

# Chapter 1

## Introduction

### 1.1 Music Information Retrieval

Music Information Retrieval (MIR) is an emerging, interdisciplinary science [21] that aims at retrieving information from music. It draws on fields like musicology, cognitive psychology, linguistics, library science, and last, but not least, computer science.

Michael Kassler mentioned the term “Musical Information Retrieval” (MIR) as early as 1966 [31]. He describes an assembler-like programming language called MIR which can be used to navigate music scores and find positions that fulfill certain criteria. He realized that his language had somewhat limited capabilities, for example that it would be very difficult to write an MIR program that could recognize whether a musical piece is a parody of another one. Interestingly enough, he claimed that the problem of Optical Music Recognition (OMR) was solvable at a cost of about one million dollars, and that an OMR system would be able to transcribe printed music scores at a rate of several thousand musical symbols per minute. Unfortunately, he did not offer any guess on how many of those symbols would be recognized correctly. Forty years after Kassler’s talk, Don Byrd presented a paper [8] at ISMIR 2006 (see page 8 for an explanation of “ISMIR”) that does not mention the transcription speed, which really is not a very serious concern, but shows that even the best contemporary OMR programs still produce double-digit numbers of errors per page including, for example, incorrectly recognized pitches for up to 20 % of the notes, and several percent of incorrectly recognized note durations.

Still, in the late 20th century, along with other areas of Multimedia Information Retrieval such as Image Retrieval or Video Retrieval, Music Information Retrieval started to flourish. Around the end of the 20th century, storage space (hard drives, flash memory) became cheap and abundant, and file formats such as MP3 were developed that made it possible to store vast amounts of music in good quality. These advances, along with the fact that processing power has become cheap and abundant as well, have created both the need for automatically retrieving information from music and new possibilities to address this need. Unlike many other art forms such as painting, music stays very enjoyable even if it is stored or transmitted digitally. This leads to large collections of digital music that can be difficult to manage.

Traditional methods of organizing digital music collections by using metadata quickly reach their limits because metadata are frequently unreliable, missing, or extremely expensive to create (a good example is Pandora, see 2.2.11), and also because it is not always easy to invent good categories one could use for metadata. For example, people with different tastes will need very different sets of music genres for characterizing their collection. For some people, it is important to distinguish between categories such as Trance, House, Dance-Pop, Acid, maybe some more categories and Classical Music, while for others a categorization such as Popular Music, Baroque, Classical Music, Romanticism, and music from the 20th century (just to name a few) would be more useful. Unfortunately, users from these two groups even use similar labels (“Classical”) for very different concepts. By not relying on manually attached labels, but rather on the music itself for retrieval and clustering tasks, one can avoid not only a lot of work, but also tricky problems like this.

One can gain a good overview of typical MIR tasks by looking at the MIREX competition<sup>1</sup> for MIR algorithms. The tasks from its first two rounds in 2005 and 2006 can be grouped into classification, feature extraction, alignment, and retrieval.

- Classification:
  - **Audio Artist Identification.** Map audio recordings to labels identifying the artist who created the recording.
  - **Audio/Symbolic Genre Classification.** Determine the musical genre for audio or MIDI recordings.
- Feature Extraction:
  - **Audio Melody Extraction.** Extract the main melodic line from polyphonic audio. This involves two subtasks: Voicing detection (deciding whether a particular time frame contains a “melody pitch”) and pitch detection (deciding the most likely melody pitch for each time frame).
  - **Audio Onset Detection.** List the onset times of notes in audio recordings.
  - **Audio Drum Detection.** Determine the onset times and corresponding drum class names of drum events in polyphonic music; that is, unlike the previous task, only detect drum onsets instead of those of any note, and also analyze which class the drum belongs to.
  - **Audio Tempo Extraction.** Determine the perceived tempo for audio recordings.
  - **Audio and Symbolic Key Finding.** Determine the key for a given audio recording or MIDI file.
  - **Audio Beat Tracking.** Unlike the drum detection or tempo extraction task, the problem here is to determine the beat locations in an audio recording.
- Alignment: **Score Following.** Real-time alignment of a music signal (audio or MIDI) to a music score.
- Retrieval:

<sup>1</sup>See [http://www.music-ir.org/mirexwiki/index.php/Main\\_Page](http://www.music-ir.org/mirexwiki/index.php/Main_Page)

- **Audio Music Similarity and Retrieval.** Calculate a similarity matrix for a list of given audio files.
- **Query by Humming<sup>2</sup>, Symbolic Melodic Similarity.** Given a query, search a collection for pieces that contain melodically similar musical material.

