PROPOSAL FOR TUTORIAL (ISMIR 2011)

AUDIO CONTENT-BASED MUSIC RETRIEVAL

Meinard Müller
Saarland University and MPI Informatik
Saarbrücken, Germany
meinard@mpi-inf.mpg.de

Joan Serrà
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
joan.serraj@upf.edu

1 OUTLINE OF THE TUTORIAL CONTENT

Even though there is a rapidly growing corpus of available music recordings, there is still a lack of audio content-based retrieval systems allowing to explore large music collections without manually generated annotations. In this context, the query-by-example paradigm is commonplace: given an audio recording or a fragment of it (used as *query* or *example*), the task is to automatically retrieve all documents from a given music collection containing parts or aspects that are similar to it. Here, the notion of similarity used to compare different audio recordings (or fragments) is of crucial importance, and largely depends on the application in mind as well as the user requirements.

In this tutorial, we present and discuss various content-based retrieval tasks based on the query-by-example paradigm. More specifically, we consider audio identification, audio matching, version (or cover song) identification and category-based retrieval. A first goal of this tutorial is to give an overview of the state-of-the-art techniques used for the various tasks. However, a further goal is to introduce a taxonomy that allows for a better understanding of the similarities, and the sometimes subtle differences, between such different retrieval scenarios. In particular, we elaborate on the differences between fragment-level and document-level retrieval, as well as on various match specificity levels found in the music search/match process.

1. **Audio identification.** In content-based retrieval, various levels of specificity can be considered. At the highest specificity level, the retrieval task consists in identifying a particular audio recording within a given music collection using a relatively small audio fragment as query [1, 4, 5, 18, 13, 36]. This task, which also aims at temporally locating the query fragments within the identified recording, is often referred to as *audio identification* or *audio fingerprinting*. In the last years, a number of applications for audio identification have become of commercial interest, including broadcast monitoring, automatic organization of music collections, filtering of copyrighted material or tune identification. Even though recent algorithms show a significant degree of robustness towards noise, MP3 compression artifacts and uniform temporal distortions, the notion of similarity used in the scenario of audio identification is rather close to the identity. This way, audio identification can be seen as an *exact-duplicate detection* task.

   In this part of the tutorial, we explain the most important requirements of audio identification systems, including robustness, reliability, granularity, scalability, and efficiency. We then explain the main ideas behind classical audio fingerprinting techniques. In particular, we look at the audio descriptors of two widely used fingerprinting techniques. Firstly, we explain the fingerprints proposed by Wang [36], which are based on “constellations” of spectral peaks. Secondly, we discuss the fingerprints introduced by Haitsma and Kalker [13], which are referred to as fingerprint blocks, and consist of short sequences of frame-based bit vectors. Finally, we also overview the indexing and hashing techniques used in these two fingerprinting systems.

2. **Audio matching.** While the problem of audio identification can be regarded as largely solved even for large scale music collections, semantically more advanced retrieval tasks are still mostly unsolved. Indeed, existing algorithms for audio identification cannot deal with strong non-linear temporal distortions or with other musically motivated variations that concern, for example, the articulation or the instrumentation. The task of *audio matching* can be seen as an extension of audio identification. Here, given a short query (i.e. an audio fragment), the goal is to automatically retrieve all fragments that musically correspond to the query from all documents within a given collection (e.g. audio recordings, video clips), see [17, 22]. For related problems, which also concern matching across different music representations, see also [10, 14, 24, 32]. In the audio matching scenario, and opposed to the audio identification scenario, one particularly admits semantically motivated variations as they typically occur in different performances and arrangements

of a musical piece. For example, two performances may exhibit significant global and (non-linear) local differences in tempo, articulation and phrasing. Such local tempo differences are, for example, the result of variations in executing ritardandi, accelerandi, fermatas, or ornamentations. Additionally, one has to deal with considerable deviations in dynamics and timbre, which are the results of differences in instrumentation, loudness, tone color, accentuation and so on.

