A Music Data Mining and Retrieval Primer

Dan Berger
dberger@cs.ucr.edu

May 27, 2003

Abstract

As the amount of available digitally encoded music increases, the challenges of organization and retrieval become paramount. In recent years, an active research community has embraced these challenges and made significant contributions in the problem space.

This paper attempts to summarize the current conceptual state of music (as contrasted with the more general term "audio") data mining and retrieval.

1 Introduction

Since the invention of the compact disc by James Russell in the late 1960's [6] and it's mass-market release in the early 1980's[1], consumers have stored music predominantly in the digital domain. The development and wide-spread adoption of psychoacoustic encodings such as MP3 [29] and OGG Vorbis [34] have enabled large collections of music to be easily accessible.

As with other media, as the amount of available digitally encoded music increases, the challenges of organization and retrieval become paramount. In recent years, an active research community has embraced these challenges and made significant contributions in the problem space.

This paper attempts to summarize the state of music data mining and retrieval research as focused on challenges such as content-based-query, feature extraction/similarity measurement, clustering and categorization. While familiarity with data mining and information retrieval concepts is assumed, all required musical background is provided herein.

The rest of the paper is organized as follows; Section 2 motivates the problem at a high level, Section 3 briefly provides musical background relevant to the topics discussed, and Section 4 describes the characteristics of various digital representations of music in common use - including a brief introduction to psychoacoustic models. Section 5 outlines the various forms of query which retrieval systems attempt to facilitate, Section 6 discusses work done on the problem of content based query, including similarity measures and feature extraction, and Section 7 revisits some similar issues while looking at the state of categorization and clustering. Finally, Section 8 poses some open questions and concludes the paper.

2 Motivation

> "Voices." The founder from Los Angeles was staring at Case. "We monitor many frequencies. We listen always. Came a voice, out of the babel of tongues, speaking to us. It played us a mighty dub."
>
> "Call 'em Winter Mute," said the other, making it two words.
>
> Case felt the skin crawl on his arms.
>
> ...
>
> "Listen," Case said, "that's an AI, you know? Artificial intelligence. The music it played you, it probably just tapped your banks and cooked up whatever it thought you'd like..."
>
> – William Gibson; *Neuromancer*, 1984

As the size of the digital music collection available to an individual grows, their ability to directly organize and navigate the collection diminishes. In addition to currently possible queries based on obvious meta-data (such as artist, title, and genre), it becomes increasingly valuable to be able to express high-level queries such as "find me other songs in my collection similar to the one I'm listening to", or "I'd like a coherent play list of light jazz, instrumental blues, and ambient techno." [2]

[1] For an interesting history of the compact disc starting in 1841 see http://www.oneoffcd.com/info/historycd.cfm

[2] Work described in [24] can almost answer this query today, but relies on high quality meta-data not generally available.

For music professionals - musicians, foley artists, producers and the like - the ability to rapidly and effectively search a music database to find the "right" piece of music for their current requirements is substantively similar. (While not strictly on topic for this survey, [33] provides an overview of a system called SoundFisher which does for "simple" audio many of the things discussed here for music.)

From a commercial perspective there are already companies attempting to leverage results in this space, including Hit Song Science [1], which uses "spectral deconvolution" to compare submitted music with a database of top 30 hits to deliver a rating which they claim represents the likelihood of the submitted song being a hit. Relatable [5] offers "TRM" - an audio fingerprinting technology which reports to be "based on an analysis of the acoustic properties of the audio itself."

Commerce ventures such as Amazon.com and the recently launched Apple iTunes music store have deployed collaborative filtering in an effort to match buyers to music - they would view much of this work in the context of enabling consumers to locate artists and albums which are similar to works for which they have already expressed an affinity.

3 Music Background

3.1 Types of Music

Music is often divided into three categories based on the amount of concurrency present:

1. **Monophonic**: music in which only one note sounds at a time. Very little "real" music fits into this category, but [14] discusses a method whereby more complex music can be decomposed into several correlated monophonic scores.

