Modulo7 : A full stack Music Information Retrieval and Querying Engine using Music Theoretic Principles

by

Arunav Sanyal

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

December, 2015

Abstract

Music Information Retrieval (MIR) is an interdisciplinary science of extracting non trivial information and statistics from music data sources. In today's digital age, music is stored in a variety of digitized formats - e.g midi, musicxml, mp3, digitized sheet music etc. Music Information Retrieval Software aim at extracting features from one or more of these source. MIR research helps in solving problems like automatic music classification, recommendation engine design etc. Users can then query the acquired statistics to acquire relevant information.

The author proposes and implements a new Music Information Retrieval and Query Engine called Modulo7. Unlike other MIR software which deal with low level audio features, Modulo7 operates on the principles of music theory and a symbolic representation of music. Modulo7 is a full stack deployment, with server components that parse various sources of music data into its own efficient internal representation and a client component that allows consumers to query the system with sql like queries which satisfies certain music theory criteria (and as a consequence Modulo7 has a

ABSTRACT

custom relational algebra with its basic building blocks based on music theory).

Primary Reader: Dr David Yarowsky

Secondary Reader: Dr Yanif Ahmad

Acknowledgments

I would like to thank Dr David Yarowsky for giving me the opportunity to work on this project. His detailed insights have immensely helpful to me to power through my work. I would like to thank Dr Yanif Ahmad for his crucial help in the systems aspects of my query engine and the implementation of the server side components.

Dedication

This thesis is dedicated to my family and to all the music lovers in the world.

Contents

CONTENTS

Chapter 1

Introduction

Why does a person like a particular song? What are the inherent aspects of a song that pleases a person's musical taste? Is it the complexity of a song, the beat the song or just a particular melodic pattern ? More so if a person likes a song, can we predict if he/she will like a similar song?

Music has been created since the dawn of civilization and these questions have plagued mankind just as long. In response to this, man has created elaborate systems of formal study for music and classification techniques in almost every ethnic community since antiquity. Two notable examples are the western system of solfege and classical music theory and the Indian system of raagas. These elaborate systems are based on very simple fundamental building blocks of melody and harmony and simple rules that govern the interplay of these building blocks. However very complex pieces of

music can be created with these simple rules depending on the skill and virtuosity of artists. Composers use these rules and concepts to create novel music for mass consumption.

In the modern era industry and academia have attempted to address the problem of music recommendation and music classification. The industry has predominantly favored approaches that look at user preferences and history. For example Amazon Music recommendation works on users shopping history. Pandora on the other hand hires an army of musicologists to ascertain how a song is similar to another song and creates software that leverages this adhoc generated data. These approaches are either expensive in the human labor needed or in the amount of data processed that is input from a large number of users. More recently, companies like Echo Nest has extensively extracted features from music sources and mined cultural information on the web but leave it at consumers how best to leverage the data. Hence symbolic MIR is not traditionally used in industry and music theory is an after thought.

Academia on the other hand attempts to solve very particular problems in MIR. Typical examples would be cover song detection, processing information via signal processing, audio feature extraction, optical music recognition etc. In most cases the applications are of a very specific domain and does not fully scale with bulk music data. Generic frameworks like the jMIR[1] (which also happens to be a major inspi-

ration for Modulo7) suite for automatic music classification exists, which is meant to facilitate research in MIR with a machine learning focus. However academia is disconnected with industry and no full scale MIR engines can satisfy the scale of industry applications.

This work is attempt to bridge both communities. Modulo7 is a full stack deployment of Music Information Retrieval Software, providing both a server architecture and a sql like client to query based on music theory criteria. Modulo7 is a big data and information retrieval framework to explore the possibilities of exploring music theoretical aspects of music sources. Modulo7 does not attempt to solve very complex music theoretic problems (e.g study orchestral music to identify counter point information). Rather Modulo7 acts a framework on which such analysis can be built upon. Most importantly, Modulo7 addresses the issue of scale and allows a fast comparison between songs on certain music theoretic criteria. Modulo7 also acts to address deficiencies in existing software, such as filling up incompleteness in music sources. Certain problem statement of this sort would be Key estimation, Tempo estimation etc.