This task list shows mainly MIR tasks that are neither satisfactorily solved nor extremely far from a solution. For example, the identification of recordings based on excerpts is not a MIREX task because there are already efficient and effective methods known, such as Shazam’s method (see Section 2.2.15) or MusicDNS<sup>3</sup>. On the other hand, some tasks are still far from a satisfying solution, for example the automatic conversion of audio recordings to MIDI files or even an intermediate step towards that goal, the separation of several audio sources (various instruments or voices in a polyphonic piece of music) that are found within one recording.

## 1.2 Topics of this thesis

The main topic of this thesis is a method for the “Symbolic Melodic Similarity” task, that is, measuring melodic similarity for notated music such as MIDI files. Chapter 2 describes several methods for solving this and similar tasks, along with examples of Music Information Retrieval systems that implement these methods.

In Chapter 3, a music search algorithm is developed and studied which views music as sets of notes that are represented as weighted points in the two-dimensional space of time and pitch. Two point sets can be compared by calculating how much effort it would take to convert one into the other; effort is measured by determining how much weight has to be moved over what distances. The point sets are more similar if there are fewer and smaller movements of weight needed. This transportation-based similarity measure has some desirable properties such as continuity and the ability to match any combination of polyphonic and monophonic music.

To make these point set comparisons efficient enough for searching large databases, the distances between every item (point set) in the database and a small, fixed set of special (vantage) point sets can be pre-calculated. Whenever a new query needs to be compared to the items in the database, one can restrict the search to those items with similar distances to the special point sets. The application of this vantage indexing method to music is described in Chapter 4. Vantage indexing was first suggested for image retrieval [90].

For studying the performance of the transportation-based search algorithm and other, similar ones, the creation of a ground truth for a large music collection (RISM) is described in Chapter 5, along with a performance measure and the application of both the ground truth and the measure for the MIREX algorithm competition.

In Chapter 6, a distance measure is described that is inspired by transportation distances, but puts additional constraints on what flows are possible. If the Earth Mover’s Distance (defined in Section 3.1.2) is used for comparing a set containing some points with very large weights to another set with very light points at similar

---

<sup>2</sup>For a definition of “Query by Humming”, see Section 1.5.

<sup>3</sup>MusicDNS (<http://www.musicdns.org/>) is an open source audio fingerprinting system. Unlike Shazam, it needs more than just a few seconds of audio to identify a track. See Section 2.2.7.

positions, it can happen that the heavy points are partially matched with points that lie far away in the time dimension. This usually does not make musical sense. It can be avoided by constructing a graph with nodes that represent notes and edges that connect notes. For such a graph, the solution of a maximum-flow, minimum-cost problem can be used for comparing point sets in a fashion similar to how the Earth Mover's Distance works. To make unwanted flows less likely, the weight distribution can be normalized before solving the maximum-flow, minimum cost problem. Supporting polyphony can be achieved by constructing the network accordingly.

Chapter 7 concludes the thesis with a brief overview of its contribution, open issues, and some thoughts about the future of Music Information Retrieval.

This thesis contains material that was published in peer-reviewed journals and at international conferences.

- **Chapter 2:** Rainer Typke, Frans Wiering, Remco C. Veltkamp: A Survey of Music Information Retrieval Systems. Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR), London, September 2005 [81]
- **Chapter 3:**
  - Rainer Typke, Panos Giannopoulos, Remco C. Veltkamp, Frans Wiering, Ren van Oostrum: Using transportation distances for measuring melodic similarity. Proceedings of the International Conference on Music Information Retrieval (ISMIR), pages 107–114, Baltimore, October 2003 [75, 76]
  - Frans Wiering, Rainer Typke, Remco C. Veltkamp: Transportation Distances and their Application in Music-Notation Retrieval. In: Music Query: Methods, Strategies, and User Studies (Computing in Musicology 13, 2004), pages 113-128. CCARH and MIT Press. [93]
  - Remco C. Veltkamp, Frans Wiering, Rainer Typke: Content Based Music Retrieval. In: Encyclopedia of Multimedia, Borko Furht (Ed.), ISBN: 0-387-24395-X, Springer 2006. [89]
  - Rainer Typke, Remco C. Veltkamp, Frans Wiering: Searching notated polyphonic music using transportation distances. Proceedings of the ACM Multimedia Conference, pages 128–135, New York, October 2004 [77]
  - Rainer Typke, Frans Wiering, Remco C. Veltkamp: A search method for notated polyphonic music with pitch and tempo fluctuations. Proceedings of the Fifth International Conference on Music Information Retrieval (ISMIR), pp. 281-288, Barcelona, October 2004 [79]
  - Rainer Typke, Frans Wiering, Remco C. Veltkamp: Transportation distances and human perception of melodic similarity. ESCOM Musicae Scientiæ, 2007 [83]
- **Chapter 4:** Reinier H. van Leuken, Remco C. Veltkamp, Rainer Typke: Selecting vantage objects for similarity indexing. International Conference on Pattern Recognition (ICPR) 2006, Hong Kong [87]
- **Chapter 5:**