2. **Homophonic**: music in which multiple notes may sound at once - but all notes start and finish at the same time. The left hand of a piano performance, or folk guitar performance, is often homophonic - producing chords rather than a series of individual notes.

3. **Polyphonic**: the most general form of music, in which multiple notes may sound independent of each other.

4 Digital Representation of Music

There are several digital representations of music in use today. These representations can be ordered in terms of the amount of **musical structure** they retain, as well as their **fidelity** - or ability for faithful reproduction.

Symbolic formats, such as MIDI [2], represent music at a relatively high level - encoding information such as note durations and intensities. It is important to note that a MIDI file doesn't contain audio data, rather it contains instructions for synthesizing instrumental audio. A large amount of musical structure is captured, but the resulting representation is unable to capture the nuance of live performance.

Sampled formats, such as Pulse Code Modulation (PCM), represent music, or any other analog data, by periodically sampling the data, quantizing the sample, and encoding the quantized value digitally.

Required parameters are the sample rate, expressed in cycles per second, and bits per sample. PCM is unable to explicitly represent any musical structure By Nyquist/Shannon's sampling theorem, however, it is possible to faithfully represent the target signal provided the sample rate is at least twice the maximum frequency to be captured. Some researchers, such as Dannenberg and Hu [13], have examined the problem of rediscovering musical structure in unstructured formats.

By far the most common sample rate is 16bit 44.1kHz, as used by compact disc, though other sample sizes and rates, such as 24bit 48kHz are also used (by Digital Audio Tape, for example).

Additionally, **compressed** formats, such as MP3 and OGG Vorbis, which use psychoacoustic models (see Section 4.1) of human hearing to discard irrelevant or imperceptible data from a PCM bit stream and produce a perceptually comprable, but distinct, bit stream significantly smaller than the "raw" PCM data. Note that while decoding a compressed source results in a PCM stream, it does not produce the input PCM bit-for-bit. Because of the lossy transformation performed by these encoding schemes, it can be argued that they retain even less of the original structure than the input PCM.

4.1 Psychoacoustics

Broadly, psychoacoustics is the study of human auditory perception, [23] gives a brief overview which is adapted here, and [12] has additional information and exposition.

The two main properties of the human auditory system which factor into psychoacoustic models are:

1. limited, frequency dependent resolution
2. auditory masking

4.1.1 Limited Resolution

Empirical results show that the human auditory system has a limited, frequency dependent resolution. We can hear sounds between 20Hz to 20,000kHz (hence the 44.1kHz compact disc sampling rate - slightly greater than twice the maximum audible frequency).

Further - results show that the audible frequency range can be divided into a number of "critical bands" - within which a listener has difficulty distinguishing the frequency of sounds. These critical bands are referred to as *barks* after the scale created from the empirical measurements. The bands range from very narrow (100Hz) at low frequencies to very wide (4kHz) at high frequencies and are non-linear.

4.1.2 Auditory Masking

Empirical results also show that when a strong audio signal is present, humans are unable to hear weak signals in it's temporal and spectral "neighborhood." This masking caused or experienced by a given signal is highly dependent on the critical band in which it falls. This behavior is the basis for lossy audio compression algorithms, though different algorithms exploit these factors in different ways.

5 Query Modes

In [27], Selfridge discusses the key differences between querying a music collection and querying a text collection - stating that most useful music queries will be "fuzzy" - and concluding that "these subtleties beg for suitable ways of searching which are likely as heterogeneous as the repertories themselves."

A few predominant query mechanisms have emerged from the literature:

1. Meta-Data based Query - simple queries based on meta-data rather than on musical content. Examples include query by title, author, genre, etc. The key challenge in this case is obtaining (or generating) and maintaining objectively accurate and correct meta-data.

2. Content based Query - more specifically referred to as "aural query" [8], "query-by-humming" [17], "sung query" [18] and "query by melody" [28]- given a (short) sample, return pieces containing, or similar to, the input. Key challenges include signal processing, in the case when the input is hummed or sung, to feature extraction and determining similarity.