Chapter 2

Literature Review

Music Information Retrieval is an active and vibrant community. Both academia and industry diligently pursue it albeit with different goals in mind. While academia's primary aim is to explore particular problems (e.g cover song detection, estimating chords from chroma vectors[2]) etc. Whereas Industry is primarily interested in solving problems like song recommendation and similarity searches. The following sections outlines the software efforts and research problems tackled by MIR community in general.

2.1 Current MIR Software

Both Industry and Academia have created an extensive set of software for solving these problems. The author presents an overview of such software and the problems they attempt to address. The following software packages were investigated

2.1.1 jMIR

jMIR,[1] or Java Music Information Retrieval tool set is a collection of java code, GUI, API and CLI tools for the purpose of feature extraction from variety of music sources (in particular audio, midi) and mine cultural information from the web. jMIR extracts and exhaustive set of features that can be used in machine learning tasks. The primary use of jMIR is automatic music classification and feature extraction and not similarity computations per se (which is one of Modulo7's core goals). Moreover jMIR does not scale to myriad sources of music in existence. Unlike Modulo7 jMIR also relies on faithful recordings and does not attempt to fill up missing information (like key signature not being encoded etc). Nevertheless its one of the best Open Source MIR software in existence especially for MIR research.

2.1.2 Marsyas

marsyas[3] (Music Analysis, Retrieval and Synthesis for Audio Signals) is a software stack for audio processing with specific emphasis on Music Information Retrieval and

music signal extraction. Marsyas is a heavily developed and a widely developed state of the art framework for audio processing but also has a steep learning curve. Modulo7 has different goals (multiple format support, music similarity etc).

2.1.3 SIMILIE

SIMILIE[4] is a set of tools for music similarity measures used for single melodies and features multiple ways to construct vector space models for melodies. The techniques used in SIMILIE are novel. Modulo7 uses a subset of these similarity measures as basis for an extended and improved model of similarities based on polyphonic music and harmonic elements. Moreover SIMILIE needs its own file format (called .mcsv) for analysis. Although the software package gives a converter for different sources, its not as variegated as Modulo7's format support is.

2.1.4 Echo Nest APIs

Echo Nest is a company that specializes in big data music intelligence. Echo Nest power many music platforms like last.fm, Spotify etc. In particular Echo Nest provides APIs for extraction of audio features, acquiring artists similar to a particular artist etc. Echo Nest API is used for some sub tasks in Modulo7 (which is discussed in the Software Architecture Chapter).

Echo Nest also maintains the worlds biggest music database as well as data mined from them along with extracted audio features, web mined information, user preference etc).

2.1.5 Humdrum

Humdrum[5] is a set of tools for assistant in music research. Humdrum has the capability for solving very complex questions using music theoretic concepts. Humdrum supports its own file format for analysis. Humdrum is specifically designed for musicologists for automating tasks that they otherwise would have required manual analysis and not music classification or music similarity analysis as an end goal.

2.1.6 Gamera

Gamera[6] is Optical Symbol Recognition Open Source software based on supervised and hybrid learning approaches for training. Gamera is designed with the particular aim of symbol recognition of old documents. Gamera also supports creating of new plugins for custom tasks. For the purpose of Music Information Retrieval, gamera can be used to solve the problem of Optical Music Recogntion (OMR) since sheet music images are also a format for music source.

2.1.7 Audiveris

Audiveris is an Open source software for Optical Music Recognition. Unlike Gamera, Audiveris can be directly consumed as a service for the purpose of OMR. Audiveris is used as service in many leading Notation Platforms like Musescore etc. As such, Audiveris is used as a subcomponent of Modulo7's architecture for OMR.

