INVITED PAPER

Content-Based Music Information Retrieval: Current Directions and Future Challenges

Current retrieval systems can handle tens-of-thousands of music tracks but new systems need to aim at huge online music collections that contain tens-of-millions of tracks.

By Michael A. Casey, *Member IEEE*, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney, *Senior Member IEEE*

ABSTRACT | The steep rise in music downloading over CD sales has created a major shift in the music industry away from physical media formats and towards online products and services. Music is one of the most popular types of online information and there are now hundreds of music streaming and download services operating on the World-Wide Web. Some of the music collections available are approaching the scale of ten million tracks and this has posed a major challenge for searching, retrieving, and organizing music content. Research efforts in music information retrieval have involved experts from music perception, cognition, musicology, engineering, and computer science engaged in truly interdisciplinary activity that has resulted in many proposed algorithmic and methodological solutions to music search using content-based methods. This paper outlines the problems of content-based music information retrieval and explores the state-of-the-art methods using audio cues (e.g., query by humming, audio fingerprinting, content-based music retrieval) and other cues (e.g., music notation and symbolic representation), and identifies some of the major challenges for the coming years.

KEYWORDS | Audio signal processing; content-based music information retrieval; symbolic processing; user interfaces

I. INTRODUCTION

Music is now so readily accessible in digital form that personal collections can easily exceed the practical limits on the time we have to listen to them: ten thousand music tracks on a personal music device have a total duration of approximately 30 days of continuous audio. Distribution of new music recordings has become easier, prompting a huge increase in the amount of new music that is available. In 2005, there was a three-fold growth in legal music downloads and mobile phone ring tones, worth $1.1 billion worldwide, offsetting the global decline in CD sales; and in 2007, music downloads in the U.K. reached new highs [1]–[3].

Traditional ways of listening to music, and methods for discovering music, such as radio broadcasts and record stores, are being replaced by personalized ways to hear and learn about music. For example, the advent of social networking Web sites, such as those reported in [4] and [5], has prompted a rapid uptake of new channels of music discovery among online communities, changing the nature of music dissemination and forcing the major record labels to rethink their strategies.

Manuscript received September 3, 2007; revised December 5, 2007. This work was supported in part by the U.K. Engineering and Physical Sciences Research Council under Grants EPSRC EP/E02274X/1 and EPSRC GR/S84750/01.

M. A. Casey and **C. Rhodes** are with the Department of Computing, Goldsmiths College, University of London, SE14 6NW London, U.K. (e-mail: m.casey@gold.ac.uk; c.rhodes@gold.ac.uk).
R. Veltkamp is with the Department of Information and Computing Sciences, Utrecht University, 3508TB Utrecht, The Netherlands (e-mail: Remco.Veltkamp@cs.uu.nl).
M. Goto is with the National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki 305-8568, Japan (e-mail: m.goto@aist.go.jp).
M. Leman is with the Department of Art, Music, and Theater Sciences, Ghent University, 9000 Ghent, Belgium (e-mail: Marc.Leman@UGent.be).
M. Slaney is with Yahoo! Research Inc., Santa Clara, CA 95054 USA (e-mail: malcolm@ieee.org).

Digital Object Identifier: 10.1109/JPROC.2008.916370

It is not only music consumers who have expectations of searchable music collections; along with the rise of consumer activity in digital music, there are new opportunities for research into using large music collections for discovering trends and patterns in music. Systems for trend spotting in online music sales are in commercial development [6], as are systems to support musicological research into music evolution over the corpus of Western classical music scores and available classical recordings [7], [8]. Musicology research aims to answer questions such as: which musical works and performances have been historically the most influential?

Strategies for enabling access to music collections, both new and historical, need to be developed in order to keep up with expectations of search and browse functionality. These strategies are collectively called music information retrieval (MIR) and have been the subject of intensive research by an ever-increasing community of academic and industrial research laboratories, archives, and libraries. There are three main audiences that are identified as the beneficiaries of MIR: industry bodies engaged in recording, aggregating and disseminating music; end users who want to find music and use it in a personalized way; and professionals: music performers, teachers, musicologists, copyright lawyers, and music producers.

At present, the most common method of accessing music is through textual metadata. Metadata can be rich and expressive so there are many scenarios where this approach is sufficient. Most music download services currently use metadata-only approaches and have reached a degree of success with them. However, when catalogues become very large (greater than a hundred thousand tracks) it is extremely difficult to maintain consistent expressive metadata descriptions because many people created the descriptions and variation in concept encodings impacts search performance. Furthermore, the descriptions represent opinions, so editorial supervision of the metadata is paramount [9].

An example of a commercial metadata-driven music system is *pandora.com* where the user is presented with the instruction "type in the name of your favorite artist or song and we'll create a radio station featuring that music and more like it." The system uses metadata to estimate *artist similarity* and *track similarity*; then, it retrieves tracks that the user might want to hear in a personalized radio station. Whereas the query was simple to pose, finding the answer is costly. The system works using detailed human-entered track-level metadata enumerating musical-cultural properties for each of several hundred thousand tracks. It is estimated that it takes about 20–30 minutes per track of one expert's time to enter the metadata.[1] The cost is therefore enormous in the time taken to prepare a database to contain all the information necessary to perform similarity-based search. In this case, it would take approximately 50 person-years to enter the metadata for one million tracks.

Social media web services address the limitations of centralized metadata by opening the task of describing content to public communities of users and leveraging the power of groups to exchange information about content. This is the hallmark of Web 2.0. With millions of users of portals such as *MySpace*, *Flickr*, and *YouTube*, group behavior means that users naturally gravitate towards those parts of the portal—categories or groups—with which they share an affinity; so they are likely to find items of interest indexed by users with similar tastes. However, the activity on social networking portals is not uniform across the interests of society and culture at large, being predominantly occupied by technologically sophisticated users, therefore social media is essentially a type of editorial metadata process.

In addition to metadata-based systems, information about the content of music can be used to help users find music. Content-based music description identifies what the user is seeking even when he does not know specifically what he is looking for. For example, the *Shazam* system (*shazam.com*), described in [10], can identify a particular recording from a sample taken on a mobile phone in a dance club or crowded bar and deliver the artist, album, and track title along with nearby locations to purchase the recording or a link for direct online purchasing and downloading. Users with a melody but no other information can turn to the online music service *Nayio* (*nayio.com*) which allows one to sing a query and attempts to identify the work.

In the recording industry, companies have used systems based on symbolic information about musical content, such as melody, chords, rhythm, and lyrics, to analyze the potential impact of a work in the marketplace. Services such as Polyphonic HMI's *Hit Song Science* and *Platinum Blue Music Intelligence* use such symbolic information with techniques from Artificial Intelligence to make consultative recommendations about new releases.

Because there are so many different aspects to music information retrieval and different uses for it, we cannot address all of them here. This paper addresses some of the recent developments in content-based analysis and retrieval of music, paying particular attention to the methods by which important information about music signals and symbols can be automatically extracted and processed for use with music information retrieval systems. We consider both audio recordings and musical scores, as it is beneficial to look at both, when they are available, to find clues about what a user is looking for.

The structure of this paper is as follows: Section II introduces the types of tasks, methods, and approaches to evaluation of content-based MIR systems; Section III presents methods for high-level audio music analysis; Section IV discusses audio similarity-based retrieval; symbolic music analysis and retrieval are presented in

[1]http://www.pandora.com/corporate.

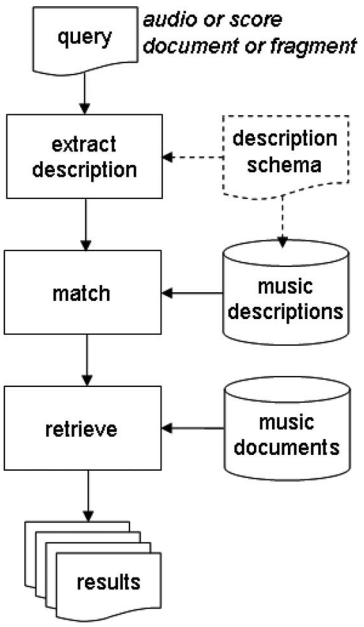

Fig. 1. *Flowchart of canonical content-based query system.*

Section V; Section VI presents an overview of advances in music visualization and browsing; Section VII discusses systems for music information retrieval; and we conclude in Section VIII with a discussion of challenges and future directions for the field.

II. USES CASES AND APPROACHES

A. Use Cases: Specificities and Query Types

Content-based MIR is engaged in intelligent, automated processing of music. The goal is to make music, or information about music, easier to find. To support this goal, most MIR research has been focused on automatic music description and evaluation of the proposed methods. The field is organized around use cases which define a type of query, the sense of match, and the form of the output. Queries and output can be textual information (metadata), music fragments, recordings, scores, or music features. The match can be exact, retrieving music with specific content, or approximate, retrieving near neighbors in a musical space where proximity encodes musical similarity, for example.

The main components of an MIR system are detailed in Fig. 1. These are query formation, description extraction, matching and, finally, music document retrieval. The scope of an MIR system can be situated on a scale of specificity for which the query type and choice of exact or approximate matching define the characteristic specificity of the system. Those systems that identify exact content of individual recordings, for example, are called high-specificity systems; those employing broad descriptions of music, such as genre, have low specificity: that is, a search given a query track will return tracks having little content directly in common with the query, but with some global characteristics that match. Hence, we divide specificity into three broad categories: high-specificity systems match instances of audio signal content; mid-specificity systems match high-level music features, such as melody, but do not match audio content; and low-specificity systems match global (statistical) properties of the query. Table 1 enumerates some of the MIR use cases and their specificities (high, mid, or low). A more comprehensive list of tasks and their specificities is given in [11].

There are three basic strategies for solving MIR use cases. Each strategy is suited to a given specificity. The first is based on conceptual metadata, information that is encoded and searched like text and is suited to low-specificity queries; the second approach uses high-level descriptions of music content corresponding with intuitive or expert knowledge about how a piece of music is constructed. This approach is suited to mid specificity queries. The third strategy is based on low-level signal-based properties which are used for all specificities. We outline each of these three approaches in the following sections.

Table 1 Examples of MIR Tasks and Their Specificities

Use Case	Specificity	Description
Music Identification	H	Identify a compact disk, provide metadata about an unknown track, mobile music information retrieval: e.g. *shazam.com*
Plagiarism detection	H	Identify mis-attribution of musical performances, mis-appropriation of music intellectual property.
Copyright monitoring	H	Monitor music broadcast for copyright infringement or royalty collection
Versions	H/M	Remixes, live vs. studio recordings, cover songs. Used for database normalization and near-duplicate results elimination
Melody	H/M	Find works containing a melodic fragment
Identical Work / Title	M	Retrieve performances of same opus number or song title
Performer	M	Find music by a specific artist
Sounds like	M	Find music that sounds like a given recording
Performance Alignment	M	Mapping one performance onto another independent of tempo and repetition structure
Composer	M	Find works by one composer
Recommend-ation	M/L	Find music that matches the user's personal profile
Mood	L	Find music using emotional concepts: *Joy, Energetic, Melancholy, Relaxing*
Style / Genre	L	Find music that belongs to a generic category: *Jazz, Funk, Female Vocal*
Instrument(s)	L	Find works with same instrumentation
Music-Speech	L	Radio broadcast segmentation, Music archives cataloguing

B. Approaches

1) Metadata: Metadata is the driver of MIR systems. As such, many services exist simply to provide reliable metadata for existing collections of music, either for end users or for large commercial music collections. Most music listeners use metadata in their home listening environments. A common MIR task is to seek metadata from the internet for digital music that was "ripped" from compact disc to a computer. This is the core functionality of automatic Web-based services such as *Gracenote (gracenote.com)* and *MusicBrainz (musicbrainz.org)*—both of these metadata repositories rely on user-contributed content to scale track-level music description to millions of entries.

These services provide both *factual metadata*, namely objective truths about a track, and *cultural metadata*, which contains subjective concepts. For a metadata system to work its descriptions of music must be accurate and the meaning of the metadata vocabulary widely understood. Web 2.0 provides a partial solution in that communities of users can vote on a track's metadata. This democratic process at least ensures that the metadata for a track is consistent with the usage of one, or more, communities. Problems associated with factual information, *artist*, *album*, *year of publication*, *track title*, and *duration*, can severely limit the utility of metadata. Ensuring the generality of the associated text fields, for example, consistencies of spelling, capitalization, international characters, special characters and order of proper names, is essential to useful functioning [9].

In addition to factual metadata, subjective, culturally determined information at the level of the whole track is often used to retrieve tracks. Common classes of such metadata are *mood*, *emotion*, *genre*, *style*, and so forth. Most current music services use a combination of factual and cultural metadata. There has also been much interest in automatic methods for assigning cultural, and factual, metadata to music. Some services collect user preference data, such as the number of times particular tracks have been played, and use the information to make new music recommendations to users based on the user community. For example, Whitman and Rifkin [12] used music descriptions generated from *community metadata*; they achieved Internet-wide description by using data mining and information retrieval techniques. Their extracted data was *time aware*—reflecting changes both in the artists' style and in the public's perception of the artists. The data was collected weekly, and language analysis was performed to associate noun and verb phrases with musical features extracted from audio of each artist.

The textual approach to MIR is a very promising new direction in the field: a comprehensive review of the methods and results of such research is beyond the scope of the current paper.

For all its utility, metadata cannot solve the entirety of MIR due to the complexities outlined above. Commercial systems currently rely heavily on metadata but are not able to easily provide their users with search capabilities for finding music they do not already know about, or do not know *how* to search for. This gap is one of the opportunities for content-based methods, which hold the promise of being able to complement metadata-based methods and give users access to new music via processes of self-directed discovery and musical search that scales to the totality of available music tracks. For the remainder of this paper, we focus primarily on content-based music description rather than factual or culturally determined parameters. However, content-based methods are considered not replacements but enhancements for metadata-based methods.

2) High-Level Music Content Description: An intuitive starting point for content-based music information retrieval is to use musical concepts such as melody or harmony to describe the content of the music. In the early days of MIR, many query-by-humming systems were proposed that sought to extract melodic content from polyphonic audio signals (those with multiple simultaneous musical lines) so that a user could search for music by singing or humming part of the melody; such systems are now being deployed as commercial services; see, for example, *naiyo.com*. A survey of sung-query methods was conducted by Hu and Dannenberg in [13].

High-level intuitive information about music embodies the types of knowledge that a sophisticated listener would have about a piece of music, whether or not they *know* they have that knowledge:

> "It is melody that enables us to distinguish one work from another. It is melody that human beings are innately able to reproduce by singing, humming, and whistling. It is melody that makes music memorable: we are likely to recall a tune long after we have forgotten its text." [14]

Even though it is an intuitive approach, melody extraction from polyphonic recordings, i.e., multiple instruments playing different lines simultaneously, remains extremely difficult to achieve. Surprisingly, it is not only difficult to extract melody from audio but also from symbolic representations such as MIDI files. The same is true of many other high-level music concepts such as rhythm, timbre, and harmony. Therefore, extraction of high-level music content descriptions is a subgoal of MIR and the subject of intensive research. Common high-level descriptors are identified in Table 2. The goal of such tasks is to encode music into a schema that conforms to traditional Western music concepts that can then be used to make queries and search music.

Automatic extraction of factual, cultural, and high-level music descriptions have been a subject of intense study in the MIREX music information retrieval experimental

Table 2 High-Level Music Features (Hard to Extract)

High-level Description	Data Source	Task Description
Timbre	Audio	Instrument Recognition Percussive, Pitched, Ensemble Recognition
Melody / Bass	Audio / Symbolic	Melody-line extraction Bass-line extraction
Rhythm	Audio	Onset detection Meter identification Meter alignment (bars) Beat (tactus) tracking Tempo tracking Average tempo
Pitch	Audio	Single fundamental freq. Multiple fundamental freq.
Harmony	Audio / Symbolic	Chord label extraction Bass-line extraction
Key	Audio / Symbolic	Modulation tracking Pitch spelling
Structure	Audio / Symbolic	Verse / chorus extraction Repeat extraction
Lyrics	Audio	Singing detection, lyrics-identification, word recognition
Non-Western music	Audio	Micro-tonal tuning systems Non-Western canon of concepts

exchange. MIREX provides a framework for formal evaluation of MIR systems using centralized tasks, datasets, platforms, and evaluation methods [15], [16]. As such, MIREX has become a very important indicator of the state of the art for many subtasks within the field of MIR. A summary of the results of the 2007 high-level music tasks is given in Table 3; the query-by-humming task has a particularly high score. It is interesting to note that the best-performing system on this task used low-level audio matching methods rather than extracting a high-level melody feature from audio [18]. The authors suggest that low-level audio methods outperform symbolic methods even when clean symbolic information is available as in this task.