Additionally, browsing, or exploration is recognized as a legitimate mode of use. Save for few examples, such as [25], which describes the implementation of a system intended sole for exploration; and [22], which focuses on visualizing music archives via self-organizing maps, there exists little work on exploration per-se. Rather browsing is often mentioned in the context of similarity, clustering and categorization work.

6 Content Based Query

The ultimate goal of a music query system is to be able to return *subjectively meaningful* results to similarity queries. In general this is a hard problem when dealing with highly dimensional data, it is even more so when dealing with music.

In many respects, feature extraction and the related problem of similarity measures are crux of the content based query problem. While human hearing is becoming a solved problem at the level of psychoacoustics, higher order understanding - "listening" - is still very much an open question. A small number of "machine listening" groups, such as the one at the MIT Media Lab [3], have begun attacking these problems.

While music can be though of as a time series, it's extraordinarily high dimensionality (a single PCM stream of a 5 minute song at cd quality has over 13 million samples) seems to preclude directly treating it as such. Hence, techniques for dimensionality reduction are key to efficient mining of musical data.

Additionally, while no explicit references were found in the research literature, the adoption of psychoacoustic based compression algorithms further complicate the matter. Two decoded encodings of the same input source can differ substantively. As a simple example, the two encodings could be at different bit-rates - causing more interpolation to occur during decoding.

More subtle variations are possible as well; different implementations of the psychoacoustic model will cause different information to be discarded during encoding - so the PCM streams that result from decoding encodings of same source using two different encoders, even at the same bit rates, may vary dramatically. Hence dimensionality reduction techniques must be robust to these factors.

Attempts to capture and quantify features suitable for use in similarity comparisons range from signal processing techniques - such as tempo and beat detection and analysis [26] - to attempts to leverage higher-order information such as tempo/rhythm and melody [31]. We examine a handful of the more notable efforts here.

In [32] Welsh et. al propose a set of admittedly ad-hoc features which they use to decompose a collection of 7000 songs, stored in MP3 format, into 1248 feature dimensions. They choose to represent a song by it's tonal histograms and transitions, noise, volume, and tempo and rhythm. They discuss their results and point readers to an on-line jukebox into which they have incorporated their work. Unfortunately, that on-line jukebox is no longer accessible.

In [7] Aucouturier and Pachet propose a similarity measure based on timbre[3] and evaluate it's performance in a content based query setting. They then conclude by proposing that similarity measures are not useful individually, but only useful in their intersection - when multiple measures are juxtaposed.

Burges, Plat and Jana discount the use of what they call "heuristic audio features," and propose a dimensionality reduction technique called Distortion Discriminant Analysis in [10]. They demonstrate the ability to identify audio clips in an audio stream against stored audio with good accuracy.

To further complicate attempts to use more high-level characteristics, acts which human listeners perform intuitively - such as determining what instrument created a given sound, or identifying a singer by their voice, turn out to be quite difficult.

A fairly early step toward sound source recognition was taken by Ellis in [15], when he described a psychoacoustic model for detecting events in an acoustic signal that a human listener would perceive as different objects. While the system was admittedly ad-hoc and problem specific, it served as the foundation for future, more general, work.

In [9] the author discusses training a computer to recognize sounds created by two specific woodwind instruments (oboe and saxophone) and compares her results to human listening tests. More generally, [20] examines the acoustic characteristics of instruments which might be suitable for source recognition, and [21] builds a theoretical basis for performing sound source recognition and describes a system which can listen to a recording of an instrument and classify it as one of 25 known "non-percussive orchestral" possibilities.

In [19], Kim and Whitman propose a voice-coding based technique for identifying the singer in pop music recordings. Their initial results are better than chance, but in their words "fall well short of expected human performance." They conclude the work by enumerating possibilities for improving the accuracy of their technique.

7 Categorization and Clustering

A related but distinct problem in the music retrieval space is that of classification and clustering. As in the case of the query problem, clustering can be based on intrinsic or extrinsic characteristics. We will focus primarily on the intrinsic case in this section.