2.2 Typical problems of MIR

On top of the generic software created by researchers and industry experts, researchers have tackled specific problems in Music Cognition, Classification

Chapter 3

Basics of Music Theory

Music theory is defined as the systematic study of the structure, complexity and possibilties of what can be expressed musically. More formally its the academic discipline of studying the basic building blocks of music and the interplay of these blocks to produce complex scores (pieces of music). Modulo7 is built on top of western theoretic principles and hence only western music theory is explored. Also music theory is an extremely complicated subject and hence only the basics and relevant portions to the modulo7 implementation are discussed here.

Traditionally music theory is used for providing directives to a performer to play a particular song/score.

This section is primarily meant for people with a weak or lack of understanding

of music theory. The following section talks about the basic building blocks of music theory:-

3.1 Building Blocks

Music is built on fundamental quantities (much like matter is built on fundamental quantities like atoms and molecules). The following are the core concepts in order of atomicity (i.e successive blocks build on the preceding ones)

Pitch/Note: A pitch is a deterministic frequency of sound played by a musical voice (instrument or human). In western music theory, certain deterministic pitches are encoded as Notes. For example the note A4 is equal to 440 Hz. In other words Notes are symbolic representations of certain pitches. With certain notable exceptions, most music is played on these set frequencies.

Each note is characterized by two entities. First is the note type and the second is the octave. An octave can be considered as a range of 12 notes. There are 8 octaves numbered 0 to 7 which are played by traditional instruments or vocal ranges. Then is the note type. Notes are categorized into 7 major notes (called A, B, C, D, E, F, G) and 5 minor notes (also called as accidentals). They can be characterized by increasing or decreasing the frequency of the notes by a certain amount

(called sharps(#) and flats(b) respectively). For example the accidental lying in between (A and B is called A# or Bb). Similarly accidentals lie in between C, D; D, E; F, G and G, A. (Note that there are no accidentals in between B and C and E and F).

Semitone and Tone: A semitone is an increment or a decrement between two notes. For instance there is one semitone in between A and A#. Similarly there are 3 semitones in between A and C. A tone is an increment in between two major notes. Another characterization of a tone is two semitones.

Beat/Tick: A beat or tick is a rhythmic pulse in a song. Beats in sequence is used to maintain a steady pulse on which the rhythmic foundations of a song is based.

Pitch duration: A pitch duration is a relative time interval the pitch persists on a musical instrument. For example a whole note will persist twice as longer as a half note.

Attack/Velocity: The intensity or force with which a pitch is played. This parameter influences the loudness of the note and in general the dynamics of the song (covered in a subsequent section)

Chord: A chord is a set of notes being stacked together (being played together at or

almost at the same time). Chords are the basic building blocks of a concept called harmony (which will be discussed further on.). Traditionally a chord is constructed by stacking together notes played on a single instrument, but a chord can be constructed by different instruments simultaneously playing different notes.

Rests: Rests are pauses in between notes (with no sound being played at that point of time) for a fixed duration.

Melody: A melody is a succession of notes and rests which sound pleasing. There are many rules about what makes a melody sound good which we will get to in the subsequent reading.

Harmony: A harmony is a succession of chords (also known as a chord progression) along with the principles that govern the relationships between different chords.

Voice: A voice is an interplay of notes, chords and stops by a single instrument/vocalist. The reader can think of a voice as a hybrid or generalization of the melody and harmony concepts.

Score/Song: A score or a song is an interplay of voices. It is the final product of music that is delivered to the audience. Songs are of different types based on cultural

context and complexity (for example an orchestra is a large number of voices being coordinated by a conductor. In contrast a folk song might be played by a single person on a guitar or a duet between a vocalist and an instrumentalist).

Interval: An interval is the relative semitone distance between any two notes. Intervals are categorized as melodic(semi tone distance between successive notes in a melody) and harmonic intervals (semi tone distance between notes within a chord).

3.2 General Concepts in Music Theory

On top of the building blocks of music, there are certain generic ideas or concepts on which music is based. The following sections describe them :-

Polyphony/Homophony: A homophonic song involves exactly one voice in the song. An example would be a single person singing a tune. A polyphonic song is one which involves two or more voices transposed with one another. An example of polyphonic music would be a Western classical orchestra of a rock band performing a chorus.

