.e. music instrument) and the selected genres are having a lot of common music instruments, the timbre-based similarity returns only 47% of samples from the same genre. Harmony profile works indifferent in terms of genre information with 39% samples from the same genre, which is close to the randomized results of 33%. The Harmony similarity profile returns 46% of samples with the same key as a query sample. Here, the evaluation criterion could be extended to the related keys to treat those as similar in terms of harmony as well.

4 Soundslike

Soundslike[5] combines automatic metadata extraction with the music similarity search. The combination of both systems shows a flexible way for annotating huge databases while being able to quickly query for similar songs. The database can be extended by additional metadata from other annotation sources. The similarity search can therefore be filtered depending on the users needs enabling fast and reliable results. The database can be stored locally using SQLite[6] or decentralized on database servers with MongoDB[7] or similar systems allowing scalability depending on the number of users and music recordings.

4.1 Creating the database

The first step in using Soundslike is to build up a database. Therefore music recordings plus additional metadata have to be fed into the database as shown in Figure 1. For this process the described features are extracted from the audio file and the automatic classification on the aforementioned categories is performed. Finally, the newly extracted metadata is added

[5] https://www.idmt.fraunhofer.de/en/institute/projects_products/q_t/soundslike.html
[6] https://www.sqlite.org
[7] https://www.mongodb.com

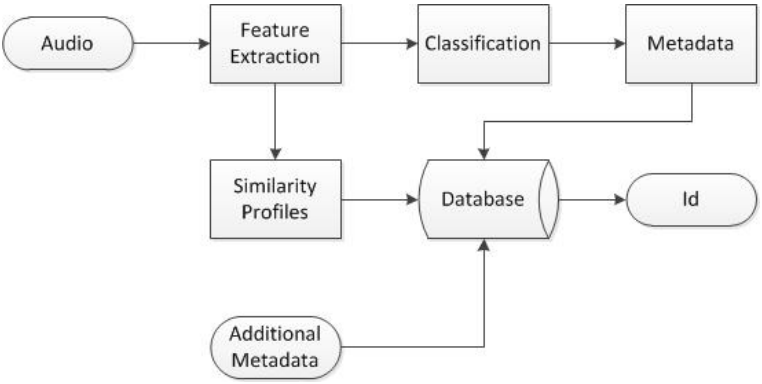

Fig. 1: Adding music recording to database.

together with existing metadata to the database. This additional metadata can be imported from different sources like ID3-Tags or iTunes. This creates a vast amount of possibilities for sorting and filtering large amounts of music data. The extracted features are also used by the selected similarity profile(s). As previously mentioned, these profiles can be adjusted to the use-case. For each added audio file, a unique id is returned which will be used for later queries.

The results of the automatic classification are also available as XML or JSON and can be used independently of the similarity results. Hereby, both automatic classification and similarity profiles can be adapted to the needs of the user (e.g., the choice of metadata categories or similarity aspects).

4.2 Recommendation

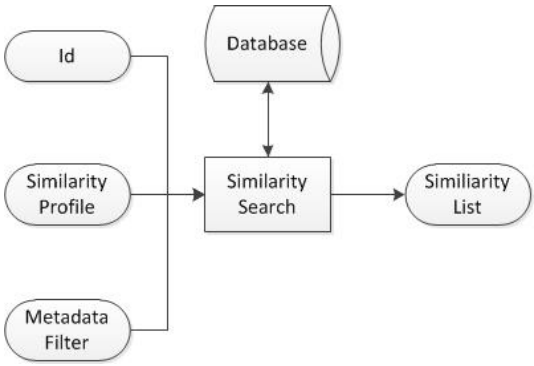

Fig. 2: Querying files for recommendation.

After a database is built, the recommendation process can be started. A previously assigned id can be used as an input for the query. Furthermore, the desired Soundslike profile can be selected and additional filters by metadata categories can be applied, see Fig. 2. The resulting similarity list can be exported in different formats like XML and JSON. Only querying audio files without inserting them beforehand to the database is also supported.

5 Conclusions

This paper presents a combined approach to automatic content-based music annotation and recommendation implemented in a Soundslike system. For the automatic music annotation, the evaluation results show room for improvement. This could be achieved by extending the training sets for each category. With enough training data, the current feature-based machine learning approach could be changed to the promising data-driven methods like CNNs. Furthermore, the extracted categories could be extended for other application-specific use-cases. For automatic music search and recommendation, we propose using predefined similarity profiles. The evaluation shows i.e. that the harmony-based similarity is indifferent to genre information but provides harmony-relevant similarity results. The music similarity can be improved by thoroughly refining the similarity profiles. The feature extraction part can be optimized with the help of feature learning and unsupervised deep learning techniques. Finally, the similarity distances can be adapted to measure the similarities of the sequences and thus enable the search of music elements like harmony progressions.

Soundslike, as described in Section 4, is used in industrial products like Jamahook, `http://www.jamahook.com`. Jamahook's Sound Match algorithm recognizes attributes of different musical elements that "match" according to key/harmony, bpm/tempo, rhythmic patterns and mood thus bring together sounds from different sources to start and/or complete a music production. This project aims to bring together creative people in the production process of music and gives users opportunities to collaborate (online platform) through matching their sounds with others.

Soundslike profits from the combined approach to music annotation and similarity. The various output formats, modular structure and support for different recommendation profiles make it flexible for various application fields. The system itself will benefit from all possible advancements in its subsystems. It can be easily updated and therefore improved by future developments.