Table 3 Summary of Results of Best-Performing Classification and Recognition Systems in MIREX 2007 Exchange

MIREX 2007 Task	Evaluation Metric	Best Result
High-Level Tasks		
Mood Recognition	Accuracy	61.5%
Classical Composer Recognition	Accuracy	53.72%
Cover Song (Work) Recognition	Accuracy	52%
Artist Recognition	Accuracy	48.14%
Low-Level Tasks		
Polyphonic Pitch Tracking (Melody / Bass) (see *Section III*)	F-measure	0.614
Onset (see *Section III*)	F-measure	0.81
Similarity Tasks		
Query by Humming	Mean reciprocal rank	0.92
Audio (Track) Similarity (see *Section IV*)	Avg. fine score (0-1)	0.56
Melody (Symbolic) Similarity (see *Section V*)	Avg. fine score (0-1)	0.59

Because there is a great number of music recordings available that can be used as a first stage input to a high-level music description system, this motivates work on extracting high-level music features from low-level audio content. The MIREX community extends the range of tasks that are evaluated each year, allowing for valuable knowledge to be gained on the limits of current algorithms and techniques.

3) Low-Level Audio Features: The third strategy for content-based music description is to use the information in the digital audio. Low-level audio features are measurements of audio signals that contain information about a musical work and music performance. They also contain extraneous information due to the difficulty of precisely measuring just a single aspect of music, so there is a tradeoff between the signal-level description and the high-level music concept that is encoded.

In general, low-level audio features are segmented in three different ways: frame based segmentations (periodic sampling at 10 ms-1000 ms intervals), beat-synchronous segmentations (features are aligned to musical beat boundaries), and statistical measures that construct probability distributions out of features (bag of features models). Many low-level audio features are based on the short-time spectrum of the audio signal. Fig. 2 illustrates how some of the most widely used low-level audio features are extracted from a digital audio music signal using a windowed fast Fourier transform (FFT) as the spectral extraction step. Both frame-based and beat-based windows are evident in the figure.

a) Short-Time Magnitude Spectrum: Many low-level audio features use the magnitude spectrum as a first step for feature extraction because the phase of the spectrum is not as perceptually salient for music as the magnitude. This is generally true except in the detection of onsets and in phase continuation for sinusoidal components.

b) Constant-Q/Mel Spectrum: The ear's response to an acoustic signal is logarithmic in frequency and uses nonuniform frequency bands, known as *critical bands*, to resolve close frequencies into a single band of a given center frequency. Many systems represent the constant bandwidth critical bands using a constant-Q transform, where the Q is the ratio of bandwidth to frequency [17], [19]. It is typical to use some division of the musical octave for the frequency bands, such as a twelfth, corresponding to one semitone in Western music, but it is also common to use more perceptually motivated frequency spacing for band centers. Fig. 3 shows the alignment between a set of linearly spaced frequency band edges and the corresponding logarithmically spaced twelfth-octave bands. The

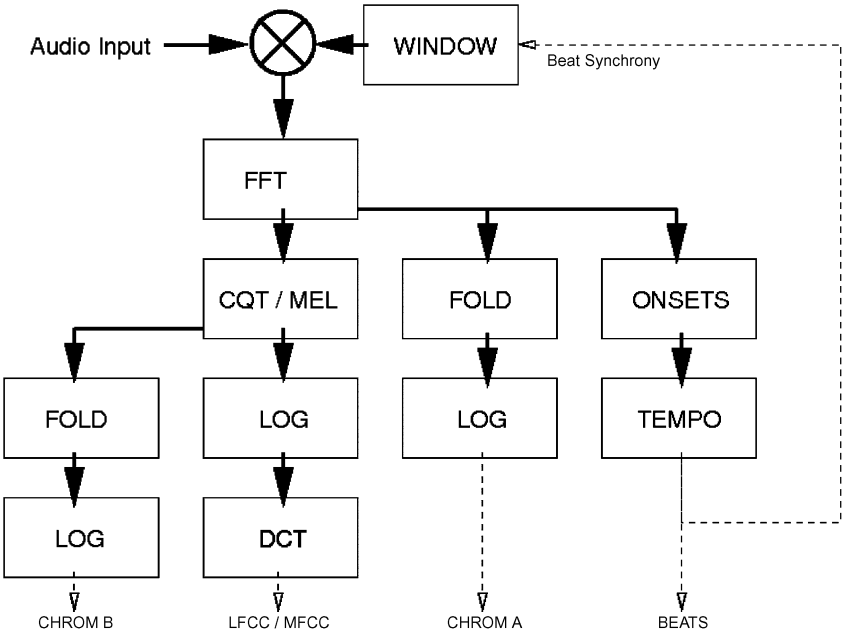

Fig. 2. *Schematic of common audio low-level feature extraction processes. From left-to-right: log-frequency chromagram, Mel-frequency cepstral coefficients, linear-frequency chromagram, and beat tracking. In some cases, the beat tracking process is used to make the features beat synchronous, otherwise segmentation uses fixed-length windows.*

Mel frequency scale has linearly spaced filters in the lower frequency range and logarithmically spaced filters above 1300 Hz. Both logarithmic and Mel frequency scales are used. The Mel or Constant-Q spectrum can be obtained from a linear spectrum by summing the powers in adjacent frequency bands. This approach has the advantage of being able to employ the efficient FFT to compute the spectrum.

c) Pitch-Class Profile (Chromagram): Another common type of frequency folding is used to represent the energy due to each pitch class in twelfth-octave bands called a pitch-class profile (PCP) [20]–[22]. This feature integrates the energy in all octaves of one pitch class into a single band. There are 12 equally spaced pitch classes in Western tonal music, independent of pitch height, so there are typically 12 bands in a chromagram representation. Sometimes, for finer resolution of pitch information, the octave is divided into an integer multiple of 12 such as 24, 36, or 48 bands. Tuning systems that use equally spaced pitch classes are called *equal temperament*. Recently, some studies have explored extracting features for tuning systems that do not use equally spaced pitch classes: a necessary extension for application to non-Western music.

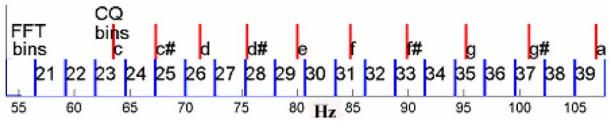

Fig. 3. *Folding of a set of linearly spaced frequency bands (lower graph) onto a set of logarithmically spaced frequency bands (upper graph). The x-axis shows frequency of bands, upper lines are labeled by their Western music notation pitch class and lower lines by their FFT bin number (for 16 384 bins with 44.1 kHz sample rate).*

d) Onset Detection: Musical events are delineated by onsets; a note has an attack followed by sustain and decay portions. Notes that occur simultaneously in music are often actually scattered in time and the percept is integrated by the ear-brain system. Onset detection is concerned with marking just the beginnings of notes. There are several approaches to onset detection, employing spectral differences in the magnitude spectrum of adjacent time points, or phase differences in adjacent time points, or some combination of the two (complex number onset detection) [23]–[25]. Onset detection is one of the tasks studied in the MIREX framework of MIR evaluation shown in Table 3.

e) Mel/Log-Frequency Cepstral Coefficients: Mel-frequency cepstral coefficients (MFCC) take the logarithm of the Mel magnitude spectrum and decorrelate the resulting values using a Discrete Cosine Transform. This is a real-valued implementation of the complex cepstrum in signal processing [19]. The effect of MFCCs is to organize sinusoidal modulation of spectral magnitudes by increasing modulation frequency in a real-valued array. Values at the start of the array correspond to long wave spectral modulation and therefore represent the projection of the log magnitude spectrum onto a basis of formant

peaks. Values at the end of the MFCC array are the projection coefficients of the log magnitude spectrum onto short wavelength spectral modulation, corresponding to harmonic components in the spectrum. It is usual to use about 20 MFCC coefficients, therefore representing the formant peaks of a spectrum; this extraction corresponds, in part, to musical *timbre*—the way the audio sounds other than its pitch and rhythm.

f) Spectral Flux: The spectral flux of a musical signal estimates the fine spectral-temporal structure in different frequency bands by measuring the modulation amplitudes in mid-to-high spectral bands [26], [27]. The resulting feature is a two-dimensional matrix, with frequency bands in the rows and modulation frequency in columns, representing the rate of change of power in each spectral band.

g) Decibel Scale (Log Power): The decibel scale is employed for representing power in spectral bands because the scale closely represents the ear's response. The decibel scale is calculated as ten times the base-10 logarithm of power.

h) Tempo/Beat/Meter Tracking: As shown in Fig. 2, beat extraction follows from onset detection, and it is often used to align the other low-level features. Alignment of features provides a measurement for every beat interval, rather than at the frame level, so the low-level features are segmented by musically salient content. This has recently proven to be exceptionally useful for mid-specificity MIR tasks such as cover songs and versions identification. Low-level audio features in themselves cannot tell us much about music; they encode information at too fine a temporal scale to represent perceptually salient information. It is usual in MIR research to collect audio frames into one of several aggregate representations. Table 4 describes this second-stage processing of low-level audio features which encodes more information than individual audio frames. An aggregate feature is ready for similarity measurement whereas individual low-level audio feature frames are not. The chosen time scale for aggregate features depends on the specificity and temporal acuity of the task.

These low-level audio features and their aggregate representations are used as the first stage in bottom-up processing strategies. The task is often to obtain a high-level representation of music as a next step in the processing of music content. The following section gives a summary of some of the approaches to bridging the gap between low-level and high-level music tasks such as these.

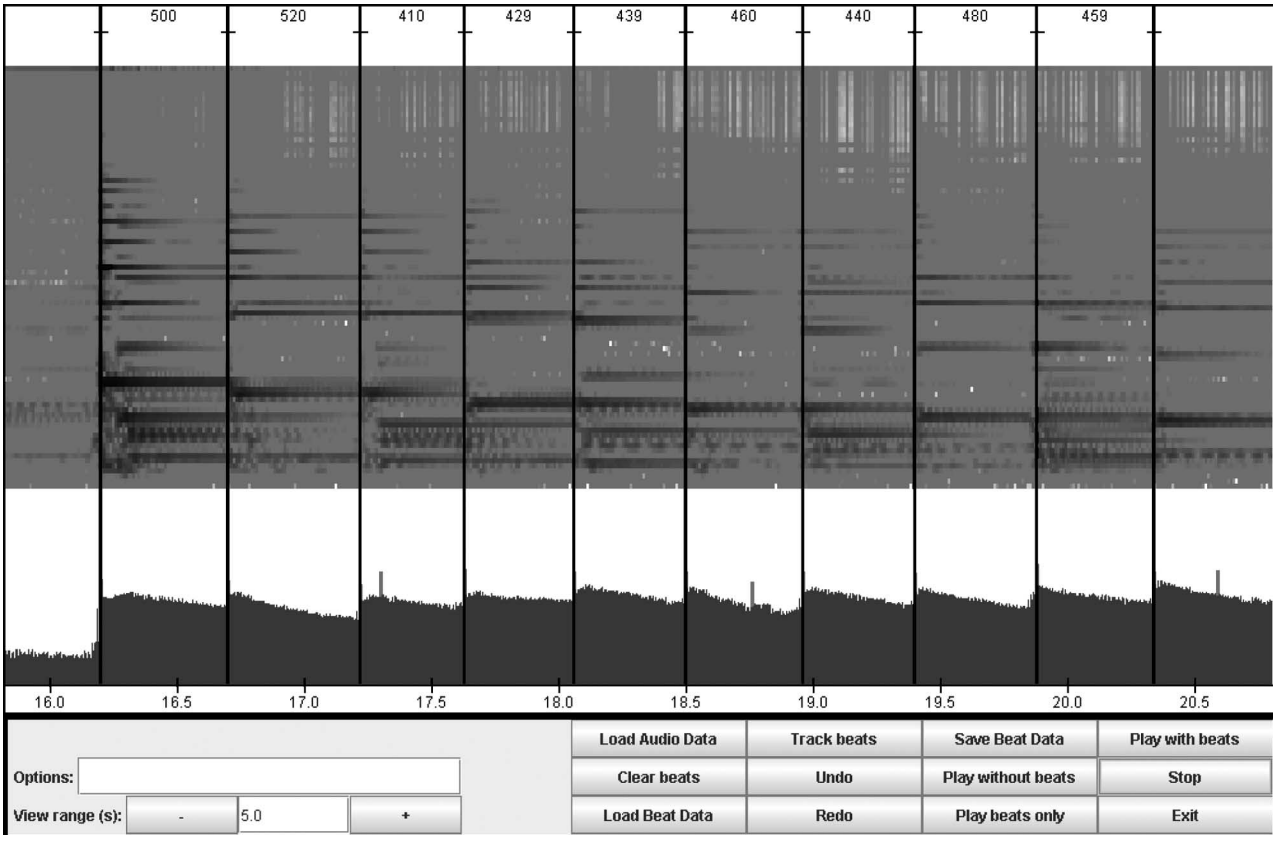

Fig. 4. ***Beat tracking a music file with beatroot tool. Lower portion of graph shows audio power and upper portion shows spectrogram. Dark vertical lines are beat positions. Numbers running along top of graph are inter-beat intervals in milliseconds.***

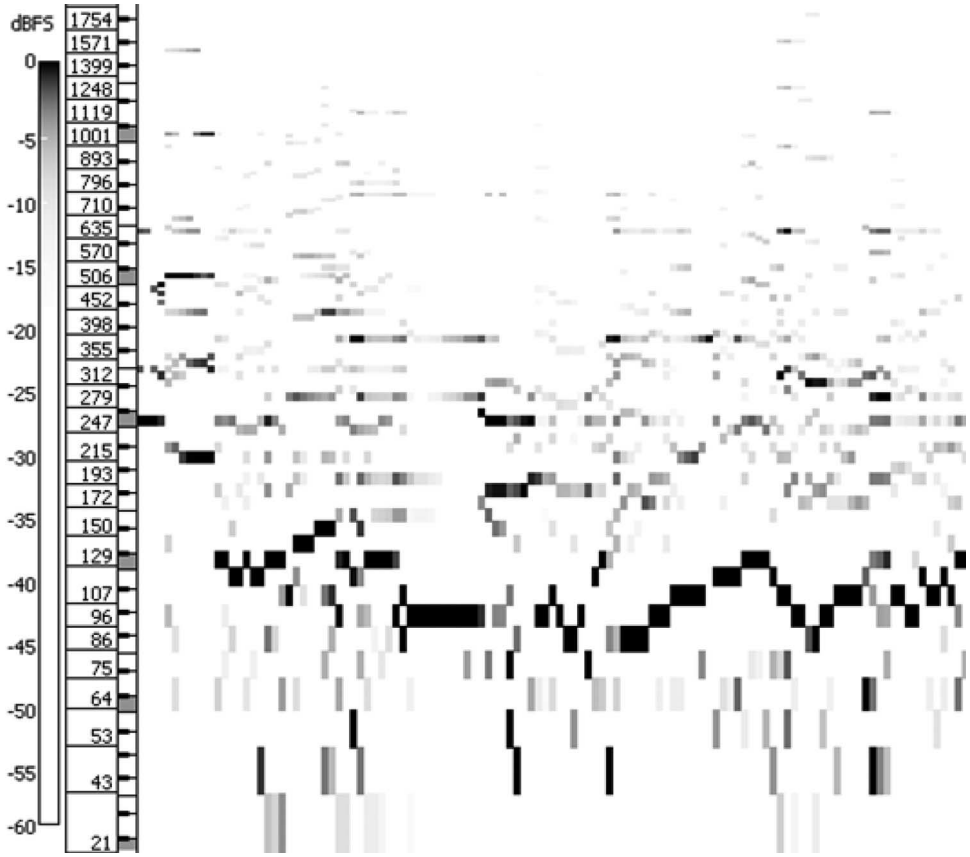

Fig. 5. *Predominant f0 trajectory of Charlie Haden jazz track shown in SonicVisualiser. Darker regions are maxima of this time versus logarithmic frequency spectrogram. Vertical axis is laid out as a piano keyboard for musical pitch reference.*

III. AUDIO ANALYSIS

In contrast to speech recognition and text IR systems, most music has several streams of related information occurring in parallel. As such, music is organized both horizontally (in time), and vertically (in frequency). Furthermore, the information in music is constructed with hierarchical schemas. Systems for analyzing and searching music content must seek to represent many of these viewpoints simultaneously to be effective.

As discussed in Section II-B, one of the intermediate goals of MIR is to extract high-level music content descriptions from low-level audio processes. The following sections describe research into extracting such high-level descriptions with a view to transforming musical audio content into representations that are intuitive for humans to manipulate and search. We begin the discussion with the related high-level music description problems of beat tracking, tempo estimation, and meter tracking.

Table 4 Frame Aggregation Methods for Low-Level Audio Features

Aggregate Type	Description	Extraction method
State models	Fine Spectral-Temporal Structure	Hidden Markov model (HMM) obtained by training on a corpus of feature data. The HMM maps d-dimensional features to 1-dimensional states. It can segment features into homogeneous regions.
Bag-of-Frames Models	Global track statistics	Single Gaussian or Gaussian mixture models of collection of audio features for a track.
Audio Shingles (sequences)	Temporally-ordered sets of feature frames	Audio shingles use, possibly overlapping, sequences of features to represent coarse temporal context.