Phrase: A musical phrase is a subset of the song that has a complete musical sense of its own. One could think of phrases as musical sentences, whereas a voice could be

considered a paragraph. As a side effect a musical phrase can be played independently and still be considered as a song albeit an incomplete one.

Meter: The meter of a song is an expression of the rhythmic structure of a song. In context of western classical music, its a representation of the patterns of accents heard in the recurrence of measures of stressed and unstressed beats. Meters dictate the rhythm or tempo in which a song is played.

Key/Tonality: Tonality or key of a song is a musical system in which pitches or chords are arranged so as to include a hierarchy of relation between musical pitches, stabilities and attractions between various pitches. For example if the song is in the key of C, C is the most stable pitch in that song and other pitches like B have a tendency to go towards C (also called resolution of a phrase) to inculcate a sense of completeness. Moreover other pitches in relation to this pitches have various degrees of stability.

Scale: A scale of a song is an ordered set of notes starting from a fundamental frequency or pitch. If viewed ascendingly or descendingly (increasing/decreasing frequency of the pitches respectively) on this ordering, a scale describes a relationship between successive notes and their semitone distances from each other. A scale restricts the set of notes being played once the fundamental pitch is determined.

Key Signature: A key signature is a key along with a scale defined for a song (or in other words the fundamental pitch of the scale of the song is the same as the key of the song). A key signature is an expression of coherence for a song as well as a well defined set of notes that can be played for this piece, and as a result a song does not have notes that are outside of this key signature.

Chromatic Music: Chromatic music is any music that does not have a well defined key signature. Alternatively chromatic music can be categorized as music which is in the chromatic scale (chromatic scale is a scale in which all semitones in western music is present). Chromatic music is more difficult to analyze due to its lack of structure.

Melodic Contour: Melodic contour is the "shape" of melody. A melody with pitches going monotonically upward in frequency is called an ascending contour. Similarly a melody going monotonically downwards in frequency is called a descending contour. There are many other kinds of contour in music theory.

Dynamics: Dynamics is a coarse idea which indicate the variety of relative loudness between notes, speed of notes being played across phrases and other such ideas.

Chapter 4

Mathematics of Modulo7

The following sections describe the mathematical concepts used and implemented in Modulo7.

4.1 Vector Space Models of Music

In traditional text based information retrieval retrieval systems, documents are indexed and a vector space representation of documents are created. Typical approaches for counting term frequencies or some weighting scheme like Term Frequency-Inverse Document Frequency Approach (TF-IDF). Analogous to text based IR, Music data can also be expressed as a vector space based on the approach taken. Some of these approaches are taken from the SIMILIE[?] but generalized for polyphonic music. Many approaches are novel based on the author's music theoretic studies.

4.1.1 Pitch Vector

A voice can be expressed as a sequence of pitches $n_i = (p_i, t_i)$ where p_i is the pitch and/or the set of pitches at instant of time t_i. The symbolic representation of music essentially discretizes these values from music sources and hence a vector representation can be made. A voice V can be represented as a vector :-

$$P =< n_1, n_2, ...n_n > \tag{4.1}$$

A similar vector representation could be when the time information is eschewed in favor on only the pitches. This vector is called the raw pitch vector and is denoted as the follows :-

$$R =< p_1, p_2, ..., p_n > \tag{4.2}$$

4.1.2 Pitch Interval Vector

Another way to look at elements is the interval spacing between elements. This is same as the interval concept in the music theory chapter. Mathematically an interval is defined as $\Delta p_i = p_i - p_{i-1}$. And thus an pitch interval vector is defined as :-

$$PI =< \Delta p_1, \Delta p_2, ..., \Delta p_n > \tag{4.3}$$

4.1.3 Rhythmically weighted Pitch Interval Vector

In order to include the rhythmic information in the pitch interval Vector, define rhythmically weighted pitch as $rp_i = \Delta p_i \times t_i$. Now the rhythmically weighted pitch vector can be represented as:-

$$RPI =< rp_1, rp_2, ...rp_n > \tag{4.4}$$

4.1.4 Normalized Tonal Histogram Vector

The tonal histogram is a vector or map of 12 distinct intervals present in western music theory. Each position in the vector corresponds to the total number of times that interval has occurred in a voice. Mathematically define $\Delta P_i^{voice_j} = \sum_{i=1}^{len(voice)} p_i^{voice_j}$

Chapter 5

Software architecture

The following sections present the software architecture of Modulo7.