In the second part of the tutorial, we highlight the difference between the audio identification and audio matching task by presenting a number of suitable audio examples from the classical and popular music domain. To cope with the abovementioned variations in timbre and instrumentation, most matching procedures rely on chroma-based audio features [2, 11, 20]. We explain various variants of these features, while emphasizing the importance of the feature design step. Furthermore, to account for global and local tempo distortions, one typically resorts to alignment procedures. In particular, we discuss a subsequence variant of dynamic time warping, which allows for deriving a matching curve [20]. Such a curve not only indicates the desired matches, but also turns out to be a powerful tool for expressing the matching capability of various feature representations [21]. Finally, we indicate how the matching procedure may be extended using indexing methods to scale to medium size datasets [17].

3. **Version identification.** Audio identification and audio matching are instances of *fragment-level* retrieval scenarios, where the goal is to retrieve all musically related fragments contained in the documents of a given music collection. In contrast, in *document-level* retrieval, a single similarity measure is considered to globally compare entire documents. One recently studied instance of document-level retrieval is referred to as cover song or, more properly, *version identification*, where the goal is to identify the different versions of the same musical piece that are present in a collection [6, 8, 28, 29]. Therefore, in the version identification scenario, we are dealing with a *near-duplicate detection* task. As in the audio matching scenario, a version may differ from the original recording in many ways, possibly including changes in timbre, instrumentation, tempo, main tonality, harmony, melody and lyrics. However, in the version identification scenario, the differences can be extreme: a version may represent a different genre, it may be performed live and adapted to a particular singer, or it may be a remix or cover song with a different musical structure.

   In the third part of our tutorial, we give a detailed overview of the version identification problem. In particular, we critically discuss different aspects of the methods for identifying versions of the same piece [3, 8, 19, 28, 29, 33]. These methods usually combine the extraction of the temporal evolution of the tonal information, together with blocks that deal with changes in the tempo, the key or the structure of the recording [28]. After this basic overview, we drive our attention towards an example of an accurate version identification algorithm [30] and towards different pre- and post-processing techniques that are suitable for the task of version identification [9, 26, 31].

4. **Category-based music retrieval.** Finally, considering even less specific matches between entire music documents, there are a number of document-level retrieval tasks which we group under the term *category-based retrieval*. This term encompasses retrieval of documents whose relationship can be described by cultural or musicological categories [7]. Some categories which have been the subject of substantial research efforts are genre [27, 35], rhythm styles [12] or mood or emotions [15, 16, 34]. Music recommendation or general music similarity assessments [23, 25] can be seen as further document-level retrieval tasks with even less specificity.

   In the final part of this tutorial, we give a brief overview of these general document-level retrieval scenarios, where the notion of similarity is weaker and sometimes fuzzy. We discuss the general pipelines used in category-based retrieval, which consist of extracting several audio content-based descriptions and in applying a classification scheme. Such descriptions are usually global representations of the musical piece reflecting one or several music characteristics that are, on average, present through the whole recording. The classification scheme usually starts by automatically selecting the descriptions that are best suited to the task at hand and then training a classifier to identify such categories.

It is the goal of our tutorial to give an overview of the various audio retrieval tasks, to explain the commonalities of and differences between these tasks and to explain the main techniques used in state-of-the-art approaches. We want to emphasize that in the abovementioned problems one has to deal with a trade-off between efficiency and specificity. The more specific the search task is, the more efficient it can be solved using indexing techniques. In the presence of significant spectral and temporal variations, the feature extraction as well as the matching steps become more delicate and cost-intensive (e.g. local warping and alignment procedures). Here, the scalability to very large data collections consisting of millions of documents still poses many yet unsolved problems.