A. Beat Tracking

Automatic estimation of the temporal structure of music, such as musical beat, tempo, rhythm, and meter, is not only essential for the computational modeling of music understanding but also useful for MIR. Temporal properties estimated from a musical piece can be used for content-based querying and retrieval, automatic classification, music recommendation, and playlist generation. If the tempo of a musical piece can be estimated, for example,

it is easy to find musical pieces having a similar tempo without using any metadata. Once the musical beats are estimated, we can use them as the temporal unit for high-level beat-based computation instead of low-level frame-based computation. This facilitates the estimation of other musical descriptions, such as the music structure and chorus sections [22]. Since the estimated beats can also be used for normalizing the temporal axis of musical pieces, the beat-based time alignment facilitates time-scale invariant music identification or cover song identification [28], [29].

Here, we define beat tracking (including measure or bar-line estimation) as the process of organizing musical audio signals into a hierarchical beat structure [30]. The typical beat structure comprises the quarter-note level (the tactus level represented as almost regularly spaced beat times) and the measure level (bar lines). The basic nature of tracking the quarter-note level is represented by two parameters.

Period: The period is the temporal difference between the times of two successive beats. The tempo (beats per minute) is inversely proportional to the period.

Phase: The phase corresponds to actual beat positions and equals zero at beat times.

On the other hand, the measure level is defined on beat times because the beat structure is hierarchical: the beginnings of measures (bar-line positions) coincide with beat times. The difficulty of beat tracking depends on how explicitly the beat structure is expressed in the target music: it depends on temporal properties such as tempo changes and deviations, rhythmic complexity, and the presence of drum sounds.

1) Tracking Musical Beats (Quarter-Note Level): The basic approach of estimating the period and phase of the quarter-note (tactus) level is to detect onset times and use them as cues. Many methods assume that a frequently occurring inter-onset interval (IOI), the temporal difference between two onset times, is likely to be the beat period and that onset times tend to coincide with beat times (i.e., sounds are likely to occur on beats).

To estimate the beat period, a simple technique is to calculate the histogram of IOIs between two adjacent onset times or cluster the IOIs and pick out the maximum peak or the top ranked cluster within an appropriate tempo range. This does not necessarily correspond to the beat period, though. A more sophisticated technique is to calculate a windowed autocorrelation function of an onset-time sequence, power envelope, or spectral flux of the input signal, or continuous onset representation with peaks at onset positions, and pick out peaks in the result. This can be considered an extended version of the IOI histogram because it naturally takes into account various temporal differences such as those between adjacent, alternate, and every third onset times. Another sophisticated technique is to apply a set of comb-filter resonators, each tuned to a possible period, to the time-varying degree of musical accentuation [31].

For audio-based beat tracking, it is essential to split the full frequency band of the input audio signal into several frequency subbands and calculate periodicities in each subband. Goto [30] proposed a method where the beat-period analysis is first performed within seven logarithmically equally spaced subbands and those results are then combined across the subbands by using a weighted sum. Scheirer [31] also used the idea of this subband-based beat tracking and applied a set of comb-filter resonators to the degrees of musical accentuation of six subbands to find the most resonant period. To locate periodicity in subband signals, Sethares and Staley [32] used a periodicity transform instead of using comb-filter resonators.

After estimating the beat period, the phase should be estimated. When onset times are used to estimate the period, a windowed cross-correlation function is applied between an onset-time sequence and a tentative beat-time sequence whose interval is the estimated period. The result can be used to predict the next beat in a real-time beat-tracking system. On the other hand, when the degrees of musical accentuation are used, the internal state of the delays of comb-filter resonators that have lattices of delay-and-hold stages can be used to determine the phase [31]. To estimate the period and phase simultaneously, there are other approaches using adaptive oscillators [33].

2) Dealing With Ambiguity: The intrinsic reason that beat tracking is difficult is due to the problem of inferring an original beat structure that is not expressed explicitly. This causes various ambiguous situations, such as those where different periods seem plausible and where several onset times obtained by frequency analysis may correspond to a beat.

There are variations in how ambiguous situations in determining the beat structure are managed. A traditional approach is to maintain multiple hypotheses, each having a different possible set of period and phase. A beam search technique or multiple-agent architectures [30], [34] have been proposed to maintain hypotheses. A more advanced, computationally intensive approach for examining multiple hypotheses is to use probabilistic generative models and estimate their parameters. Probabilistic approaches with maximum likelihood estimation, MAP estimation, and Bayes estimation could maintain distributions of all parameters, such as the beat period and phase, and find the best hypothesis as if all possible pairs of the period and phase were evaluated simultaneously. For example, Hainsworth and Macleod [35] explored the use of particle filters for audio-based beat tracking on the basis of MIDI-based methods by Cemgil and Kappen [36]. They made use of Markov chain Monte Carlo (MCMC) algorithms and sequential Monte Carlo algorithms (particle filters) to estimate model parameters, such as the period and phase.

3) Estimating Beginnings of Measures (Measure Level): Higher level processing using musical knowledge is necessary to determine the measure level of the hierarchical beat structure. Musical knowledge is also useful for selecting the best hypothesis in the above ambiguous situations.

Various kinds of musical knowledge have been studied. Goto [30] used musical knowledge concerning chord changes and drum patterns with a focus on popular music. By estimating the chord changes using signal processing technique, for example, the beginnings of measures can be determined so that chords are more likely to change on those positions. On the other hand, Klapuri *et al.* [38] used musical knowledge concerning temporal relationship among different levels of the hierarchical beat structure and encoded this prior knowledge in HMMs that could jointly estimate periods at different hierarchical levels and then separately estimate their phases.

4) Conclusion: The topic of beat tracking still attracts many researchers because it includes fundamental issues in understanding temporal aspects of music, contributes to a number of practical applications, and it is difficult to achieve perfect beat tracking for various kinds of music. Therefore, new approaches are proposed every year, including holistic beat tracking [39] where information about music structure estimated before beat tracking helps to track beats by adding a constraint that similar segments of music should have corresponding beat structure.

B. Melody and Bass Estimation

Automatic estimation of melody and bass lines is important because the melody forms the core of Western music and is a strong indicator for the identity of a musical piece, see Section II-B, while the bass is closely related to the harmony. These lines are fundamental to the perception of music and useful in MIR applications. For example, the estimated melody (vocal) line facilitates song retrieval based on similar singing voice timbres [41], music retrieval/classification based on melodic similarities, and music indexing for query by humming which enables a user to retrieve a musical piece by humming or singing its melody. Moreover, for songs with vocal melody, once the singing voice is extracted from polyphonic sound mixtures, the lyrics can be automatically synchronized with the singing voice by using a speech alignment technique and can be displayed with the phrase currently being sung highlighted during song playback, like the Karaoke display [42].

The difficulty of estimating melody and bass lines depends on the number of channels: the estimation for stereo audio signals is easier than the estimation for monaural audio signals because the sounds of those lines tend to be panned to the center of stereo recordings and the localization information can help the estimation. In general, most methods deal with monaural audio signals because stereo signals can be easily converted to monaural signals. While a method depending on stereo information cannot be applied to monaural signals, a method assuming monaural signals can be applied to stereo signals and can be considered essential to music understanding, since human listeners have no difficulty understanding melody and bass lines even from monaural signals.

Here, melody and bass lines are represented as a continuous temporal-trajectory representation of fundamental frequency (F0, perceived as pitch) or a series of musical notes. It is difficult to estimate the F0 of melody and bass lines in monaural polyphonic sound mixtures containing simultaneous sounds of various instruments, because in the time-frequency domain the frequency components of one sound often overlap the frequency components of simultaneous sounds. Even state-of-the-art technologies cannot fully separate sound sources and transcribe musical scores from complex polyphonic mixtures. Most melody and bass estimation methods therefore do not rely on separated sounds or transcribed scores but directly estimate the target melody and bass lines from music that has distinct melody and bass lines, such as popular songs.

1) Estimating Melody and Bass Lines by Finding the Predominant F0 Trajectory: Since the melody line tends to have the most predominant harmonic structure in middle- and high-frequency regions and the bass line tends to have the most predominant harmonic structure in a low-frequency region, the first classic idea of estimating melody and bass lines is to find the most predominant F0 in sound mixtures with appropriate frequency-range limitation. In 1999, Goto [43], [44] proposed a real-time method called PreFEst (Predominant-F0 Estimation method) which detects the melody and bass lines in monaural sound mixtures. Unlike most previous F0 estimation methods, PreFEst does not assume the number of sound sources.

PreFEst basically estimates the F0 of the most predominant harmonic structure—the most predominant F0 corresponding to the melody or bass line—within an intentionally limited frequency range of the input sound mixture. It simultaneously takes into consideration all possibilities for the F0 and treats the input mixture as if it contains all possible harmonic structures with different weights (amplitudes). It regards a probability density function (PDF) of the input frequency components as a weighted mixture of harmonic-structure tone models (represented by PDFs) of all possible F0s and simultaneously estimates both their weights corresponding to the relative dominance of every possible harmonic structure and the shape of the tone models by maximum *a posteriori* probability (MAP) estimation considering their prior distribution. It then considers the maximum-weight model as the most predominant harmonic structure and obtains its F0. The method also considers the F0's temporal continuity by using a multiple-agent architecture.

Because this original PreFEst simply estimates the predominant F0 trajectory every moment and does not distinguish between sound sources, Fujihara *et al.* [45] extended it to discriminate between vocal and nonvocal sounds with the focus on songs with vocal melody. This method evaluates the "vocal" probabilities of the harmonic structure of each F0 candidate by using two Gaussian mixture models (GMMs) for vocal and nonvocal. This extension resulted in improvement of the estimation accuracy. In addition, the original PreFEst inevitably detects the F0 of a dominant accompaniment part in the absence of a melody line. Fujihara *et al.* [45] extended it to identify vocal sections where the vocal melody line is actually present by using a two-state Hidden Markov model (HMM) with vocal and nonvocal states.

Marolt [46] also used another implementation of the PreFEst with some modifications to represent the melody line as a set of short vocal fragments of F0 trajectories. The advantage of this method is that F0 candidates are tracked and grouped into melodic fragments, which are then clustered into the melody line. The melodic fragments denote reasonably segmented signal regions that exhibit strong and stable F0 and are formed by tracking temporal trajectories of the F0 candidates.

Paiva *et al.* [47] proposed a method of obtaining MIDI-level note sequence of the melody line, while the output of PreFEst is a simple temporal trajectory of the F0. The basic idea is the same as the PreFEst concept that the F0 of the most predominant harmonic structure is considered the melody. The method first estimates predominant F0 candidates by using correlograms, quantizes their frequencies to the closest MIDI note numbers, and forms temporal trajectories of the F0 candidates. Then, the trajectories are segmented into MIDI-level note candidates by finding a sufficiently long trajectory having the same note number and by dividing it at clear local minima of its amplitude envelope. After eliminating inappropriate notes, the melody note sequence is finally obtained by selecting the most predominant notes according to heuristic rules.

Li and Wang [48] proposed a method of detecting the vocal melody line. It first uses a 128-channel gammatone filter bank and splits these channels into two subbands at 800 Hz. To extract periodicity information, an autocorrelation is calculated on the filter output of each channel in the low subband while this is done on the output envelope of each channel in the high subband. Plausible peaks are then selected in both autocorrelation results and are used to score a collection of F0 hypotheses. The method finally tracks the most probable temporal trajectory of the predominant F0 in the scored F0 hypotheses.

2) Knowledge-Based or Classification-Based Estimation of Melody Lines: Eggink and Brown [49] proposed a knowledge-based method of detecting the melody line with the emphasis on using various knowledge sources to choose the most likely succession of F0s as the melody line. Unlike other methods, this method is specialized for a classical sonata or concerto, where a solo melody instrument can span the whole pitch range, so the frequency-range limitation is not feasible. In addition, because the solo instrument does not always have the most predominant F0, additional knowledge sources are indispensable. The knowledge sources include local knowledge about an instrument recognition module and temporal knowledge about tone durations and interval transitions. Those sources can both help to choose the correct F0 among multiple concurrent F0 candidates and to determine sections where the solo instrument is actually present.

Poliner and Ellis [50] proposed a classification-based method that uses a set of support vector machine (SVM) classifiers. Each classifier is assigned to a particular F0 quantized to the semitone level (i.e., a different MIDI note number) and is trained on polyphonic sound mixtures with correct melody annotations so that it can judge whether each audio frame has the melody of the assigned F0. The overall melody trajectory is finally smoothed by using an HMM. The advantage of this method is that it does not make any assumptions beyond what is learned from its training data.

3) Conclusion: Although even state-of-the-art technologies for automatic transcription or sound source segregation have had significant difficulty dealing with complex polyphonic sound mixtures containing singing voices and sounds of various instruments (even drums), most above-mentioned state-of-the-art technologies for automatic melody/bass estimation can deal with such music recordings and are useful in many applications. This practical sub- (or reduced) problem (so-called predominant melody/F0 estimation/detection) that was first proposed by Goto [43], therefore, has attracted a lot of researchers since 1999 and has resulted in various approaches. In a recent interesting approach by Ryynanen and Klapuri [51], even a multiple F0 estimation method designed for polyphonic music transcription is used as a front-end feature extractor for this problem; comparisons of various methods are discussed in [56].

C. Chord and Key Recognition

Musical chords and key information are an important part of Western music and this information can be used to understand the structure of music. Currently, the best performing chord- and key-recognition systems use HMMs to unify recognition and smoothing into a single probabilistic framework.

In an HMM, a musical performance is assumed to travel through a sequence of states. These states are hidden from the recognizer because the only thing we can see is the acoustic signal. Thus, an HMM consists of a transition matrix—a probabilistic model of the state sequence—and an output model—a probabilistic distribution that encodes

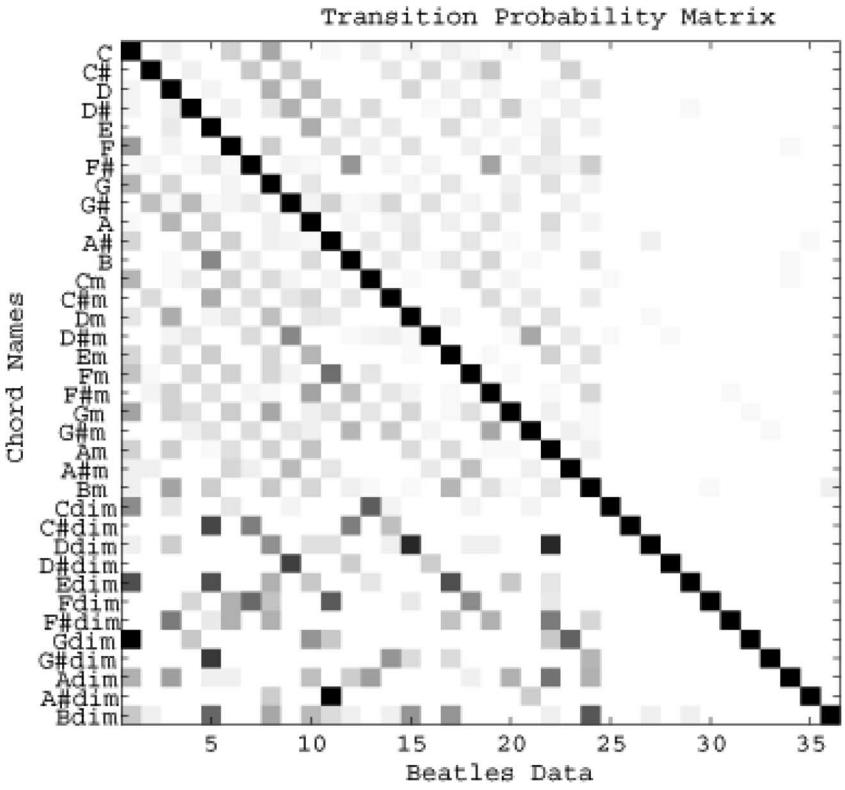

Fig. 6. *A 36 × 36 transition probability matrices obtained from 158 pieces of Beatles' music. For viewing purposes, we show logarithm of original matrix. Axes are labeled in the order of major, minor, and diminished chords. Right third of this matrix is mostly zero because a musical piece is unlikely to transition from a major or minor chord to a diminished chord; bottom left third of matrix shows that once in a diminished chord, music is likely to quickly transition back to a major or minor chord. Reprinted with permission from Lee [54].*

the probability that one of the states produces the signal that we measure. In a chord-recognition system there are, typically, 24 or 36 chords, each represented by one state. The acoustic signal is represented as a set of chromagram frames so the output model represents the probability that each state (chord) produces any given chromagram signal. The beauty of an HMM is that it is relatively easy to evaluate this model and find the sequence of states that best fits the observed data. By training different chord-recognition systems for each key, we can simultaneously recognize a musical key and the best chord sequence.