5.1 Server Side architecture

Modulo7 is designed with the purpose of scalability. A block diagram of the components of the server side architecture is presented below :-

1. Source Converter : Converts music sources (e.g. music XML, midi etc) into modulo7's binary representation.

2. Music Theory Models : The model is a description of music theoretic criteria that can be applied on top of a song. Examples would be melodic contour, tonal histogram etc.

3. Distributed Storage Mechanism : The modulo7 internal representation is a

conversion to create a song representation with all the meta data of the song (Key, Scale, etc) along with the sequences of note events stored as lists. This representation is then serialized and stored in and Hadoop Distributed File System. This allows for fault tolerance and a distributed deployment of the input data.

4. Lyrics Indexer : A distributed index of songs lyrics. This acts as a base on which standard techniques for similarity analysis might be applied. Alternatively it can provide a framework on which custom models (e.g. semantic intent of the song, correlation between music theory models and lyrics might also be applied).

5. Lyrics similarity models : A set of similarity models that can be applied to an index.

6. Query Engine : An SQL like interface to a client that allows you to gather and ascertain useful information (based on music theoretic criteria).

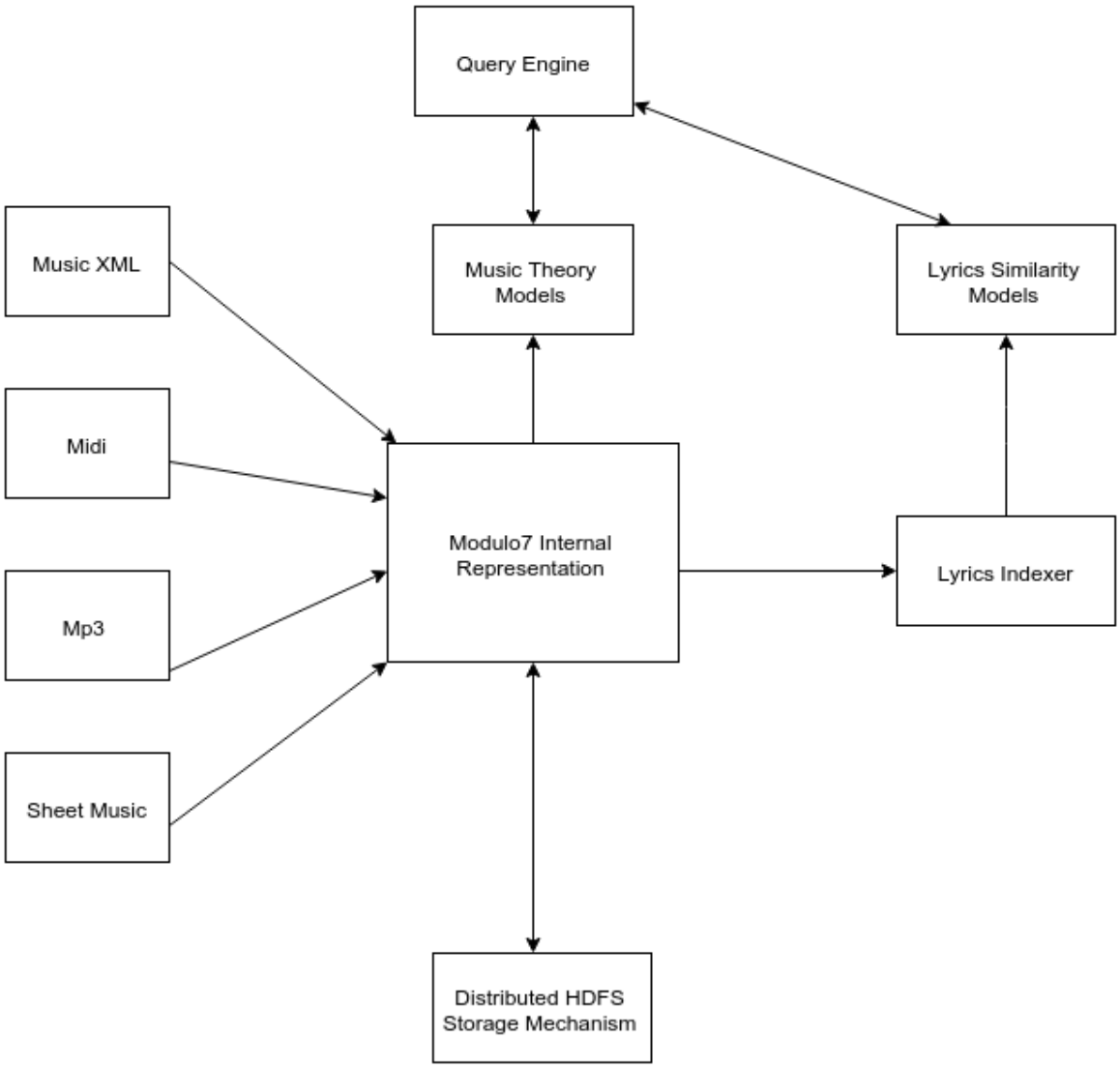

Figure 5.1: Modulo7 architectural design

5.2 Client architecture

The server exposes a sql like interface as well as a consumable API. Some sample queries would be :-

1. select midi files from database where $melodic\_complexity > somethreshold$
2. select * from database where $artist = led\_zepplin\ and\ harmonic\_movement > harmonic\_movement(stairway\_to\_heaven)$
3. select $num\_voices\ from\ Database\ where\ songName = someSong.midi$

An API will also be exposed to the client along a remote invocation procedure. The API would primarily target single sources for specifics. Some example API would be :-

1. int getNumVoices(String midiFilePath)
2. double melodicContourMovement(String pngSheetFilePath)
3. double compareAverageAttack(String musicXMLFile)