References

[AS03] Atlas, Les; Shamma, Shihaba A.: Joint acoustic and modulation frequency. EURASIP Journal on Applied Signal Processing, 7:668–675, 2003.

[Ba16] Balke, Stefan; Driedger, Jonathan; Abeßer, Jakob; Dittmar, Christian; Müller, Meinard: Towards Evaluating Multiple Predominant Melody Annotations in Jazz Recordings. In:

Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). pp. 246–252, 2016.

[BJ15] Bonnin, Geoffray; Jannach, Dietmar: Automated generation of music playlists: Survey and experiments. ACM Computing Surveys (CSUR), 47(2):26, 2015.

[BKW15] Böck, Sebastian; Krebs, Florian; Widmer, Gerhard: Accurate Tempo Estimation Based on Recurrent Neural Networks and Resonating Comb Filters. In: ISMIR. pp. 625–631, 2015.

[Bo84] Borda, Jean C de: Mémoire sur les élections au scrutin. Histoire de l'Academie Royale des Sciences pour 1781, Paris, 1784.

[BP05] Bello, Juan Pablo; Pickens, Jeremy: A Robust Mid-Level Representation for Harmonic Content in Music Signals. In: Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR). London, UK, 2005.

[Ca08] Casey, Michael A; Veltkamp, Remco; Goto, Masataka; Leman, Marc; Rhodes, Christophe; Slaney, Malcolm: Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE, 96(4):668–696, 2008.

[Ch11] Charbuillet, Christophe; Tardieu, Damien; Peeters, Geoffroy et al.: GMM supervector for content based music similarity. In: International Conference on Digital Audio Effects, Paris, France. pp. 425–428, 2011.

[DBG07] Dittmar, Christian; Bastuck, Christoph; Gruhne, Matthias: Novel Mid-Level Audio Features for Music Similarity. In: Proceedings of the International Conference on Music Communication Science (ICOMCS). Sydney, Australia, pp. 38–41, 2007.

[Dw01] Dwork, Cynthia; Kumar, Ravi; Naor, Moni; Sivakumar, Dandapani: Rank aggregation methods for the web. In: Proceedings of the 10th international conference on World Wide Web. ACM, pp. 613–622, 2001.

[Eg15] Eghbal-zadeh, Hamid; Lehner, Bernhard; Schedl, Markus; Widmer, Gerhard: I-Vectors for Timbre-Based Music Similarity and Music Artist Classification. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). pp. 554–560, 2015.

[Fl14] Flexer, Arthur: On Inter-rater Agreement in Audio Music Similarity. In: Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). pp. 245–250, 2014.

[Fu90] Fukunaga, Keinosuke: Introduction to Statistical Pattern Recognition, Second Edition (Computer Science and Scientific Computing Series). Academic Press, 1990.

[Ga07] Gatzsche, Gabriel; Mehnert, Markus; Gatzsche, David; Brandenburg, Karlheinz: A Symmetry Based Approach for Musical Tonality Analysis. In: Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR). Vienna, Austria, pp. 207–210, 2007.

[GD09] Gruhne, Matthias; Dittmar, Christian: Comparison of harmonic mid-level representations for genre recognition. In: Proceedings of the 3rd Internation Workshop on Learning Semantics of Audio Signals (LSAS). Graz, Austria, pp. 91–102, 2009.

[GDG09] Gruhne, Matthias; Dittmar, Christian; Gärtner, Daniel: Improving rhythmic similarity computation by beat histogram transformations. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR). Kobe, Japan, pp. 177–182, 2009.

[HBL13] Humphrey, Eric J; Bello, Juan P; LeCun, Yann: Feature learning and deep architectures: New directions for music informatics. Journal of Intelligent Information Systems, 41(3):461–481, 2013.

[KL51] Kullback, S.; Leibler, R. A.: On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 1951.

[KS13] Knees, Peter; Schedl, Markus: Music similarity and retrieval. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 1125–1125, 2013.

[LDB08] Lukashevich, Hanna; Dittmar, Christian; Bastuck, Christoph: Applying statistical models and parametric distance measures for music similarity search. In: Advances in Data Analysis, Data Handling and Business Intelligence. 2008.

[Le06] Lee, Kyogu: Automatic Chord Recognition from Audio Using Enhanced Pitch Class Profile. In: Proceedings of the International Computer Music Conference (ICMC). 2006.

[Le12] Lerch, Alexander: An introduction to audio content analysis: Applications in signal processing and music informatics. John Wiley & Sons, 2012.

[Mc12] McFee, Brian: More like this: machine learning approaches to music similarity. PhD thesis, University of California, San Diego, 2012.

[Mü15] Müller, Meinard: Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. Springer, 2015.

[Pe04] Peeters, Geoffroy: , A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project: Technical Report, 2004.

[Po09] Pohle, Tim; Schnitzer, Dominik; Schedl, Markus; Knees, Peter; Widmer, Gerhard: On Rhythm and General Music Similarity. In: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR). Kobe, Japan, pp. 525–530, 2009.

[PR03] Peeters, Geoffroy; Rodet, Xavier: Hierarchical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instruments Databases. In: Proceedings of the 6th International Conference on Digital Audio Effects (DAFx). London, UK, 2003.

[UH03] Uhle, Christian; Herre, Jürgen: Estimation of tempo, micro time and time signature from percussive music. In: Proceedings of the 6th International Conference on Digital Audio Effects (DAFx). London, UK, 2003.

[YC12] Yang, Yi-Hsuan; Chen, Homer H.: Machine Recognition of Music Emotion. ACM Transactions on Intelligent Systems and Technology, 3(3):1–30, 2012.