Sheh and Ellis [52] proposed the first HMM-based chord recognizer, and this idea was refined by Bello and Pickens [53]. Both of these systems were hampered by a shortage of labeled training data, so Bello and Pickens built part of their model using human knowledge. A system by Lee and Slaney [54] used a large database of symbolic musical scores to obtain chord ground truth (via symbolic musical analysis) and acoustic waveforms (via synthesis) and match their performance. About the same time, Harte *et al.* [55] proposed a new 6-D feature vector—called tonal centroid—based on harmonic relationships in western music. This feature proved to have even better performance in Lee's system, and the switch was straightforward because the system is entirely based on training data. This machine-learning based solution is promising because speech recognition has taught us that it is always better to have more data, and learning from symbolic music data is more cost effective.

The results of a chord-recognition model are shown in Figs. 6 and 7. Fig. 6 shows the transition matrix—the probability of moving from any one chord to any other chord on a frame-by-frame basis. Thus, when starting with a C chord (the top row of the figure) one is most likely to move to a F chord in Lee's classical music or a G chord in Beatles music. The resulting musical segmentation is shown in Fig. 7, which shows a sequence of chords over 22 seconds.

D. Music Structure

Segmentation in MIR is both an end in itself and a means to improve performance in some other task; indeed, we have already discussed two kinds of segmentation (beat tracking and chord recognition) in Section III-A, where the segments are chords, and III-C, where the segments extracted are the temporal regions between beats or the duration of a chord. Beat-based segmentation has been found to improve performance in cover song identification [28]

by enabling matching between tracks that is invariant to both tempo and, more relevantly for this section, reordering of musical structure; for example, verse-chorus-verse in one version matches verse-verse-chorus-verse in another. Recently, automatic structure extraction has been used to facilitate editing of audio in recording workflows [57].

Audio similarity, discussed further in Section IV, also benefits from segmenting the tracks beforehand; even a simple segmentation, removing the starts and ends of tracks as in [58], allows a similarity measure based on audio signal analysis to improve in performance by removing unrepresentative material (intro and outro) from consideration; a segmentation of the form outlined below allows this removal of unrepresentative content in a less *ad hoc* manner. This has implications for recommender systems, as in the *SoundBite* system [59], which uses a structural segmentation to generate representative thumbnails for presentation and search.

A segment, in music information, is a region with some internal similarity or consistency. Segments include both regions that are approximately homogeneous in a feature space, such as timbre or instrumentation, and those which have some relationship with other regions either in the same or different musical works. With this definition, a segment implies that it has temporal boundaries at its start and end; however, a pair of boundaries does not necessarily imply that there is a segment between them, as segments in general may overlap or have gaps between them, and there is no one single hierarchical arrangement for all segments.

The scales of interest for musical segments range from the region between successive onsets in a track (of the order of 100 ms) to entire movements (thousands of seconds) within a large, monolithic audio recording, a complete opera for example. There are tasks relating to all timescales between these extremes, such as segmenting a track into beats at the tactus level or musical phrases; performing a structural segmentation of a popular music track into sections which can be identified as the verse, chorus, or bridge; and separating speech from audio or distinct tracks from each other (for example, within a streaming broadcast [60] for royalty distribution or within a large archive such as archive.org for detection of copyright infringement).

1) Structural Analysis: One segmentation task is to extract the high-level structure of a track from just the audio; applications for this include chorus-detection for music sales kiosks [61] or retrieval result presentation [62]. An important component of these tasks is how knowledge of the kind of music being analyzed, and the kind of segments being retrieved, affects the features and algorithms used.

Fig. 8 shows an automatic segmentation using the method detailed in [63] of a pop track, along with an expert's annotation of segments from the same track. Note that the segmentation matches the annotation closely: apart from the identification of the solo ending and the transition, each annotation is algorithmically given a distinct label. However, despite the close match between algorithmic segmentation and human annotation in Fig. 1,

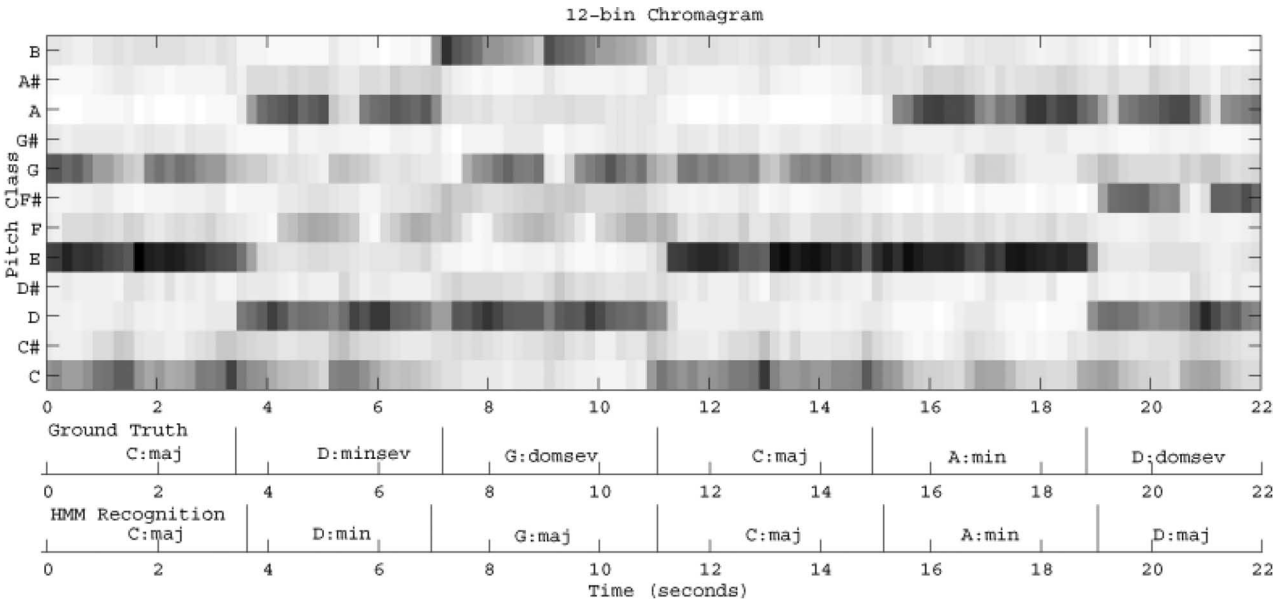

Fig. 7. ***Recognition results for Bach's Prelude in C Major performed by Glenn Gould. Below 12-bin chromagram are ground truth and recognition result using a C major key HMM trained on classical symbolic music. Reprinted with permission from Lee [54].***

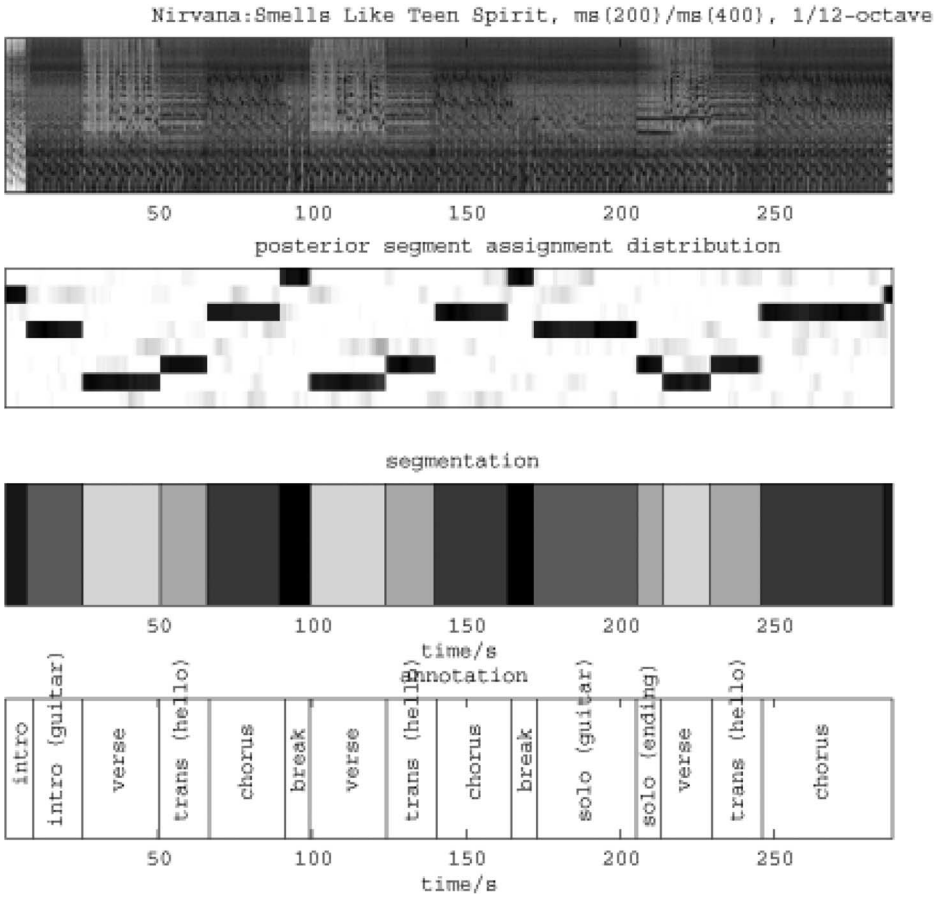

Fig. 8. ***Segmentation of Nirvana's Smells Like Teen Spirit (top panel: spectrogram; segmentation in middle two panels) along with annotation by a human expert (bottom panel).***

care must be taken in evaluating the performance of a segmentation method for segment retrieval.

It is not sufficient to compare the output of a segmentation method against single, unambiguous "ground truth," because many different divisions of music tracks can be considered "correct." Different expert listeners will differ in their judgments, for instance segmenting at different hierarchy levels [64] or with different criteria for segment boundaries. Rather than a direct comparison with a single ground truth, it may be useful to elicit responses to segmentations using a forced-choice [65] or recognition-memory [66] paradigm, or else to ask several experts for quantitative ratings of segmentations in a manner similar to [67].

2) Smoothing: Fig. 9 shows a segmentation using the same algorithm and textural feature type as in Fig. 8. In this case, the machine-generated segmentation does not directly reflect the ground truth structural annotation; indeed, it would be surprising if it did, as sections in classical music are not generally associated with a single texture. Nevertheless, there is a clear correspondence between the segmentation and the annotation, and it is possible to recover the repeated structure labels from this segmentation, even given the expressive performance [68], [69], because the segmentation has acted to smooth over small variations. Again, the subsequent task (for which the segmentation acts as a preprocessing or smoothing stage) will dictate, or at least suggest, a particular segmentation algorithm or set of parameters: for instance, in the task above (detecting repeated structure in classical music) it is important that the segments generated are sufficiently large that the subsequent matching is robust against small variations, while being smaller than the repeated sections being searched for.

3) Application to Content-Based Retrieval: Segmentation additionally allows for control of the characteristics of content-based retrieval applications in two ways: firstly, segmentation of the database allows for indexing to be done over regions of homogeneity, yielding potential space savings over a simple frame-based feature vector with no loss of retrieval performance. Secondly, segmentation of the provided query datum, as well as giving the potential

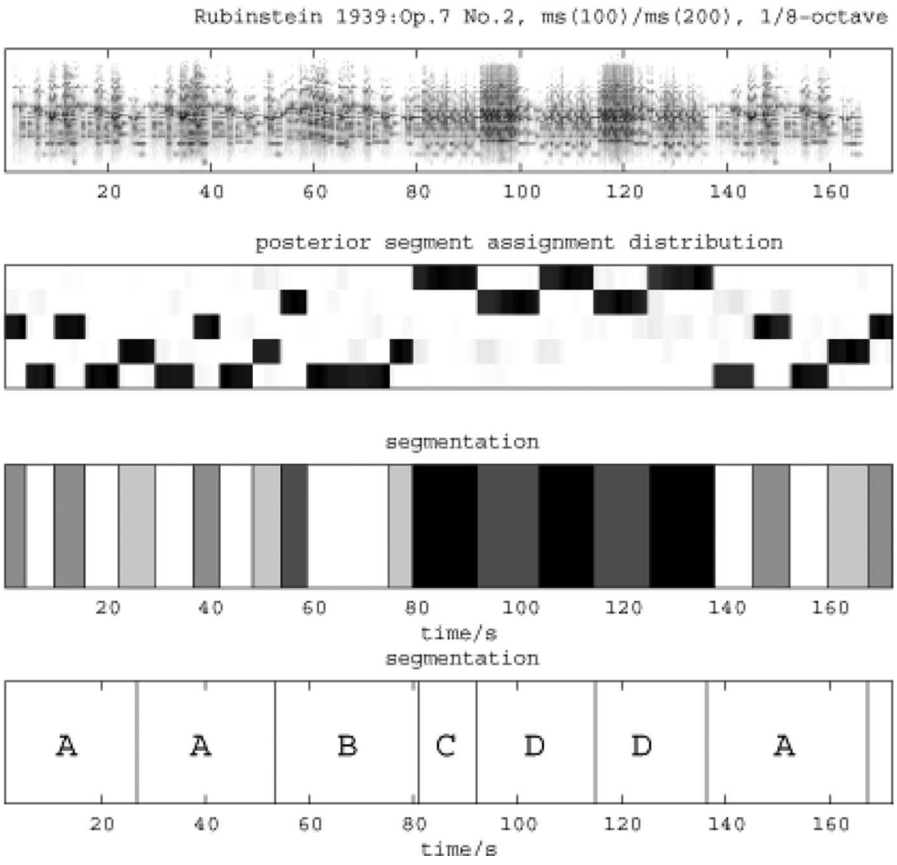

Fig. 9. *Segmentation of Rubenstein's 1939 performance of Chopin's Mazurka Op. 7 No. 2. (top panel: spectrogram; segmentation in middle two panels) along with annotation by a human expert (bottom panel).*

for direct matching, permits implementation of searches of different specificity, matching the sequence of segments directly, or arbitrarily permuted, or matching a subset of query segments within a single database record; all have application in MIR.

IV. AUDIO SIMILARITY-BASED RETRIEVAL

The results of the systems described in Sections II and III have been applied to the wide spectrum of tasks collectively called audio similarity-based retrieval. Fig. 10, adapted from [73], illustrates the specificity spectrum for these tasks as well as the audio/music semantic gap that is inherent to it. The tasks on the left of the semantic gap are those requiring matching of specific audio content (high-to-mid specificity systems) and those on the right require matching of high-level music content or concepts only (mid-to-low specificity systems).

Work in audio similarity started at the two extremes of the specificity spectrum and is gradually working its way

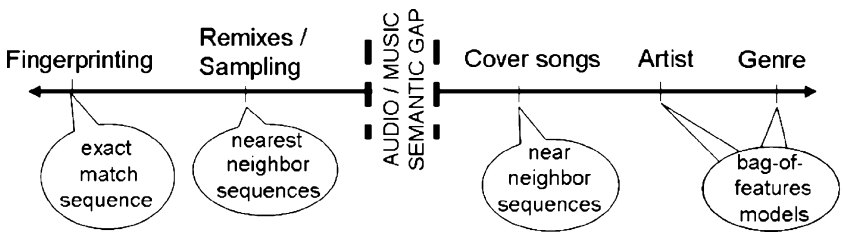

Fig. 10. *Audio similarity specificity spectrum. Tasks on the left of audio/music semantic gap locate specific audio content in a new context. Tasks on the right are invariant to specific audio content but are sensitive to high-level music content. Lower text shows aggregation approach for low-level features.*

inwards toward the semantic gap. It would seem that tasks near the audio/music semantic boundary are among the hardest for MIR. Situated on the right-hand side of the specificity spectrum, an often-cited work is Logan and Salomon [19], who conducted experiments on audio similarity in a database of 8000 tracks, consisting of popular songs in different genres using histograms of 16 clusters of MFCC features per track with clusters extracted using a k-means algorithm. The distance between two Gaussian distributions can be measured using the Kullback-Leibler divergence but this is not possible between two different *sets* of probability distributions, as in histograms of per-track clusters. Instead, they used the earth mover's distance (EMD). Their results reported an average precision of 50% relevant songs in the top five ranked songs for each query.

This work was extended by Aucouturier and Pachet [70] to use Gaussian mixture models (GMM) of MFCC features. Among the evaluation methods used was measuring the precision in terms of the number of songs returned that were labeled with the same *genre* as the query. Relevance was therefore measured by similar cultural metadata to the query.

A large number of other studies has explored the idea of retrieving musical tracks by global statistical features, so many, that in 2006 and 2007 the MIREX exchange conducted evaluations on low-specificity audio similarity [15], [16]. The task was to produce a similarity ranking between 5000 pop music tracks chosen from two pop music test sets. Each entry was tasked to produce a 5000 $\times$ 5000 inter-song distance matrix.