This API can be consumed for specific song analysis. As design this API will not work on a bulk of files like its sql counterpart.

Moreover the client also exposes a highly customized search engine based on the custom vector space representation of features extracted by Modulo7.

5.3 Song sources

At the heart of Modulo7's design is its song sources adaptors (or converters) into its own internal binary format. Each music source is a different representation and while certain sources ascribe what how music should be played (e.g musicxml, sheet music), other formats ascribe what is actually being played (e.g midi, mp3). There are many other music sources in existence (e.g guitar tablature, GUIDO format, humdrum format), but for the purposes of breadth and ubiquity, these four sources have been targeted as input for Modulo7. Note that acquiring features from each format is a domain specific challenge and inaccuracies are inherent because of that. The following subsections describe the individual formats in detail and the challenges encountered in parsing them.

5.3.1 Midi format

MIDI (short for Musical Instrument Digital Interface), is a technical specification for encoding of events on a midi enabled instrument and a protocol for interfacing and communicating between various midi enabled instruments. Typically any midi enabled electronic instrument when played, relays to its internal circuitry a message. Examples of such messages could be a particular note is being hit on a keyboard, a note is being hit off after being hit on, tempo based messages on the number of ticks per second etc. While MIDI is a technical specification for encoding music the

score is being played, Modulo7 treats it as a symbolic representation of music. Midi was also a simple and popular encoding format for music and gaming industry in the nineteen ninties.

A symbolic representation is a codification of music which acts a higher level of abstraction (individual notes or chords being played) as compared to lower level representations like audio files (which codify information like waveforms). Modulo7's internal representation is also a symbolic representation. Symbolic representations are easier to manipulate when applying a music theoretic criteria.