In general the clustering and classification problem overlaps significantly with the content-based query problems discussed in Section 6. After applying a suitable dimensionality reduction to the input music data, the results are clustered or categorized by traditional methods such as hierarchical or partitional clustering, K-means, etc. Here we look at a selection of the more novel or influential contributions.

In [11] the authors break a bit from the traditional mold and present a scheme for classifying folk music into known categories (corresponding to the songs country of origin) using hidden Markov models. They compared their classification technique using four different representations of the melodies - which were obtained in highly structured symbolic formats called `**kern` and `EsAC`. While interesting, the reliance on highly structured input data diminishes the value of this as a general purpose technique.

Foote, in [16], presents results from using a "supervised tree-based vector quantizer trained to maximise (sic) mutual information (MMI)." Notable in this approach is it's complete disregard for perceptual criteria - it is complete data driven, and (unlike the aforementioned work using hidden Markov models) computationally efficient. The presented results are graphically well clustered, but no claims are made as to the subjective correctness of the produced clusters.

Tzanetakis et. al use a 17-dimensional feature vector composed of 9 so-called "surface features" and 8 rhythmic features to automatically classify music into genres in [30]. They include information on two different user interfaces built on their proposed tech-

[3] defined as "...that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."

nique - one called "GenreGram" which provides a dynamic real-time display of the current classification probabilities during on-line classification, and one called "GenreSpace" which allows the user to navigate through a 3 dimensional reduction of the feature space.

In [22] Pampalk et. al present *Islands of Music*, a system which uses self-organizing maps (SOMs) as the interface to visualize and explore the clusters created by their proposed feature extraction scheme which incorporates specific loudness and rhythm patterns as inputs to determine similarity. While the scheme works as designed, the choice of similarity metric often results in unintuitive clustering decisions on the 259 element input set. Additionally, it is unclear if the SOM approach will meaningfully scale to more realistic collection sizes.[4]

8 Conclusion

We have attempted to present an introductory overview of the state of data mining and information retrieval as applied to digital representations of music. While the psychoacoustic model of human hearing is getting closer to being a solved problem, the higher order comprehension of music is far from being so. Ironically, it is precisely this higher order understanding which we would most like to exploit in managing large music collections.

An interesting, and seemingly unexplored question is the effect that various psychoacoustic encoders have on the similarity metrics which have been proposed to date. While a single individuals collection may be encoded with a single implementation of the encoder, it is unlikely that large collections - such as the 300,000 titles offered by on-line music service PressPlay [4] would have that characteristic. Given that, how should one best insulate their similarity metrics against the differences in encoded output? It may be the case that psychoacoustic similarity metrics are fairly immune to these differences, but that requires empirical verification.

While progress is being made in the areas of feature extraction, similarity measures, and categorization and clustering - the ultimate goal of imparting some semblance of "musical appreciation" to our software systems seem far off.

[4] The current authors personal music collection consists of over 2900 tracks - an order of magnitude more than the sample set in this work.

References

[1] Hit song science. http://www.hitsongscience.com.

[2] Midi: Musical Instrument Digital Interface. http://www.midi.org.

[3] Music, mind and machine. http://sound.media.mit.edu.

[4] Pressplay. http://www.pressplay.com.

[5] relatable. http://www.relatable.com.

[6] James T. Russell. http://web.mit.edu/invent/iow/russell.html, December 1999.

[7] Jean-Julien Aucouturier and Francois Pachet. Music Similarity Measures: What's the Use? In *Proceedings of Int. Symposium on Music Information Retrieval (ISMIR)*, 2002.

[8] W. Birmingham, R. Dannenberg, G. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, and W. Rand. Musart: Music Retrieval Via Aural Queries. In *Proceedings of Int. Symposium on Music Information Retrieval (ISMIR)*, 2001.

[9] Judith Brown. Computer Identification of Musical Instruments using Pattern Recognition with Cepstral Coefficients as Features. http://sound.media.mit.edu/ brown/iid.ps, 1997.