2 INTENDED AND EXPECTED AUDIENCE

In this tutorial, we cover basic principles as well as state-of-the-art techniques for audio content-based music retrieval in a non-technical way. Our main goal is to give a comprehensive overview of the different music retrieval tasks while introducing some taxonomy that allows for a better understanding of the commonalities and differences between these tasks. By providing many illustrative audio examples and by working with pictures (rather than with formulas), we will make an effort to convey the main ideas, in particular to non-experts and to researchers who are new to the field. By doing so, the tutorial appeals to a wide and interdisciplinary audience working in different fields ranging from musicology and music perception to information retrieval and signal processing. Furthermore, we will provide handouts of our slides that will also contain pointers to the most relevant literature for the various retrieval tasks.

3 SHORT BIOGRAPHY OF THE PRESENTERS

Meinard Müller studied mathematics (Diplom) and computer science (Ph.D.) at Bonn University, Germany. In 2002/2003, he conducted postdoctoral research in combinatorics at the Mathematical Department of Keio University, Japan. In 2007, he finished his Habilitation at Bonn University in the field of multimedia retrieval writing a book titled *Information Retrieval for Music and Motion*, which appeared as Springer monograph. Currently, Meinard Müller is a member of the Saarland University and the Max-Planck Institut für Informatik, where he is leading the research group *Multimedia Information Retrieval & Music Processing* within the Cluster of Excellence on *Multimodal Computing and Interaction*. His recent research interests include content-based multimedia retrieval, audio signal processing, music processing, music information retrieval, and motion processing.

Joan Serrà obtained both the degrees of Telecommunications and Electronics at Enginyeria La Salle, Universitat Ramón Llull, Barcelona, Spain, in 2002 and 2004, respectively. After working from 2005 to 2006 at the research and development department of Music Intelligence Solutions Inc, he joined the Music Technology Group of Universitat Pompeu Fabra, Barcelona, where he received the MSc and PhD in Information, Communication and Audiovisual Media Technologies in 2007 and 2011, respectively. He is currently a post-doc researcher with the aforementioned group and a part-time associate professor with the Dept. of Information and Communication Technologies of the same university. In 2010 he was a guest scientist with the Research Group on Nonlinear Dynamics and Time Series Analysis of the Max Planck Institute for the Physics of Complex Systems in Dresden, Germany. His research interests include music retrieval and understanding, signal processing, time series analysis, complex networks, complex systems, information retrieval and music perception, psychology and cognition.

4 SPECIAL REQUIREMENTS

No special requirements are needed besides the usual equipment (beamer, internet connection and loudspeakers connectable to a standard laptop). However, we need to have *stereo*, where the volume of the two channels is adjustable separately.

5 CONTACT INFORMATION

Name:	**Priv.-Doz. Dr. Meinard Müller**
Position:	Senior Researcher
Nationality:	Germany
Institution:	Saarland University and MPI Informatik
Address:	Campus E1-4, 66123 Saarbrücken, Germany
Phone:	+49-681-9325405
Fax:	+49-681-9325499
E-mail:	`meinard@mpi-inf.mpg.de`
Web:	`http://www.mpi-inf.mpg.de/~mmueller/`

Name:	**Dr. Joan Serrà**
Position:	Part-time Associate Professor & Researcher
Nationality:	Spanish
Institution:	Universitat Pompeu Fabra (MTG)
Address:	Roc Boronat 138, 08018 Barcelona, Spain
Phone:	+34-935-422864
Fax:	+34-935-422455
E-mail:	`joan.serraj@upf.edu`
Web:	`http://joanserra.weebly.com/`

6 REFERENCES

[1] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, and M. Cremer. AudioID: Towards content-based identification of audio material. In *Proc. 110th AES Convention*, Amsterdam, NL, 2001.

[2] M. A. Bartsch and G. H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, Feb. 2005.

[3] J. P. Bello. Audio-based cover song retrieval using approximate chord sequences: testing shifts, gaps, swaps and beats. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pages 239–244, 2007.

[4] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of algorithms for audio fingerprinting. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 169–173, St. Thomas, Virgin Islands, USA, 2002.