- Rainer Typke, Marc den Hoed, Justin de Nooijer, Frans Wiering, Remco C. Veltkamp: A Ground Truth For Half A Million Musical Incipits. Proceedings of the 5th Dutch-Belgian Information Retrieval Workshop (DIR) 2005, Utrecht, the Netherlands, pages 63–70. [73, 72]
- This publication was selected to also appear in the Journal of Digital Information Management:  
Rainer Typke, Marc den Hoed, Justin de Nooijer, Frans Wiering, Remco C. Veltkamp: A Ground Truth For Half A Million Musical Incipits. Journal of Digital Information Management 3(1), 2005, pages 34–39 [74]
- Rainer Typke, Remco C. Veltkamp, Frans Wiering: A measure for evaluating retrieval techniques based on partially ordered ground truth lists. International Conference on Multimedia & Expo (ICME) 2006, Toronto, Canada [78]
- Rainer Typke, Frans Wiering, Remco C. Veltkamp: Evaluating the Earth Movers Distance for measuring symbolic melodic similarity, MIREX abstract, 2005 [80]
- Rainer Typke, Frans Wiering, Remco C. Veltkamp: MIREX Symbolic Melodic Similarity and Query by Singing/Humming, MIREX abstract, 2006 [82]

### 1.3 The usefulness of melodic similarity measures

As the list of MIREX tasks shows, one can distinguish two main groups of methods for content-based searching of music databases: methods for audio data and methods for notated music. Some Music Information Retrieval systems combine the two by first converting an audio signal into a symbolic description of notes and then searching a database of notated music. Since many researchers work on tasks for automatically creating symbolic descriptions of audio recordings (such as beat detection, onset detection, melody extraction, and even complete transcription to MIDI of the notes in an audio recording), methods that perform well for symbolic data are eventually going to be useful for audio data as well.

Melody search engines (search engines which serve the information need for pieces of music that contain musical material which is melodically similar to a given query) can be useful for a variety of purposes and audiences:

- Query-by-Humming: in record stores, it is not uncommon for customers to only know a tune from a record they would like to buy, but not the title of the work, composer, or performers. Salespeople with a vast knowledge of music who are willing and able to identify tunes hummed by customers are scarce, and it could be interesting to have a computer do the task of identifying melodies and suggesting records. A Query-by-Humming device would also be interesting for libraries, as a phone service similar to Shazam (see Section 2.2.15), or to aid internet users with the task of retrieving entertaining MP3 files for buying them online.
- A search engine that finds musical scores similar to a given query can help musicologists find out how composers influenced one another or how their works are related to earlier works of their own or by other composers. This task has been done manually by musicologists over the past centuries. If

computers could perform this task reasonably well, more interesting insights could be gained faster and with less effort.

- Copyright issues could be resolved, avoided or raised more easily if composers could easily find out if someone is plagiarizing them or if a new work exposes them to the risk of being accused of plagiarism.

For example, the symbolic melodic similarity retrieval algorithm described in Chapter 3 is used as one of the starting points for Frans Wiering's "Witchcraft" project<sup>4</sup>, which aims at creating a search tool that helps test hypotheses about oral transmission of folk songs. This tool is also going to be integrated in the Nederlandse Liederbank to provide access to the folksong collection of "Onder de Groene Linde", taking into account the problem of oral variation.

## 1.4 Important features for melody, invariances in perception

Melody is one of the most memorable and characteristic features of Western music. The main topic of this thesis is a retrieval method for melodically similar items, so a few words about what makes melodies similar and what does not are in place.

From the cognitive theory of music [71], we know that melodic motion (characterized by successive pitch intervals) and contour are very important for the perception of a melody. For the rhythmical aspect, patterns are perceived in relation to an underlying pulse that defines the tempo.

Melodic motion and contour as well as the rhythmic patterns do not change if the tempo is changed. Rhythm is always defined in relation to the pulse, and if the pulse gets slower or faster, the rhythm stays the same (unless the tempo change is extreme). For pitch intervals and melodic contour, it is obvious that they are not affected by tempo changes. Also, if a melody is transposed to a different key, these things do not change. Therefore, a basic requirement for a melody search engine is that its distance measures are invariant under transposition and augmentation or diminution.<sup>5</sup> Another musical feature that does not influence the perception of melodies is timbre; making a melody search engine ignore timbre is trivial if it uses notation.