Of the totality of 5000 queries, 60 tracks were randomly selected as queries and the first five most highly ranked songs out of the 5000 were extracted for each query. For each query, the returned results from all participants were grouped and were evaluated by a team of human graders, 24 in 2006. Graders provided a categorical score (*not similar, somewhat similar , and very similar*) and one fine similarity score (0–10) for pairs of tracks. The performance was evaluated by comparing the system results to the human evaluated similarities. The best performing system as evaluated by the human fine-scale similarity judgment scored an average of 0.43 in 2006 and 0.51 in 2007, as shown in Table 3.

On the left-hand side of the audio similarity specificity spectrum are queries for the same audio content. Here, there are audio fingerprinting systems which seek to identify specific recordings in new contexts, as in [60] and [72]; and moving right, towards the center of the spectrum, there are high-to-mid specificity systems where the correct answers to the audio query are not always from the same recording. In this regard, systems for identifying remixed audio content were introduced in [73] and [75]. The motivation behind such systems is to normalize large music content databases so that a plethora of versions of the same recording are not included in a user search and to relate user recommendation data to all versions of a source recording including: radio edits, instrumental, remixes, and extended mix versions.

Finally, to the right of the audio/music semantic gap, in the mid-to-low audio similarity specificity range, the user seeks specific musical content of the query audio but not necessarily the same audio content. These are among the most challenging problems for audio similarity-based MIR systems. Examples are: find the same work by the same performer recorded at a different time or place; find the same work by a different performer (cover songs) [28], [29]; find a different work containing similar melodic content (musical quotations); or find a different work containing similar harmonic sequences (musical idioms). The cover song task was studied in MIREX 2006 and 2007 with an average precision of 52% [29] in 2007.

It must be noted that all of the tasks in the mid-specificity part of the audio-similarity spectrum are solved by approaches using sequences of features; the tasks on the far right of the spectrum are solved by bag-of-frames models that disregard sequences of features.

In the following sections we present two use cases for mid-specificity queries: misattribution, where a recording has been accidentally or deliberately given the incorrect factual information, and opus retrieval, where the goal is to retrieve different versions of the same classical work (a form of cover song retrieval).

1) Use Case 1: Apocrypha (Detecting Misattribution): Apocrypha in audio recordings are those works that are falsely attributed to an artist when they have been performed or composed by a different artist. A recent commercial example of this occurred in the Classical music repertoire, in which a significant number of recordings (100 CDs) were released by the Concert Artists recording label during the 1990s in the U.K. falsely claiming new complete recordings of various Classical composer's repertoires (such as Chopin Mazurkas) [76], [77]. It was eventually discovered that these recordings were re-releases of decades-old performances by different artists under the control of different labels.

It took many years for the experts to discover that these recordings were not authentic [77]. This was in part due to the apocrypha recordings' signal treatment; they were filtered and time-compress/expanded to make them sound slightly different to the originals. The goal of the misattribution task, then, is automatic detection by audio similarity on a large collection of recordings of performances of the same works. Those recordings falling within a predetermined distance threshold, obtained by measuring similarities between recordings known to be different, are candidate misattributions or *apocrypha*. The specificity of this task is similar to that of audio fingerprinting—to establish whether two recordings are acoustically identical, but for some degree of signal transformation and distortion such as filtering or time compression/expansion.

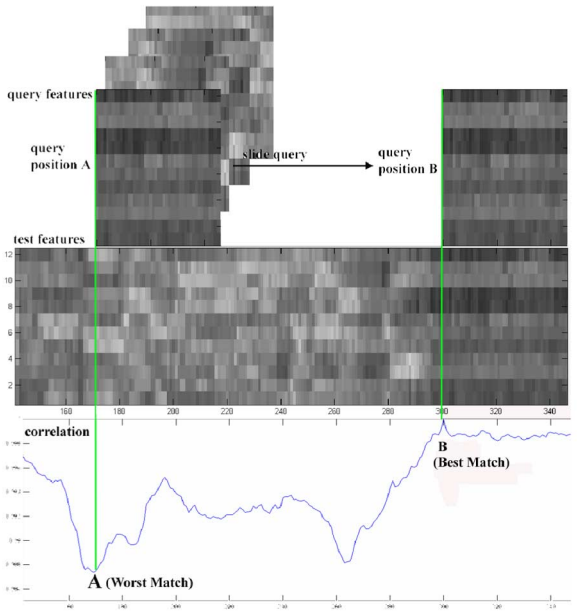

Fig. 11. *Audio shingling; audio features are concatenated into sequences of 1 to 30 seconds duration. Input query sequence is compared with all subsequences of database of the same length. Nearest neighbor search is hence performed by a multidimensional matched filter. Naïve implementations have O(n) time complexity but sublinear algorithms, such as LSH [118], [119], are employed in real-world systems.*

To admit temporally specific similarity, audio features are concatenated into sequences of features called shingles as illustrated in Fig. 11; this is the approach used in [28], [29] and [73]–[75]. Because silent regions between tracks would otherwise match, shingles are ejected from the database if the power falls below a given absolute threshold. To compute the similarity of two tracks, invariant to global structural changes, the pair-wise similarity of all the shingles are computed and counts of the number that fall below a distance threshold. This threshold is determined by measuring the PDF of known nonsimilar shingles for the task. For example, those tracks with a high number of shingles in common are likely candidates for Apocrypha. Fig. 12 shows how the decision boundary for similar shingles is calculated from a sample of known nonsimilar shingles. The decision boundary is the point at which the null hypothesis—that shingles are drawn from the set of nonsimilar shingles—is rejected. Therefore, distances below this value are due to similar shingles.

Using the statistical test method outlined above, all 49 known misattributed recordings in the database of 2700 Chopin Mazurka recordings by 125 artists were retrieved with rank 1; these were the known apocryphal Joyce Hatto, actually Eugene Indjic, recordings that had been reported in the press. However, in the experiment another 11 mis-attributed recordings were found by a second apocryphal pianist, Sergei Fiorentino, whose work

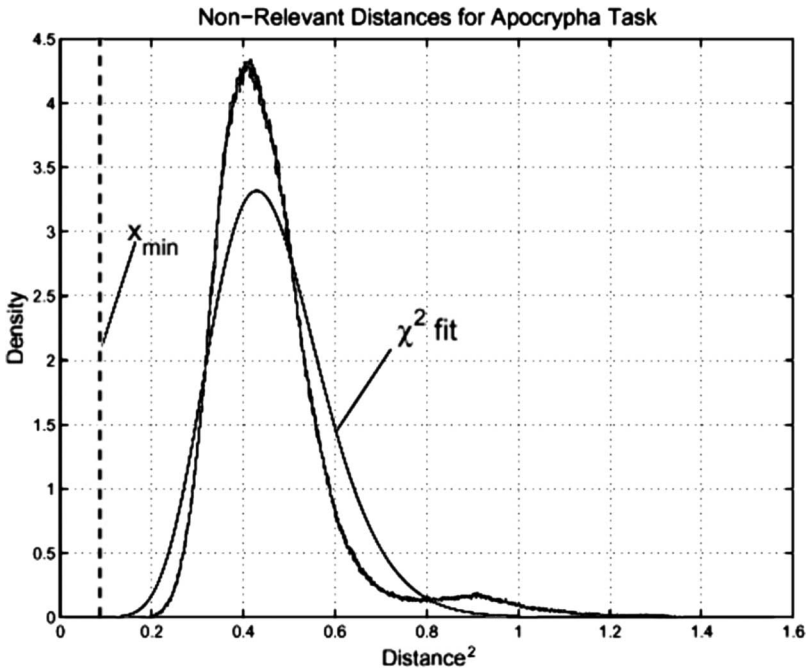

Fig. 12. *Fit of distribution of nonapocrypha shingle distances to a $\chi 2$ distribution using maximum likelihood estimation. Distances below xmin, here 0.09 which is 1% of the CDF, are from similar shingles. Reprinted with permission from Casey et al. [75].*

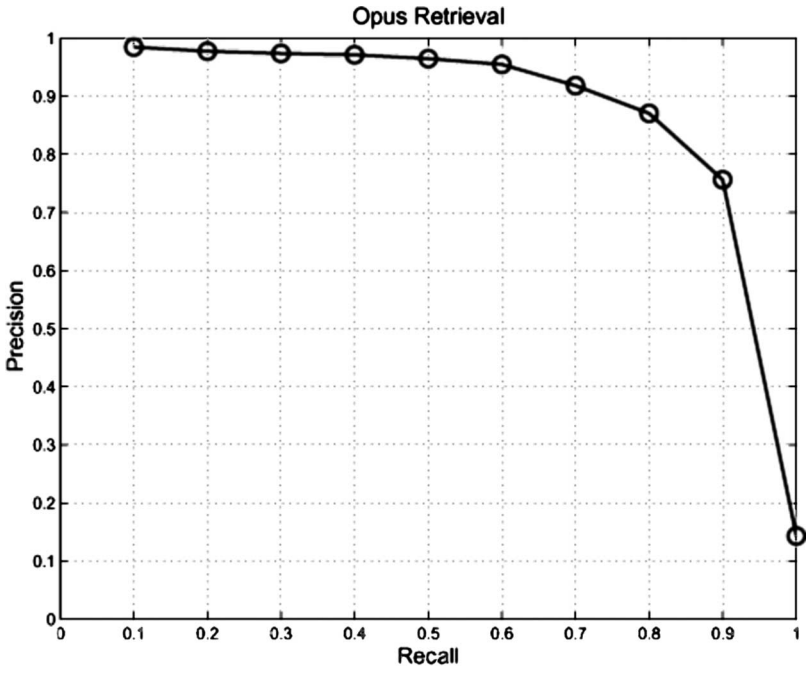

Fig. 13. ***Precision-recall graph for Opus retrieval task. Each of 49 works had between 31 and 65 relevant items out of a 2257-track database. Results indicate 88% precision at 80% recall. Rapid falloff in precision is due to outlying recordings made in early 20th century. Reprinted with permission from Casey et al. [75].***

had been sampled from three different artists. The new mis-attributions were not previously disclosed in the classical music literature. At least four well-known classical music labels' rights had been infringed in this single collection of Chopin recordings. In each case, the infringement was by the same record label.

2) Use Case 2: Opus Retrieval: Opus retrieval starts with a query performance, in this case of one of the Chopin Mazurkas, and retrieves performances of the same work from a large database containing different performances of the same work. The task is difficult because performances by different artists possess significant differences in expressive interpretation of the music. Furthermore, each performance differs in its structure due to choices in performing repeats. In the Apocrypha task, the audio was in common between similar tracks, in the Opus task, it is the musical content that is in common, not the audio.

A precision-recall graph for the Opus retrieval task is shown in Fig. 13. Overall, the precision was very high for recall rates below 90%. For most of the 49 Mazurkas, there were two to three outliers in the database. On inspection, these were typically early recordings that were transferred from 78 r/min shellac media and contained a high degree of surface noise and extreme wideband filtering; additionally, the cutoff frequency for these tracks was typically much lower than the remaining tracks. These results suggest that a near-perfect score can be obtained for Opus retrieval if outlying recordings are first removed or preprocessed to make them compatible with the retrieval method.

V. NOTATED MUSIC

Representing music as a pointset in a 2-D space has a tradition of many centuries. Since the 13th century on, music has been written as a set of notes, represented by points in a 2-D space, with time and pitch as coordinates. Various characteristics are associated with the notes using different symbols for different note durations, for example. The look of written music has changed somewhat over the past centuries, but the basic idea of representing music as a weighted point set has been followed for almost a millennium, and it has served composers and performers well.

In addition to audio analysis and retrieval, MIR on notated music also has applications in commerce, research, and education. Whereas the audio retrieval tasks are primarily useful at the level of genre and specific instances of music, the analysis and retrieval of notated music is more targeted to the level of composers and artists and their works. Because music notation is more symbolic than audio, the feature extraction and analysis are different. However, the matching of derived features can be very similar in both domains.

A. Symbolic Melody Similarity

1) String-Based Methods for Monophonic Melodies: Monophonic music can be represented by 1-D strings of characters, where each character describes one note or one pair of consecutive notes. Strings can represent interval sequences, gross contour, sequences of pitches and the

like, and well-known string matching algorithms such as algorithms for calculating edit distances, finding the longest common subsequence, or finding occurrences of one string in another have been applied, sometimes with certain adaptations to make them suitable for matching melodies.

Some MIR systems only check for exact matches or cases where the search string is a substring of database entries. For such tasks, standard string searching algorithms like Knuth-Morris-Pratt and Boyer-Moore can be used. *Themefinder* [78] searches the database for entries matching regular expressions. In this case, there is still no notion of distance, but different strings can match the same regular expression. For approximate matching, it can be useful to compute an edit distance with dynamic programming. Musipedia is an example of a system that does this [79]. Simply computing an edit distance between query strings and the data in the database is not good enough, however, because these strings might represent pieces of music with different lengths. Therefore, it can be necessary to choose suitable substrings before calculating an edit distance.

More recently, Cilibrasi *et al.* [80] have suggested using an approximation to Kolmogorov distance between two musical pieces as a means to compute clusters of music. They first process MIDI representation of a music piece to turn it into a string from a finite alphabet. Then, they compute the distance between two music pieces using their normalized compression distance (NCD). NCD uses the compressed length of a string as an approximation to its Kolmogorov complexity. The Kolmogorov complexity of a string is not computable, but the compressed length approximation gives good results for musical genre and composer clustering.

For finding substrings that match exactly, the standard methods for indexing text can be used (for example, inverted files, B-trees, etc.). The lack of the equivalent of words in music can be overcome by just cutting melodies into N-grams [81], where each N-gram is a sequence of N pitch intervals. For most edit distances that are actually useful, the triangle inequality holds. Therefore, indexing methods that rely on the triangle inequality property of the distance measure can be used, for example metric trees, vantage point trees, or the vantage indexing method described in [82].

2) Geometry-Based Methods for Polyphonic Melodies: Unlike string-based methods, set-based methods do not assume that the notes are ordered. Music is viewed as a set of events with properties like onset time, pitch, and duration. Clausen *et al.* [83] propose a search method that views scores and queries as sets of notes. Notes are defined by note onset time, pitch, and duration. Exact matches are supersets of queries, and approximate matching is done by finding supersets of subsets of the query or by allowing alternative sets. By quantizing onset times and by segmenting the music into measures, they make it possible to use inverted files.

Since weighted point sets seem to be so well suited to representing music, it feels natural to measure melodic similarity directly by comparing weighted point sets instead of first transforming the music into 1-D abstract representations.

Fig. 14 shows two melodies compared by matching the distribution of weighted points. Typke *et al.* [82] also view scores and queries as sets of notes, but instead of finding supersets, they use transportation distances such as the EMD for comparing sets. They exploit the triangle inequality for indexing, which avoids the need for quantizing. Distances to a fixed set of vantage objects are precalculated for each database entry. Queries then only need to be compared to entries with similar distances to the vantage objects. This approach was very successful in the symbolic melodic similarity track of the MIREX 2006 evaluation.

Ukkonen *et al.* [84] propose a number of algorithms for searching notated music. One method finds translations of

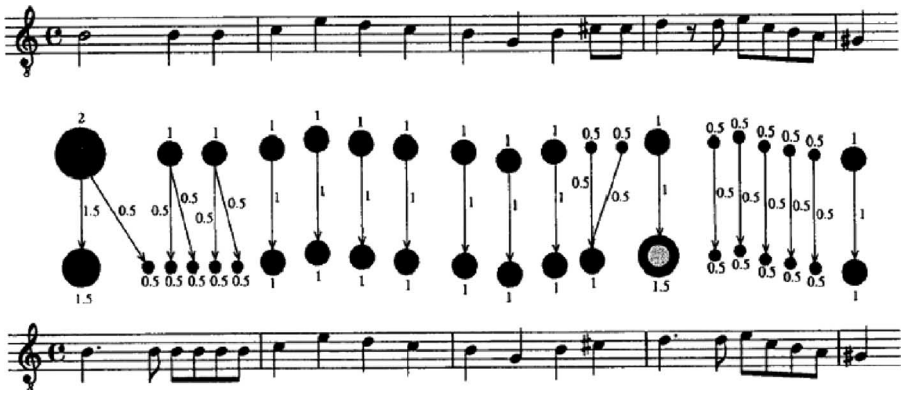

Fig. 14. *Melodies of Handel and Kerll compared as we weighted point sets. Upper point set represents upper melody, lower point set the lower melody. Arrows indicate weight flow that minimizes transportation cost, i.e., corresponding to EMD.*

the query pattern such that all onset times and pitches of notes in the query match with some onset times and pitches of notes in the database documents. Another method finds translations of the query pattern such that some onset times and pitches of the query match with some onset times and pitches of database documents. A third one finds translations of the query pattern that give longest common shared time (i.e., maximize the times at which query notes sound at the same time and with the same pitch as notes from the database documents). This algorithm does not take into consideration whether onset times match.