[10] Christopher J.C. Burges, John C. Platt, and Soumya Jana. Extracting Noise-Robust Features From Audio Data. In *Proceedings of Int. Conference on Acoustics Speech and Signal Processing.* IEEE, 2002.

[11] Wei Chai and Barry Vercoe. Folk Music Classification Using Hidden Markov Models. In *Proceedings of the Int. Conference on Artificial Intelligence*, 2001.

[12] Alex Chen, Nader Shehad, Aamir Virani, and Erik Welsh. W.A.V.S. compression. http://is.rice.edu/ welsh/elec431/psychoAcoustic.html.

[13] R. Dannenberg and N. Hu. Discovering musical structure in audio recordings. In *Proceedings of Int. Conference on Music and Artificial Intelligence (ICMAI)*, 2002.

[14] Shyamala Doraisamy and Stefan M. Rger. An Approach Towards A Polyphonic Music Retieval System. In *Proceedings of Int. Symposium on Music Information Retrieval (ISMIR)*, 2001.

[15] Daniel PW Ellis. A computer implementation of psychoacoustic grouping rules. Technical report, MIT Media Lab, 1994.

[16] Jonathan Foote. A similarity measure for automatic audio classification. In *Proceedings of Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*. American Association for Artificial Intelligence (AAAI), 1997.

[17] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of ACM Multimedia*, pages 231–236, 1995.

[18] Ning Hu and Roger Dannenberg. A comparison of melodic database retrieval techniques using sing queries. In *Proceedings of Joint Conference on Digital Libraries*, 2002.

[19] Youngmoo E. Kim and Brian Whitman. Singer Identification in Popular Music Recordings Using Voice Coding Features. In *Proceedings of Int. Symposium on Music Information Retrieval (ISMIR)*, 2002.

[20] Keith D. Martin. Toward Automatic Sound Source Recognition: Identifying Musical Instruments. In *Proceedings of NATO Advanced Study Instituite On Computational Hearing*, 1998.

[21] Keith D. Martin. *Sound-Source Recognition: A Theory and Computational Model*. PhD thesis, Machine Listening Group, MIT Media Lab, 1999. http://sound.media.mit.edu/Papers/kdm-phdthesis.pdf.

[22] E. Pampalk, A. Rauber, and D. Merkl. Content-based Organization and Visualization of Music Archives. In *Proceedings of ACM Multimedia*, 2002.

[23] Davis Pan. A Tutorial on MPEG/Audio Compression. *IEEE Multimedia*, pages 60–74, 1995.

[24] John C. Platt, Christopher J.C. Burges, Steven Swenson Christopher Weare, and Alice Zheng. Learning a gaussian process prior for automatically generating music playlists. Technical report, Microsoft Research, 2002.

[25] J. Polastre, C. Heyl, and M. Noori. Loud: An Immersive Music Exploration System. Technical report, University of California, Berkeley, 2002.

[26] Eric D. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, 103(1):588–601, 1998.

[27] Eleanor Selfridge-Field. What Motivates a Musical Query? In *Proceedings of Int. Symposium on Music Information Retrieval*, 2000.

[28] Shai Shalev-Shwartz, Shlomo Dubnov, Nir Friedman, and Yoram Singer. Robust temporal and spectral modeling for query by melody. In *Proceedings of ACM SIG Information Retrieval*, pages 331–338, 2002.

[29] Thompson. Audio revolution: the story of mp3. http://www.mp3licensing.com/mp3/index.html.

[30] George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals. In *Proceedings of Int. Symposium on Music Information Retrieval (ISMIR)*, pages 205–210, 2001.

[31] A. Uitdenbogerd and J. Zobel. Melodic Matching Techniques for Large Music Databases. In *Proceedings of ACM Multi Media*, 1999.

[32] M. Welsh, N. Borisov, J. Hill, R. von Behren, and A. Woo. Querying Large Collections of Music for Similarity. Technical report, University of California, Berkeley, 1999.

[33] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 1996.

[34] xiph.org. Ogg Vorbis: open, free audio. http://www.vorbis.com/.