Studies such as Selfridge-Field's article [70] show that melodic similarity is continuous. Local melodic changes such as lengthening a note or moving it up or down a step are usually not perceived as changing the identity of a melody, and by applying more and more changes, the perceived relationship to the original becomes only gradually weaker. Also, melodies are generally quite resistant to the insertion of all sorts of ornamentation. Mozart's variations on "Ah, vous dirai-je, maman" can serve as an illustration for this. See Figure 1.1.

<sup>4</sup><http://www.cs.uu.nl/research/projects/witchcraft/>

<sup>5</sup>In Section 3.5.2, we will see that the performance of a search engine can still be improved by attaching a cost to tempo changes. The main reason for this is that there are limits to the tempo invariance of human perception. Snyder [71] explains, for example, that there are some fixed thresholds for time intervals within which events are perceived as happening concurrently; whenever one changes the tempo such that some musical events cross one of these thresholds, perception will change. So, a good distance measure for melodies should not necessarily be always invariant under tempo changes, but allow for controlling the influence of tempo changes on the resulting distances.



Figure 1.1: A few excerpts from Wolfgang Amadeus Mozart: Twelve variations on “Ah, vous dirai-je, maman”, K. 300e.

## 1.5 Some terms and basic facts

For the rest of this book, we will use the following terms without defining them again:

**Pitch** The pitch of a sound determines as how “high” it is perceived. Different areas of the cochlea are responsible for detecting different pitches. A musical instrument or singer produces a spectrum of signals with different frequencies. For a pitched instrument, there is a fundamental frequency (often abbreviated as “F0”), accompanied by overtones whose frequencies are multiples of the fundamental frequency. F0 estimation is an important task for pitch detection.

**Chroma** In a Music Information Retrieval context, chroma is a 12-dimensional vector containing the spectral energy of each of the 12 traditional pitch classes of the equal-tempered scale. This feature takes the close octave relationship in melody and harmony into account as it is prominent in Western music.

**Intensity/loudness** The amplitude and therefore the energy of the musical signal determines how intense or loud music is perceived.

**Timbre** Timbre is best defined as what it is not: it is the qualities of sounds that make them distinguishable and that do not fall under either pitch or loudness.

**Intervals** When speaking about “intervals”, we usually mean intervals between pitches. The sequence of intervals between subsequent notes is important for the perception of melody.

**Onset time, duration** A note usually starts with a very short pitchless attack, followed by the presence of a clearly detectable pitch for a certain amount of time. By “onset time”, we mean the point of time when a note starts, and a note’s

duration is determined by the time period between the onset time and the point of time when the note stops being audible (the offset time).

**Melodic contour** By “melodic contour”, we mean a sequence of interval directions in a melody. If this sequence only distinguishes between “up”, “down”, or “repeat”, we call it “gross contour” to express the fact that the information from the interval sequence is very much reduced.

**N-grams** We call a group of  $n$  subsequent notes an  $n$ -gram.

**Query by Humming** With “Query by Humming”, we mean that the user enters a search query by singing, humming, or whistling it into a microphone. That is, the query contains a pitch vector, but not necessarily a measure structure or lyrics. Neither pitches nor tempo are quantized, and onsets or notes need to be detected after the query has been entered if this information is needed by the search algorithm.

### 1.5.1 Acronyms

The following acronyms will appear frequently when talking about music information retrieval:

**ISMIR** ISMIR is the main music information retrieval conference. It started as “International Symposium on Music Information Retrieval”; now it is a veritable conference and not just a symposium. For more information, see <http://www.ismir.net>.

**MIREX** MIREX is a TREC-like series of comparisons for algorithms. MIREX stands for “Music Information Retrieval Evaluation eXchange”. See <http://www.music-ir.org/mirex2006/>.

**RISM** Répertoire International des Sources Musicales. The International Inventory of Musical Scores describes itself as “a cross-country non-profit joint venture which aims at comprehensive documentation of the worldwide existing musical scores”. See <http://rism.stub.uni-frankfurt.de/>. RISM has created a collection of musical incipits and metadata, the “RISM A/II” collection, that covers material from several centuries. We used this collection for comparing the retrieval algorithm that is described in this book to other approaches. Subsets of this collection were used for the “Symbolic Melodic Similarity” MIREX task in 2005 and 2006.

**TREC** The Text Retrieval Conference (TREC) started yearly competitions for retrieval algorithms in 1992. See <http://trec.nist.gov/>.