An interesting approach that takes the process character of melody as starting point is the concept of "melodic expectation." The concept of melodic expectation is based on psychological experiments investigating the relation between expectation and errors in the recall of melodies. Melodic expectation could help to locate the places in the melody where variants are more likely to emerge than at other places within the melody. Related to the melodic expectation is the "Implication/Realization Theory" [85]. This analytical technique was successfully used for defining melodic similarity. It works on the note-to-note level but also incorporates higher hierarchical structures of melody [86]. This approach was very successful in the melodic similarity track of the MIREX 2005 evaluations.

3) Rhythm and Beat Analysis: According to cognitive studies, metric and rhythmic structures play a central role in the perception of melodic similarity, as discussed in Section II. For instance, in the immediate recall of a simple melody studied by Foote [87] the metrical structure was the most accurately remembered structural feature. Rhythmic similarity has been used extensively in the audio domain for classification tasks. In contrast, similarity for symbolic data has been less extensively discussed so far.

Traditionally, rhythmic similarity measures have been evaluated or compared with respect to how well rhythms may be recognized [88], how efficiently they can be retrieved from a data-base [87], or how well they model human perception and cognition of rhythms [89], [90]. In contrast, Toussaint [91] compares rhythmic similarity measures with respect to how much insight they provide concerning the structural interrelationships that exist within families of rhythms, when phylogenetic trees and graphs are computed from the distance matrices determined by these similarity measures. Several similarity measures are compared, including the Hamming distance, the Euclidean interval vector distance, the interval-difference vector distance, the swap distance, and the chronotonic distance.

Inner Metric Analysis [92] describes the inner metric structure of a piece of music generated by the actual notes inside the bars as opposed to the outer metric structure

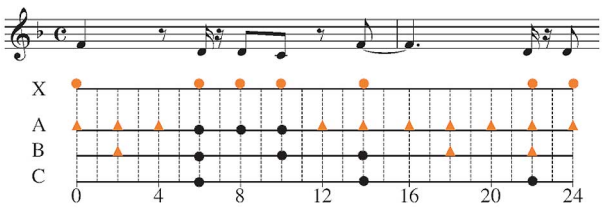

Fig. 15. ***Local meters in excerpt from "The Girl from Ipanema." From [94].***

associated with a given abstract grid such as the bar lines. The model assigns a metric weight to each note of the piece (which is given as symbolic data). The general idea is to search for all pulses (chains of equally spaced events) of a given piece and then to assign a metric weight to each note. The pulses are chains of equally spaced onsets of the notes of the piece called local meters. Consider the set of all onsets of notes in a given piece. Consider every subset of equally spaced onsets as a local meter if it contains at least three onsets and is not a subset of any other subset of equally spaced onsets. The inner metric counts the number of repetitions of the period (distance between consecutive onsets of the local meter) within the local meter. The metric weight of an onset is calculated as the weighted sum of the length of all local meters that coincide at this onset. Fig. 15 illustrates the concept of local meter. Volk *et al.* [93] demonstrate that folksong melodies belonging to the same melody group can successfully be retrieved based on rhythmic similarity. Therefore, rhythmic similarity is a useful characteristic for the classification of folksongs. Furthermore, their results show the importance of rhythmic stability within the oral transmission of melodies, which confirms the impact of rhythmic similarity on melodic similarity suggested by cognitive studies.

VI. MUSIC VISUALIZATION AND BROWSING

Music information retrieval needs user interfaces that facilitate the retrieval, classification, browsing, and management of large collections of music. The most popular and basic interface for browsing a music collection is based on lists of bibliographic (catalogue) information such as titles, artist names, and genres on a display. It typically provides a function for playing back songs in order, a function for searching songs or artists by specifying their bibliographic information, and a function for generating playlists based on bibliographic information. Although it is useful for dealing with a small personal music collection, it is not always useful for browsing a huge online music collection to encounter unexpected but interesting musical pieces or artists. Research on music visualization and browsing for a music collection or a musical piece is

therefore necessary to provide end users with comprehensive and functional interaction.

A. For a Music Collection

To break free from stereotyped thinking of how music playback interfaces must be based on lists of bibliographic information, various interfaces with novel visualization functions have been proposed. Automatic music analysis and music similarity measurement have often been used for this purpose, while the use of human annotations or metadata of musical pieces is also popular for commercial services.

To visualize and browse a collection of musical pieces, those pieces can be projected onto a plane or space by automatically analyzing their audio features. For example, Tzanetakis *et al.* [95] developed the "GenreSpace" interface for browsing music collections in a three-dimensional space into which musical pieces are projected according to their similarity. Pampalk *et al.* [96] then reported the "Islands of Music" interface featuring self-organizing maps (SOMs) that projects musical pieces onto a plane. They used a metaphor of "islands" that represent self-organized clusters of similar pieces. By using a small focused collection by a single composer, musical pieces can be mapped in the shape of the silhouette of its composer [97]. Another visualization technique "U-Map" [98] using a variant of SOM called Emergent SOM (ESOM) was also proposed. The Traveller's Sound Player [99] uses the Traveling Salesman algorithm to map musical pieces of a collection on a circle and visualizes the distribution of metadata (e.g., genre and tempo) having a certain value by changing the color of the corresponding region on the circle.

Instead of visualizing the whole collection, a part of the collection can be dynamically shown to induce active user interaction. *Musicream* [100] is such a user interface for discovering and managing musical pieces. As shown on the right side in Fig. 16, disc icons representing pieces flow one after another from top to bottom, and a user can select a disc and listen to it. By dragging a disc in the flow, the user can easily pick out other similar pieces (attach similar discs). This interaction allows a user to unexpectedly come across various pieces similar to other pieces the user likes. *Musicream* also gives a user greater freedom of editing playlists by generating a playlist of playlists. Since all operations are automatically recorded, the user can visit and retrieve a past state as if using a time machine.

While the above interfaces focus on the level of musical pieces, there are interfaces that focus on the level of artists. For example, Van Gulik *et al.* [101] reported the "Artist Map" interface with the focus on artists and small devices. It enables users to explore and discover music collections on small devices by projecting artists into a 2-D space. Artists are drawn as dots in the space so that similar artists are placed close together on the basis of a modified spring-embedder algorithm. This visualization can also be used to make playlists by drawing paths and specifying regions on top of the visualization [102].

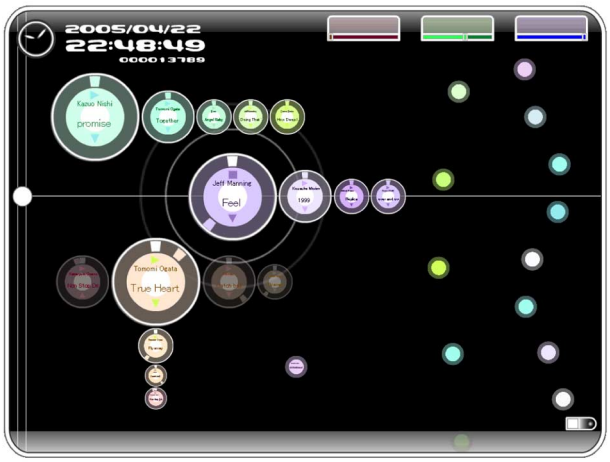

Fig. 16. *Musicream: User can actively browse a music collection to discover musical pieces.*

Artist-level similarity computed on the basis of piece-level similarity can be used with a web-based technique for automatically labeling artists with words. *MusicRainbow* [103] is such a user interface for discovering unknown artists. As shown in Fig. 17, all artists in a music collection are mapped on a circular rainbow where colors represent different styles of music. Similar artists are automatically mapped near each other by using the Traveling Salesman algorithm and summarized with word labels at three different hierarchical levels. A user can rotate the rainbow by turning a knob and find an interesting artist by referring to the word labels. By pushing the knob, the user can select and listen to the artist highlighted at the midpoint on the

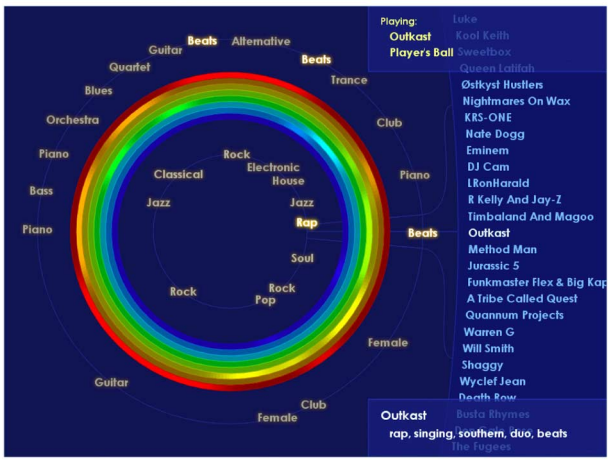

Fig. 17. *MusicRainbow: User can actively browse a music collection to discover artists.*

right side. *MusicRainbow* is based on the content-based similarity between artists, which is computed from the audio-based similarity between musical pieces. The artists are then summarized with word labels extracted from web pages related to the artists.

On the other hand, by using metadata of sound files without analyzing audio signals, Torrens *et al.* [104] reported visualization techniques where musical pieces are placed in a circle, rectangle, or tree map. When visual information related to musical pieces is available, a "collaging" technique proposed by Bainbridge *et al.* [105] is also an effective way to provide leisurely, undirected interaction with a music collection.

B. For a Musical Piece

As is apparent from music (promotion) videos, music playback accompanied by visual images enables end users to immerse themselves in the music or simply enjoy music more. In fact, recent computer-based music players often support a music visualization function that shows music-synchronized animation in the form of geometrical drawings moving synchronously with waveforms and frequency spectrums.

If the visualization could be more closely related to the musical content, it would provide end users with unique experiences. For example, by using an automatic beat tracking method, the animation could be rigidly synchronized with the beats. Cindy [37] is a music-synchronized visualization system that displays virtual dancers whose motions and positions change in time to musical beats in real time. This system has several dance sequences, each for a different dance motion mood. A user can select dance sequences one after another by pressing buttons during music playback. By using an automatic genre classification method, on the other hand, *GenreGram* [95] shows a dynamic real-time visualization consisting of genre cylinders. Each cylinder represents a different genre and is texture mapped with a representative image of its genre. It moves up and down during music playback according to a genre-classification confidence measure, revealing correlations of different genre decisions. Since the boundaries between musical genres are fuzzy in general, a display like this is informative and useful.

Automatic visualization of the music structure is also interesting and helps a user understand the structure easily. By using a method that estimates chorus sections and various repeated sections with a focus on popular music, *SmartMusicKIOSK* [61] shows a "Music Map" (Fig. 18) that is a visual representation of the entire song structure consisting of chorus sections (the top row) and repeated sections (the five lower rows). On each row, colored sections indicate similar (repeated) sections. This visualization also facilitates browsing within a musical piece. *SmartMusicKIOSK* provides a content-based playback-control interface for within-song browsing or trial listening for popular music. With this interface, a user can easily jump and listen to the chorus with just a push of a button and skip sections of no interest by interactively changing the playback position while viewing the "Music Map."

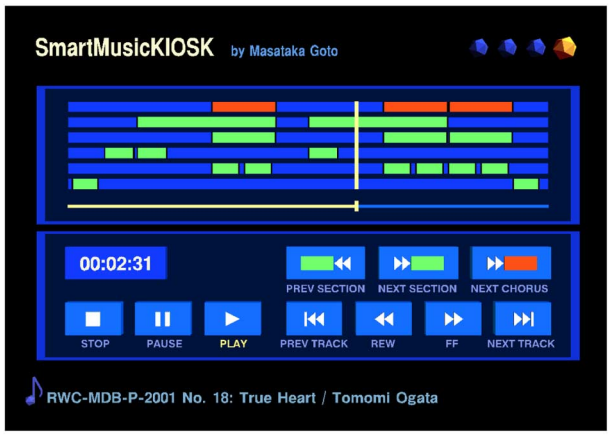

Fig. 18. ***SmartMusicKIOSK: User can actively listen to various parts of a song while moving back and forth as desired on the visualized song structure ("Music Map" in upper window).***

Timbral information within a musical piece can also been visualized. For example, *TimbreGrams* [106] are a graphical representation consisting of a series of vertical color stripes where the color of each stripe corresponds to a short-time audio feature. Similar pieces tend to have similar colors. Since time is mapped from left to right, time periodicity can easily be found in color. By using an automatic musical-instrument recognition method, on the other hand, *Instrogram* [107] shows a spectrogram-like graphical representation that enables a user to find when which instruments are used in a musical piece. *Instrogram* consists of several images, the number of which is the same as with the number of target instruments. Each spectrogram-like image with time and frequency axes corresponds to a different instrument and represents the probability that its instrument is used at each time-frequency region in polyphonic sound mixtures.

C. Conclusion

To open new possibilities for various user interfaces with better visualization and browsing functions, further research on music-understanding technologies based on signal processing as well as novel interaction and visualization techniques will play important roles. Some of the above interfaces can also be considered as active music listening interfaces [108] that enable nonmusician users to enjoy music in more active ways than conventional passive music consumption.

VII. MIR SYSTEMS

At present, a number of tools and frameworks are in development for combining MIR components into systems

that perform feature extraction, music retrieval, and automatic evaluation. Many of the tools are Open Source and, as such, the research community is encouraged to participate in their development.

A. Music Research Tools

This section briefly discusses a number of music information retrieval systems. They are divided into music research tools and music performance tools.

1) MARSYAS (sourceforge.net/projects/marsyas): Music Analysis, Retrieval and Synthesis for Audio Signals [109] is a collection of Open Source C++ tools developed at Princeton University and, subsequently, the University of Victoria, Canada, for extracting features and performing machine learning and music retrieval tasks on collections of music. It is a low-level audio framework and is primarily targeted at MIR researchers and developers.

2) CLAM (clam.iua.upf.edu): The C++ Library for Audio and Music [110] was developed at the University Pompeu Fabra (UPF), Barcelona, Spain. It is a similar to MARSYAS in that it provides a software framework for performing feature extraction and music synthesis. It extends the concepts in MARSYAS by providing an extensive set of graphical user interfaces allowing developers to produce feature rich applications targeted at nonprogrammer music users.

3) M2K (www.music-ir.org/evaluation/m2k): Music to Knowledge is a framework developed at the University of Illinois, Urbana-Champaign, that is primarily targeted at evaluating music information retrieval systems [111]. It extends the ideas of MARSYAS and CLAM by providing a graphical user interface for constructing MIR systems. M2K also supports automatic evaluation of MIR tasks and is a key component of the MIREX evaluation experiments.

4) OMRAS2 (omras2.org): Online Music Recognition and Searching II is a distributed MIR framework jointly under development at Goldsmiths, University of London, and Queen Mary, University of London. Expanding on the concepts of the previous frameworks, OMRAS2's goals are to integrate MIR system components into user-facing software tools that can help MIR researchers, musicologists, and commercial music services explore the use of content-based methods. Already released components include: *SonicVisualiser*, a low-level audio feature extraction and visualization tool that uses the audio-editor interaction paradigm for researchers to closely inspect automatically extracted information from audio files; the *MusicOntology*, a metadata scheme constructed using the World Wide Web Consortium's Resource Description Framework (RDF) for describing music and software resources on the Semantic Web; *AudioDB*, a low-level audio feature database that scales to storing and searching features for large music collections, see Fig. 20; and *FFTExtract*, a tool for batch extraction of low-level audio features as discussed in Section II of this paper. Fig. 19 shows the OMRAS2 framework with system components reflected on the Semantic Web as resources and services.

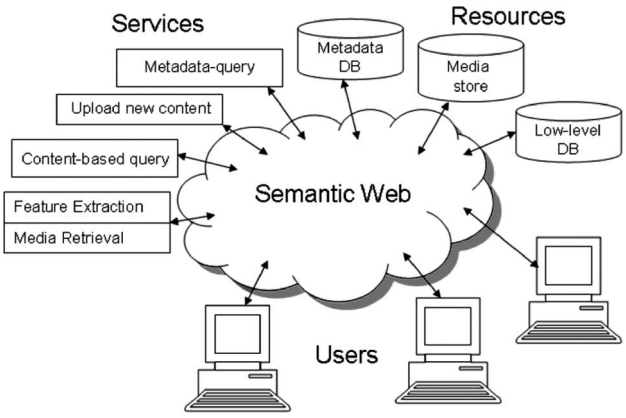

Fig. 19. ***Overview of OMRAS2 framework showing use of Semantic Web to make resources and services available to MIR researchers, music researchers, and music search services.***

B. Music Performance Tools

In addition to the MIR frameworks discussed previously, a large number of systems have been proposed, and deployed, that support music performance. The first of these systems emerged in 1984 independently by Vercoe [112] and Dannenberg [113]. These systems were used by composers to synchronize live musical performance

Fig. 20. ***Retrieval interface for AudioDB content-based search engine, part of OMRAS2 system. Interface shows alignment of best matching time points within result tracks to given time-point in query track. Clicking on musical note icon plays track from best-matching position.***

to electronic accompaniment in the context of an electronic music performance. A number of recently proposed systems extend the ideas set forth in these early MIR systems [114].