Midi is one of the easier formats to parse for musical specifications. Moreover there is a big volunteer community of midi encoders. As such acquiring and parsing non trivial amounts of midi data is not a very challenging task.

5.3.2 Western Sheet Music

Sheet music is one of the oldest forms of music in existence. Its a hand written or printed form of music that uses a specific script (a set of musical symbols on a manuscript paper) to ascribe music. Music Composers from Medieval and Modern periods of the western world use western sheet scripting to codify their work while performers play from these sources. A vast body of older work and particularly orchestral work is codified in sheet music.

Like midi, sheet music is also symbolic in nature. However unlike midi, its an ex-

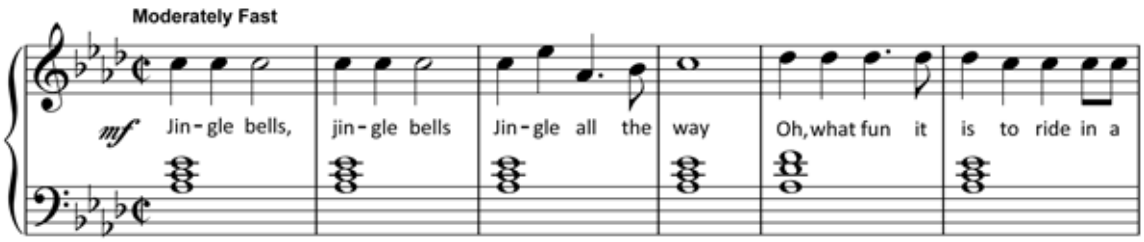

Figure 5.2: Jingle bells melody sheet music representation

pression of how a score should be played, rather than what is being played. Modulo7 converts digitized versions of these sheet music (e.g sheet music stored .tiff, .png. jpeg etc formats)

A very simple example of sheet music for describing a melody is shown below.

Parsing digitized sheet music is an extremely challenging task. It requires a solid understanding on Computer Vision and even the state of the art software in existence today cant handle all scores (especially a poorly digitized formats). Given the amount of domain knowledge required, Modulo7 uses a third party library (insert TPL here).

5.3.3 Music XML format

Music XML format is a standard open format for exchanging digital sheet music. A music XML format is unusual as its a format that is easy to parse for computers and easy for humans to understand it. MusicXML formats are heavily used by music notation applications. Music XML format is a symbolic format and can be considered

a modernization of the Sheet music format. Its disadvantage however is unlike sheet music, a performer cant read the piece and play it on the spot directly.

Just like Western Sheet music and midi, music XML is a symbolic format as well. Music XML is also a transcription format which specifies how a score should be played.

5.3.4 MP3 format

For the sake of completeness, Modulo7 also supports an audio format called mp3. Its an audio encoding format that uses lossy compression to encode audio data. Mp3 gives a reasonably good approximation to other digital audio formats of music storage with a significant savings in space for storage. Its one of the defacto standards of digital music compression and transfer and playback on most digital audio players.

5.4 Modulo7 Internal Representation

Modulo7 consists of converters that convert data into Modulo7's internal representation. This representation can be thought of a document representation on which similarity measures described in Chapter 4 can be applied to.