[5] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of audio fingerprinting. *Journal of VLSI Signal Processing Systems*, 41(3):271–284, 2005.

[6] M. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech & Language Processing*, 16(5), 2008.

[7] J. S. 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.

[8] D. P. W. Ellis and G. E. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1429–1432, 2007.

[9] R. Foucard, J. L. Durrieu, M. Lagrange, and G. Richard. Multimodal similarity between musical streams for cover version detection. In *Proc. of the IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages 5514–5517, 2010.

[10] C. Fremerey, M. Müller, F. Kurth, and M. Clausen. Automatic mapping of scanned sheet music to audio recordings. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, pages 413–418, Philadelphia, USA, Sept. 2008.

[11] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, UPF Barcelona, 2006.

[12] F. Gouyon. *A computational approach to rhythm description: audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005. Available online: http://mtg.upf.edu/node/440.

[13] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 107–115, Paris, France, 2002.

[14] N. Hu, R. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, US, October 2003.

[15] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. Ehmann. The 2007 mirex audio mood classification task: lessons learned. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2008.

[16] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: a state-of-the-art review. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 255–266, 2010.

[17] F. Kurth and M. Müller. Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):382–395, Feb. 2008.

[18] F. Kurth, A. Ribbrock, and M. Clausen. Identification of highly distorted audio material for querying large scale data bases. In *Proceedings of the 112th AES Convention*, 2002.

[19] M. Marolt. A mid-level representation for melody-based retrieval in audio collections. *IEEE Trans. on Multimedia*, 10(8):1617–1625, 2008.

[20] M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

[21] M. Müller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 18(3):649–662, 2010.

[22] M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pages 288–295, 2005.

[23] E. Pampalk. *Computational models of music similarity and their application to music information retrieval*. PhD thesis, Vienna University of Technology, Vienna, Austria, 2006. Available online: http://www.ub.tuwien.ac.at/diss/AC05031828.pdf.

[24] J. Pickens, J. P. Bello, G. Monti, T. Crawford, M. Dovey, M. Sandler, and D. Byrd. Polyphonic score retrieval using polyphonic audio. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[25] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer. On rhythm and general music similarity. In *Proc. of the Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, pages 525–530, 2009.

[26] S. Ravuri and D. P. W. Ellis. Cover song detection: from high scores to general classification. In *Proc. of the IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages 55–58, 2010.

[27] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[28] J. Serrà, E. Gómez, and P. Herrera. Audio cover song identification and similarity: background, approaches, evaluation and beyond. In Z. W. Ras and A. A. Wieczorkowska, editors, *Adv. in Music Information Retrieval*, volume 16 of *Studies in Computational Intelligence*, chapter 14, pages 307–332. Springer, Berlin, Germany, 2010.

[29] J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Trans. on Audio, Speech and Language Processing*, 16(6):1138–1152, 2008.

[30] J. Serrà, X. Serra, and R. G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11:093017, 2009.

[31] J. Serrà, M. Zanin, P. Herrera, and X. Serra. Characterization and exploitation of community structure in cover song networks. *Pattern Recognition Letters*, 2010. Submitted.

[32] I. S. H. Suyoto, A. L. Uitdenbogerd, and F. Scholer. Searching musical audio using symbolic queries. *IEEE Transactions on Audio, Speech & Language Processing*, 16(2):372–381, 2008.

[33] W. H. Tsai, H. M. Yu, and H. M. Wang. Using the similarity of main melodies to identify cover versions of popular songs for music document retrieval. *Journal of Information Science and Engineering*, 24(6):1669–1687, 2008.

[34] E. Tsunoo, T. Akase, N. Ono, and S. Sagayama. Musical mood classification by rhythm and bass-line unit pattern analysis. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010.

[35] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 5(10):293–302, 2002.

[36] A. Wang. An industrial strength audio search algorithm. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 7–13, Baltimore, USA, 2003.