Another type of music performance system is known as concatenative music synthesis. Here, the goal is to create music by retrieving musical segments, in real time, that match a live audio input stream. The sense of the matching can either be nearest neighbor or the intention can be to find audio that can accompany the current live input. Such systems include Schwarz's *Caterpillar* system [116] which uses principles of concatenative speech synthesis to constrain matching of music segments to a target; Casey's *SoundSpotter* [115] which matches a live target stream of audio shingles to a large database of audio-visual material in real time; and Collins' audio-visual concatenative synthesis [117], which combines audio and visual features for matching. Such systems open the door to collaboration between musicians and MIR in new performance contexts and for developing new types of musical experience.

VIII. CONCLUSION: CHALLENGES AND FUTURE DIRECTIONS

We conclude the paper with an account of the remaining significant challenges for the field of content-based music information retrieval. Many of the issues that remain as challenges have started to be addressed in the most current systems, but it will likely be some time before these difficult challenges are resolved.

A. Scaling Content-Based Search to Millions of Tracks

Pandora is an example of a working MIR system. It is a popular music service that has a large user base. But it does not scale to the size of a million-track database because of the effort required in describing a track and the consistency requirements on track description. This is why we currently do not see, for example, content-based MIR as part of the *iTunes* interface: the current tools and techniques do not yet scale to millions of tracks. Next generation MIR systems must address this and bring content-based methods to the larger music services and digital libraries.

As an illustration, current MIR systems operate on the scale of 10 000 tracks; for example, the MIREX audio similarity experiment required computation of 5000×5000 similarity matrices, so the task required pair-wise track comparisons on the order of millions. However, today's music download services and music and video downloads sites are on the order of tens of millions of tracks, so pair-wise comparisons require computation on the order of hundreds of trillions. Clearly, pair-wise methods are intractable at this scale on even the most advanced hardware; however, this scale is required to make content-based MIR practical for solving today's media search and retrieval problems.

One approach to reducing the time complexity of pair-wise distance computation that has much potential is locality sensitive hashing (LSH), which trades time complexity of the query for space complexity of the database and a possible slight loss of accuracy [118], [119]. The gain is significant; a standard pair-wise computation is $O(n^2)$ where n is the number of features that are compared. LSH systems perform the similarity computation in $O(n^{(1/c)})$ with c an approximation factor that is greater than one. This sublinear time complexity makes LSH an attractive proposition; the tradeoff in space complexity is not as critical as time complexity due to the ease of availability of large storage at low cost. Additionally, LSH comes with a probabilistic guarantee that the returned points are within a given radius of the query, so systems can be tuned to tradeoff precision for speed under well understood bounds. Armed with such methods, content-based MIR systems can scale to similarity computations on the order of millions of tracks, a scale that was, until only recently, entirely out of reach.

B. Integration of Tools and MIR Frameworks

The music information research community has created a significant number of tools and frameworks, as discussed in Section VII. To create new systems it is beneficial to integrate such existing tools because they have well-known properties and are widely understood by researchers and developers.

Integrating such tools is not easy. This is largely because the input, output, and parameter formats are completely different between frameworks. A significant amount of effort in recent years has been placed in attempting to standardize description formats, such as the MPEG-7 International Standard for Multimedia Content description, and in providing vocabularies for describing concepts that are used for describing content, tools, and methods [120], [121]. Standardization among MIR tools has not been widely adopted at this point, but this will become essential as new applications and ways to combine existing tools emerge. Cooperation among the communities of researchers and developers working in music and multimedia information retrieval will be required to make such integration happen.

C. Content Description of Polyphonic Music

Sections II and III of this paper discussed the exaction of low-level and high-level music features. The low-level audio methods typically treat polyphonic music *en masse* so that music with multiple instruments playing simultaneously is represented as a sum of the information across instruments. Great strides have been made in recent years in source separation methods for extracting information about individual musical parts from a polyphonic mixture. This type of processing is one of the most challenging

problems in audio processing, but new work in the area looks promising. In particular, methods based on independent components analysis and sparse coding coupled with Bayesian methods will likely lead to new results in these areas.

With the extraction of high-level music features comes the question of how they might be used to improve the results for tasks and use cases. Thus, as the field grows, the intermediate goals of high-level music extraction will give way to task-oriented research using such features.

D. Addressing the Semantic Gap

Aucouturier and Pachet [122] conducted experiments, similar to the audio similarity experiments discussed in Section IV, that suggested a glass ceiling in performance of around 65% accuracy, no matter what type of system was employed. Similarly, Paivo [123], taking melody extraction from polyphonic audio as a case study, showed bottom-up low-level audio feature approaches are likely to have reached a performance ceiling. The engineering results of such systems are far from being sufficiently robust to be considered solved problems. Similar observations have been made in rhythm and timbre recognition: the use of powerful stochastic and probabilistic modeling techniques (including hidden Markov models, Bayesian modeling, support vector machines, and neural networks) do not appear to provide a solution for closing the gap between the acoustical and the semantic level [124]. The connection between the objective measurement of audio features and the subjective description of musical perception and emotional experiences turns out to be a hard problem. Among music experts, as in [125], there is a growing understanding that the bottom-up approach may be too narrow, leaving a semantic gap that poses a difficult challenge to future systems.

To improve the performance of MIR systems, the findings and methods of music perception and cognition could lead to better understanding of how humans interpret music and what humans expect from music searches. At the intersection of MIR research and music cognition research are the following approaches.

1) Perceptive/Cognitive Approach: Research on music perception and cognition has a long tradition and strongly draws upon the achievements of psychoacoustics, auditory physiology, Gestalt psychology, and systematic musicology [126]. Traditionally, the main focus has been on the perception and representation of musical parameters: such as pitch, melody, harmony, tonality, onset, offset, beat, tempo, rhythm, and so on, as in [127]. Understanding these features in human perception is often based on the combination of two methods, namely: experimenting and modeling. Experiments contribute to the empirical and evidence-based knowledge of human perception and cognition, while modeling contributes to our understanding of the dynamics of the systems underlying perception and cognition. Representations of melody, rhythm, and tonality allow users to access music at a high semantic level corresponding to high-level musical features as discussed in Section II.

2) Emotional/Affective Approach: In line with the above approach, several authors have attempted to extend these parameters with emotional/affective concepts [128], [129], corresponding with cultural metadata discussed in Section II, the rationale being that users are often not familiar with descriptions of musical parameters. Research on music and emotions [130], [131] suggests strong correlations between emotional/affective descriptions and various musical parameters. Recent approaches in MIR and in music cognition have studied the relationship between low-level perceptual/acoustic features and semantic emotional/affective concepts such as mood [124], [128]; this is a promising new direction for the field.

E. User Preference Modeling

A further approach to improving MIR systems is user modeling. Many new services are emerging that are based around the gathering and modeling of user preference data. These have their roots in recommender systems such as found in *Amazon*, *Pandora*, *last.fm*, and many others. The advantage of modeling user preferences over attempting to solve the audio similarity problem alone is that communities of users with different tastes are addressed separately. Recent studies suggest that item-to-item similarity computed from user-preference data is a better measure of acoustic similarity than that provided by systems based on acoustic data; this approach only works when the song is well known and a large corpus of preference data is available. Such data might be a better way to get ground truth for playlist generation rather than relying on subjective acoustic similarity judgments from a small pool of experts [132]. Anonymous user preference data is available from some music services, such as *last.fm*. User preference data provides the linkages between music according to use, so it makes for a supreme set of relevance data when compared with manual evaluation exercises.

A recent study on semantic description of music [133] addresses the problem of representative population and the role of subjective background. It was found that gender, age, musical expertise, active musicianship, breadth of taste, and familiarity with the music have a large influence on semantic descriptions. Using statistical analysis on a representative sample of the population interested in music information retrieval, it was possible to accommodate for the variance in user profiles and take that into account for the semantic description of music. The study illustrates that content-based music information retrieval systems may profit from an analysis of the subjective background of users and that such top-down

knowledge may be useful in addressing musical content description.

F. User Focus

As a final observation for MIR, it is interesting to note that most of the activity in the field has been engineering-led. There have been very few user studies that attempt to understand and evaluate the way that MIR tools get used by nonresearch communities. New research is required to better understand: requirements on user control of search; integration of MIR with professional work flows such as music production and engineering, musicology research, and music archiving; and how users navigate million-song music download services.

Just as with the early development of the field, it is likely that the solutions will be forged out of interdisciplinary collaboration and understanding between the fields of musicology, music perception and cognition, information retrieval, large databases, theoretical computer science, signal processing, machine learning, audio engineering and user interaction design as well as the involvement of commercial organizations in the development of future users and markets and the roles for MIR in serving them. ■

Acknowledgment

The authors acknowledge the valuable contributions of the anonymous expert reviewers, whose insights and suggestions were extremely valuable to this paper and T. Crawford, B. Fields, M. d'Inverno, M. Magas, and P. Proutskova at Goldsmiths, University of London, and A. Volk and F. Wiering at Utrecht University.

REFERENCES

[1] A. Edgecliffe-Johnson. (2006, Feb. 10). "Rivals aim at iTunes' grip on music downloads," *Financial Times*. [Online]. Available: www.ft.com

[2] S. Douglas. (2007, Oct. 12). *U.K. music downloads reach record high*. [Online]. Available: www.investmentmarkets.co.uk

[3] B. Bland. (2007, Oct. 31). "Illegal music downloads hit record high," *Telegraph*. [Online]. Available: www.telegraph.co.uk

[4] E. Terazono. (2007, May 30). "Last.fm founders in windfall after sale," *Financial Times*. [Online]. Available: www.ft.com

[5] O. Celma, M. Ramírez, and P. Herrera, "Foafing the music: A music recommendation system based on RSS feeds and user preferences," in *Proc. 6th Int. Conf. Music Information Retrieval*, London, U.K., 2005.

[6] A. Jha, "Music machine to predict tomorrow's hits," *Guardian*, Jan. 17, 2006.

[7] B. Bell and B. Vecchione, "Computational musicology," *Computers Humanities*, vol. 27, no. 1, Jan. 1993.

[8] E. Clarke and N. Cook, *Empirical Musicology: Aims, Methods, Prospects*. New York: Oxford Univ. Press, 2004.

[9] A. Freed, "Music metadata quality: A multiyear case study using the music of Skip James," in *Proc. AES 121st Conv.*, San Francisco, CA, 2006.

[10] A. Wang, "The Shazam music recognition service," *Com. ACM*, vol. 49, no. 8, pp. 44–48, 2006.

[11] D. Byrd, *A similarity scale for content-based music IR*. [Online]. Available: www.informatics.indiana.edu/donbyrd/MusicSimilarityScale.html

[12] B. Whitman and R. Rifkin, "Musical query-by-description as a multiclass learning problem," in *Proc. IEEE Multimedia Sig. Proc. Conf.*, St. Thomas, Dec. 2002.

[13] N. Hu and R. Dannenberg, "A comparison of melodic database retrieval techniques using sung queries," in *Proc. ACM Joint Conf. Digital Libraries*, 2002, pp. 301–307.

[14] E. Selfridge-Field, "Conceptual and representational issues in melodic comparison," in *Melodic Similarity—Concepts, Procedures, and Applications*, W. B. Hewlett and E. Selfridge-Field, Eds. Cambridge, MA: MIT Press, 1998.

[15] J. S. Downie, "The music information retrieval evaluation exchange (MIREX)," *D-Lib Mag.*, vol. 12, no. 12, 2006.

[16] J. S. Downie, K. West, A. Ehmann, and E. Vincent, "The 2005 music information retrieval evaluation exchange," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005.

[17] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Amer.*, vol. 92, no. 5, pp. 2698–2701, 1992.

[18] X. Wu and M. Li, "A QBSH system based on a three-level melody representation," in *Proc. Int. Conf. Music Information Retrieval*, Vienna, Austria, 2007.

[19] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Tokyo, Japan, 2001.

[20] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. Int. Computer Music Conf.*, Beijing, China, 1999.

[21] E. Gómez and E. Chew, Eds., "Tonal description of polyphonic audio for music content processing," *INFORMS J. Computing, Special Cluster Computation in Music*, vol. 18, no. 3, 2006.

[22] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 15–18.

[23] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Conf. Acoustics, Speech and Signal Processing*, 1999.

[24] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[25] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex domain onset detection for musical signals," in *Proc. Int. Conf. Digital Audio Effects*, London, U.K., 2003.

[26] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Munich, Germany, 1997.

[27] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph.D. dissertation, Vienna Univ. Technology, Vienna, Austria, Mar. 2006.

[28] D. P. W. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. Int. Conf. Acoustic, Speech and Signal Processing*, Honolulu, HI, 2007.

[29] J. Serra and E. Gomez, "A cover song identification system based on sequences of tonal descriptors," in *Proc. Int. Conf. Music Information Retrieval*, Vienna, Austria, 2007.

[30] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Res.*, vol. 30, no. 2, pp. 159–171, 2001.

[31] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 588–601, 1998.

[32] W. A. Sethares and T. W. Staley, "Meter and periodicity in musical performance," *J. New Music Res.*, vol. 30, no. 2, pp. 149–158, 2001.

[33] W. A. Sethares and R. Arora, "Equilibria of adaptive wavetable oscillators with applications to beat tracking," in *Proc. Int. Cont. Acoustics, Speech Signal Processing*, Honolulu, HI, 2007.

[34] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. New Music Res.*, vol. 30, no. 1, pp. 39–58, 2001.

[35] S. W. Hainsworth and M. D. Macleod, "Particle filtering applied to musical tempo tracking," *EURASIP J. Applied Signal Processing*, vol. 15, pp. 2385–2395, 2004.

[36] A. T. Cemgil and B. Kappen, "Monte Carlo methods for tempo tracking and rhythm quantization," *J. Artificial Intell. Res.*, vol. 18, pp. 45–81, 2003.

[37] M. Goto and Y. Muraoka, "A virtual dancer 'Cindy'—Interactive performance of a music-controlled CG dancer," in *Proc. Lifelike Computer Characters*, 1996, p. 65.

[38] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Speech Audio Proc.*, vol. 14, no. 1, Jan. 2006.

[39] R. B. Dannenberg, "Toward automated holistic beat tracking, music analysis, and understanding," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005, pp. 366–373.

[40] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 5, pp. 1832–1844, May 2006.

[41] H. Fujihara and M. Goto, "A music information retrieval system based on singing voice timbre," in *Proc. Int. Conf. Music Information Retrieval*, Vienna, Austria, 2007.

[42] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in *Proc. IEEE Int. Symp. Multimedia*, San Diego, CA, 2006, pp. 257–264.

[43] M. Goto and S. Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," in *Working Notes Int. Joint Conf. Artifical Intelligence, Workshop Computational Auditory Scene Analysis*, 1999, pp. 31–40.

[44] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Commun.*, vol. 43, no. 4, pp. 311–329, 2004.

[45] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "F0 estimation method for singing voice in polyphonic audio signals based on statistical vocal models and Viterbi search," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, pp. V-253–V-256.

[46] M. Marolt, "Gaussian mixture models for extraction of melodic lines from audio recordings," in *Proc. Int. Conf. Music Information Retrieval*, Barcelona, Spain, 2004, pp. 80–83.

[47] R. P. Paiva, T. Mendes, and A. Cardoso, "An auditory model based approach for melody detection in polyphonic musical recordings," in *Proc. 2nd Int. Symp. Computer Music Modeling Retrieval*, 2004.

[48] Y. Li and D. Wang, "Detecting pitch of singing voice in polyphonic audio," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2005, pp. III-17–III-20.

[49] J. Eggink and G. J. Brown, "Extracting melody lines from complex audio," in *Proc. Int. Conf. Music Information Retrieval*, 2004, pp. 84–91.

[50] G. E. Poliner and D. P. W. Ellis, "A classification approach to melody transcription," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005, pp. 161–166.

[51] M. Ryynanen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. Int. Conf. Music Information Retrieval*, Victoria, Canada, 2006.

[52] A. Sheh and D. P. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proc. Int. Conf. Music Information Retrieval*, Baltimore, MD, 2003.

[53] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005.

[54] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio," *IEEE Trans. Acoust., Speech Language Process.*, to be published.

[55] C. A. Harte, M. B. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. Audio Music Computing for Multimedia Workshop*, Santa Barbara, CA, 2006.

[56] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1247–1256, Apr. 2007.

[57] G. Fazekas and M. Sandler, "Intelligent editing of studio recordings with the help of automatic music structure extraction," in *Proc. 122nd AES Convention*, Vienna, Austria, 2007.

[58] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005.

[59] M. Levy and M. Sandler, "Signal-based music searching and browsing," in *Proc. Int. Conf. Consumer Electronics*, 2007.