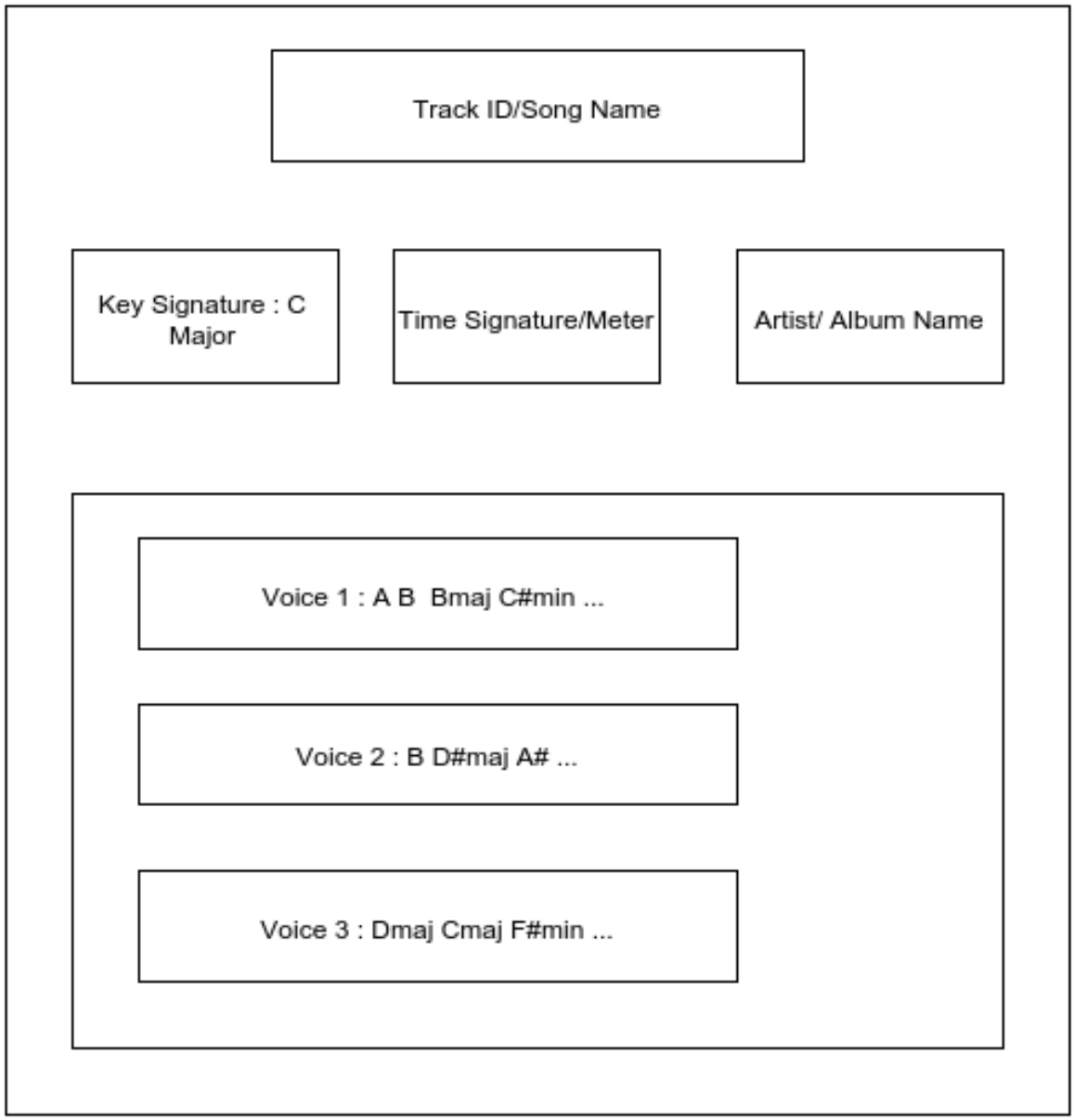

Figure 5.3: Modulo7 internal representation

Chapter 6

Experimental Evaluation

For the purposes of evaluating Modulo7

Bibliography

[1] C. McKay, *Automatic Music Classification with jMIR*. Montreal: McGill University, 2010.

[2] M. D. P. Adam M. Stark, "Real-time chord recognition for live performance."

[3] P. G. Tzanetakis, "Marsyas a framework for audio analysis," *Organized Sound*, vol. 4(3).

[4] D. M. Klaus Frieler, "The simile algorthms for melodic similarity."

[5] "The humdrum toolkit: Reference manual. menlo park, california: Center for computer assisted research in the humanities, 552 pages, isbn 0-936943-10-6." p. 552 pages.

[6] I. F. Karl MacMillan, Micheal Droettbroom, "Gamera: Optical music recogntion in a new shell."