[60] A. Wang, "An industrial-strength audio search algorithm," in *Proc. Int. Conf. Music Information Retrieval*, 2003, pp. 713–718.

[61] M. Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 5, pp. 1783–1794, May 2006.

[62] M. Levy, M. Sandler, and M. Casey, "Extraction of high-level musical structure from audio data and its application to thumnail generation," in *Proc. Int. Conf. Acoustics, Speech Signal Processing*, 2006, pp. V-1316–V-1319.

[63] S. Abdallah, M. Sandler, C. Rhodes, and M. Casey, "Using duration models to reduce fragmentation in audio segmentation," *Machine Learning J.*, vol. 62, no. 2–3, pp. 485–515, 2006.

[64] G. Peeters, A. La Burthe, and X. Rodet, "Toward automatic music audio summary generation from signal analysis," in *Proc. Int. Conf. Music Information Retrieval*, Paris, France, 2002, pp. 100–106.

[65] J. R. Saffran, E. K. Johnson, R. N. Aslin, and E. L. Newport, "Statistical learning of tone sequences by human infants and adults," *Cognition*, vol. 70, pp. 27–52, 1999.

[66] N. Tan, R. Aiello, and T. G. Bever, "Harmonic structure as a determinant of melodic organization," *Memory Cognition*, vol. 9, no. 5, pp. 533–539, 1981.

[67] D. Müllensiefen, D. Lewis, C. Rhodes, and G. Wiggins, "Evaluating a chord-labelling algorithm," in *Proc. Int. Conf. Music Information Retrieval*, 2007, pp. 317–318.

[68] M. Müller and F. Kurth, "Towards structural analysis of audio recordings in the presence of musical variations," *EURASIP J. Advances in Signal Processing*, 2007.

[69] C. Rhodes and M. Casey, "Algorithms for determining and labelling approximate hierarchical self-similarity," in *Proc. Int. Conf. Music Information Retrieval*, Vienna, Austria, 2007.

[70] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?" in *Proc. Int. Conf. Music Information Retrieval*, Paris, France, 2002.

[71] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *J. VLSI Signal Processing Systems for Signal Image Video Technology*, vol. 41, no. 3, pp. 271–284, 2005.

[72] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, T. Kastner, and M. Cremer, "Content-based identification of audio material using MPEG-7 low-level description," in *Proc. Int. Conf. Music Information Retrieval*, Bloomington, IN, 2001.

[73] M. Casey and M. Slaney, "The importance of sequences in music similarity," in *Proc. Int. Conf. Acoustics, Speech Signal Processing*, Toulouse, France, 2006.

[74] M. Casey and M. Slaney, "Song intersection by approximate nearest neighbor search," in *Proc. Int. Conf. Music Information Retrieval*, Victoria, Canada, 2006.

[75] M. Casey, C. Rhodes, and M. Slaney, "Analysis of minimum distances in high dimensional musical spaces," in *IEEE Trans. Acoust., Speech Language Process.*, accepted for publication.

[76] N. Cook and C. Sapp. (2007). *Purely coincidental? Joyce Hatto and Chopin's Mazurkas*, Centre for History and Analysis of Recorded Music, Royal Holloway, Univ. London, London, U.K. [Online]. Available: http://www.charm.rhul.ac.uk/

[77] C. Joseph and A. Luck, "Revenge of the fraudster pianist," *Mail on Sunday*, Feb. 24, 2007.

[78] A. Kornstädt, "Themefinder: A web-based melodic search tool," in *Melodic Similarity: Concepts, Procedures, and Applications, Computing in Musicology*, vol. 11, W. Hewlett and E. Selfridge-Field, Eds. Cambridge, MA: MIT Press, 1998.

[79] L. Prechelt and R. Typke, "An interface for melody input," *ACM Trans. Computer-Human Interaction*, vol. 8, no. 2, pp. 133–149, 2001.

[80] R. Cilibrasi, P. Vitanyi, and R. de Wolf, "Algorithmic clustering of music based on string compression," *Computer Music J.*, vol. 28, no. 4, pp. 49–67, 2004.

[81] J. S. Downie, "Evaluating a simple approach to music information retrieval: Conceiving melodic n-grams as text," Ph.D. dissertation, Univ. Western Ontario, Canada, 1999.

[82] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R. van Oostrum, "Using transportation distances for measuring melodic similarity," in *Proc. Int. Conf. Music Information Retrieval*, Baltimore, MD, 2003, pp. 107–114.

[83] M. Clausen, R. Engelbrecht, D. Meyer, and J. Schmitz, "PROMS: A web-based tool for searching in polyphonic music," in *Proc. Int. Conf. Music Information Retrieval*, Plymouth, MA, 2000.

[84] E. Ukkonen, K. Lemström, and V. Mäkinen, "Geometric algorithms for transposition invariant content-based music retrieval," in *Proc. Int. Conf. Music Information Retrieval*, Baltimore, MD, 2003, pp. 193–199.

[85] E. Narmour, *The Analysis and Cognition of Basic Melodic Structures*. Chicago, IL: Univ. Chicago Press, 1990.

[86] M. Grachten, J. L. Arcos, and R. Lápez de Mántaras, "Melody retrieval using the implication/realization model," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005.

[87] J. Foote, M. Cooper, and U. Nam, "Audio retrieval by rhythmic similarity," in *Proc. Int. Conf. Music Information Retrieval*, Paris, France, 2002.

[88] E. J. Coyle and I. Shmulevich, "A system for machine recognition of music patterns,"

in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, Seattle, WA, 1998.

[89] L. Hofmann-Engl, "Rhythmic similarity: A theoretical and empirical approach," in *Proc. Int. Conf. Music Perception Cognition*, 2002, pp. 564–567.

[90] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Anchor space for classification and similarity measurement of music," in *Proc. Int. Conf. Multimedia Expo*, Baltimore, MD, 2003.

[91] G. Toussaint, "A comparison of rhythmic similarity measures," School of Computer Science, McGill Univ. Montreal, Quebec, Canada, Tech. Rep. SOCS-TR-2004.6, 2004.

[92] A. Fleischer, *Die analytische Interpretation. Schritte zur Erschließung eines Forschungsfeldes am Beispiel der Metrik. dissertation*, Verlag im Internet Gmbh, Berlin, Germany, 2003.

[93] A. Volk, J. Garbers, P. van Kranenburg, F. Wiering, R. C. Veltkamp, and L. P. Grijp, "Applying rhythmic similarity based on inner metric analysis to folksong research," in *Proc. Int. Conf. Music Information Retrieval*, Vienna, Austria, 2007.

[94] E. Chew, A. Volk, and C.-Y. Lee, "Dance music classification using inner metric analysis: A computational approach and case study using 101 Latin American dances and national anthems," in *Proc. INFORMS Computer Society Conf.*, 2005.

[95] G. Tzanetakis, G. Essl, and P. Cook, "Automatic musical genre classification of audio signals," in *Proc. Int. Conf. Music Information Retrieval*, Bloomington, IN, 2001, pp. 205–210.

[96] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," in *Proc. Int. Conf. Music Information Retrieval*, Baltimore, MD, 2003, pp. 201–208.

[97] R. Mayer, T. Lidy, and A. Rauber, "The map of Mozart," in *Proc. Int. Conf. Music Information Retrieval*, Victoria, Canada, 2006.

[98] F. Morchen, A. Ultsch, M. Nocker, and C. Stamm, "Databionic visualization of music collections according to perceptual distance," in *Proc. Int. Conf. Music Information Retrieval*, 2005.

[99] M. Schedl, T. Pohle, P. Knees, and G. Widmer, "Assigning and visualizing music genres by Web-based co-occurrence analysis," in *Proc. Int. Conf. Music Information Retrieval*, Victoria, Canada, 2006.

[100] M. Goto and T. Goto, "Musicream: New music playback interface for streaming, sticking, sorting, and recalling musical pieces," in *Proc. Int. Conf. Music Information Retrieval*, 2005, pp. 404–411.

[101] R. van Gulik, F. Vignoli, and H. van de Wetering, "Mapping music in the palm of your hand, explore and discover your collection," in *Proc. Int. Conf. Music Information Retrieval*, Barcelona, Spain, 2004, pp. 409–414.

[102] R. van Gulik and F. Vignoli, "Visual playlist generation on the artist map," in *Proc. Int. Conf. Music Information Retrieval*, London, U.K., 2005.

[103] E. Pampalk and M. Goto, "MusicRainbow: A new user interface to discover artists using audio-based similarity and Web-based labeling," in *Proc. Int. Conf. Music Information Retrieval*, Victoria, Canada, 2006.

[104] M. Torrens, P. Hertzog, and J.-L. Arcos, "Visualizing and exploring personal music libraries," in *Proc. Int. Conf. Music Information Retrieval*, Barcelona, Spain, 2004, pp. 421–424.

[105] D. Bainbridge, S. J. Cunningham, and J. S. Downie, "Visual collaging of music in a digital library," in *Proc. Int. Conf. Music Information Retrieval*, Barcelona, Spain, 2004, pp. 397–402.

[106] G. Tzanetakis, "Manipulation, analysis and retrieval systems for audio signals," Ph.D. dissertation, Princeton Univ., Princeton, NJ, 2002.

[107] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrogram: A new musical instrument recognition technique without using onset detection nor F0 estimation," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Toulouse, France, 2006, pp. V-229–V-232.

[108] M. Goto, "Active music listening interfaces based on signal processing," in *Proc. Int. Conf. Acoustics Speech Signal Processing*, Honolulu, HI, 2007, pp. IV-1441–IV-1444.

[109] G. Tzanetakis and P. Cook, "MARSYAS: A framework for audio analysis," *Organized Sound*, vol. 4, no. 3, 2000.

[110] X. Amatriain, "CLAM: A framework for audio and music application development," *IEEE Software*, vol. 24, no. 1, pp. 82–85, Jan. 2007.

[111] J. S. Downie, J. Futrelle, and D. Tcheng, "The international music information retrieval systems evaluation laboratory: Governance, access and security," in *Proc. Int. Conf. Music Information Retrieval*, Barcelona, Spain, 2004.

[112] B. Vercoe, "The computer as musical accompanist," *ACM SIGCHI Bull.*, vol. 17, no. 4, pp. 41–43, 1986.

[113] R. B. Dannenberg, "An online algorithm for real-time accompaniment," in *Proc. Int. Computer Music Conf.*, San Francisco, CA, 1985, pp. 193–198.

[114] C. Raphael, "A probabilistic expert system for automatic musical accompaniment," *J. Computational Graphical Statistics*, vol. 10, no. 3, pp. 487–492, 2001.

[115] M. Casey and M. Grierson, "SoundSpotter/remix-TV: Fast approximate matching for audio-visual performance," in *Proc. Int. Computer Music Conf.*, Copenhagen, Denmark, 2007.

[116] D. Schwarz, "The Caterpillar system for data-driven concatenative sound synthesis," in *Proc. Digital Audio Effects*, London, U.K., 2003.

[117] N. Collins, "Audiovisual concatenative synthesis," in *Proc. Int. Computer Music Conf.*, Copenhagen, Denmark, 2007.

[118] M. Datar, P. Indyk, N. Immorlica, and V. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. Symp. Computational Geometry*, 2004.

[119] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for near-neighbor problems in high dimensions," in *Proc. Symp. Foundations of Computer Science*, 2006.

[120] M. Casey, "MPEG-7 sound-recognition tools," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 737–747, Jun. 2001.

[121] S. Quackenbush, S., and A. Lindsay, "Overview of MPEG-7 audio," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, Jun. 2001.

[122] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?" *J. Negative Results Speech Audio Sciences*, vol. 1, no. 1, 2004.

[123] R. Paivo, "Melody detection in polyphonic audio," Ph.D. dissertation, Univ. Coimbra, 2007.

[124] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[125] S2S2, "A roadmap for sound and music computing," in *Proc. S2S2 (Sound-to-Sense, Sense-to-Sound) Consortium*, Brussels, Belgium, 2007.

[126] M. Leman and A. Schneider, "Origin and nature of cognitive and systematic musicology: An introduction," in *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology*, M. Leman, Ed. Berlin, Germany: Springer-Verlag, 1997, pp. 13–29.

[127] D. Deutsch, *The Psychology of Music*, 2nd ed. San Diego, CA: Academic, 1999.

[128] M. Leman, *Embodied Music Cognition and Mediation Technology*. Cambridge, MA: MIT Press, 2007.

[129] M. Lesaffre, L. de Voogdt, M. Leman, B. de Baets, H. de Meyer, and J. P. Martens, "How potential users of music search and retrieval systems describe the semantic quality of music," *J. Amer. Soc. Inform. Sci. Technol.*, to be published.

[130] P. N. Juslin and J. Sloboda, *Music and Emotion: Theory and Research*. Oxford, U.K.: Oxford Univ. Press, 2001.

[131] A. Gabrielsson and P. N. Juslin, "Emotional expression in music," in *Handbook of Affective Sciences*, R. J. Davidson, H. H. Goldsmith, and K. R. Scherer, Eds. New York: Oxford Univ. Press, 2003, pp. 503–534.

[132] M. Slaney and W. White, "Similarity based on rating data," in *Proc. Int. Conf. Music Information Retrieval*, Vienna, Austria, 2007.

[133] D. J. Hargreaves and A. North, "The functions of music in everyday life: Redefining the social in music psychology," *Psychol. Music*, vol. 27, no. 1, pp. 71–83, 1999.

ABOUT THE AUTHORS

Michael A. Casey (Member, IEEE) received the Ph.D. degree from the Media Laboratory, Massachusetts Institute of Technology, Cambridge, in 1998.

He is Visiting Professor of Computer Science at Goldsmiths College, University of London, London, U.K. He was the Principal Investigator at Goldsmiths for two MIR-related grants: *Online Music Recognition and Searching II* (OMRAS2) and *Hierarchical Segmentation and Markup of Musical Audio* (SeMMA). He was a Contributing Editor of the MPEG-7 international standard for content-based multimedia description ISO15938-4 (audio) 2002 while a Research Scientist at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, before taking a position at Goldsmiths, in 2004. In 2008, Prof. Casey joined the faculty of the Music Department at Dartmouth College, Hanover, NH.

Remco Veltkamp studied computer science at Leiden University and completed the M.S. thesis at IBM Scientific Centre, Paris, France. He received the Ph.D. degree from Erasmus University, Rotterdam, in 1992.

Since 1995, he has worked at Utrecht University, where he contributes to the master program Game and Media Technology. He is currently an Associate Professor in the Department of Information and Computing Sciences, Utrecht University. Prior to Utrecht University, he worked at the Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands, and at the Technical University Eindhoven. His current research focuses on the algorithmic aspects of multimedia retrieval, like the algorithmic design and analysis, and experimental verification.

Dr. Veltkamp is Editor of the *International Journal on Shape Modeling* and *Journal Pattern Recognition*, and he is the organizer of the Dagstuhl Seminars Series on Content-Based Retrieval.

Masataka Goto received the Ph.D. degree in engineering from Waseda University, Japan, in 1998.

He then joined the Electrotechnical Laboratory (ETL), which was reorganized as the National Institute of Advanced Industrial Science and Technology (AIST), in 2001, where he has been a Senior Research Scientist since 2005. He served concurrently as a Researcher in Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST), from 2000 to 2003, and as an Associate Professor in the Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba, since 2005.

Dr. Goto has received 21 awards, including the DoCoMo Mobile Science Awards "Excellence Award in Fundamental Science," an IPSJ Best Paper Award, IPSJ Yamashita SIG Research Awards, and the Interaction 2003 Best Paper Award.

Marc Leman is a Research Professor in systematic musicology at Ghent University, Ghent, Belgium. His personal research is in epistemological and methodological foundations of human musical involvement. He is a Coordinator and Partner in several ongoing projects in the field of musical content processing, among which are MAMI, GOASEMA, and DEKKMMA (www.ipem.ugent.be). He is the author of several books and scientific articles, including *Music and Schema Theory* (Springer-Verlag, 1995), *Music, Gestalt, and Computing* (Springer-Verlag, 1997), and *Embodied Music Cognition and Mediation Technology* (MIT Press, 2007).

Dr. Leman was Editor-in-Chief of the *Journal of New Music Research* from 1987 to 2004.

Christophe Rhodes received the B.S. degree in physics, in 2000, and Ph.D. degree in applied mathematics, in 2004, both from the University of Cambridge, London, U.K.

He is a Research Assistant in the Intelligent Sound and Music Systems group, Computer Science Department, Goldsmiths College, University of London, London, U.K. His Ph.D. research formed part of the Planck satellite science program, while his current research interests span musical information retrieval and representation, user interfaces, and language design and implementation.

Malcolm Slaney (Senior Member, IEEE) is a Researcher with Yahoo! Research, Sunnyvale, CA, and a Consulting Professor at Stanford University, Stanford, CA. He is a coauthor of the book *Principles of Computerized Tomographic Imaging* and coeditor of the book *Computational Models of Hearing*.